A PROFILE OF DATA PRESERVATION ACTIVITIES IN UNIVERSITY DATA LIBRARIES AND ARCHIVES

ALICE ROBBIN
DATA AND PROGRAM LIBRARY SERVICE
UNIVERSITY OF WISCONSIN-MADISON

I. Introduction

For the last two decades, funding priorities have dictated allocation of resources to national centers as principal sources of archival data for research and teaching. However, the importance of local centers in universities for preserving, disseminating, and describing computer-readable data cannot be understated. These local, campus-based libraries and archives play a critical role in the information transfer system for the social scientific community, both within and outside their institutions. Even though technological developments make it possible to receive and transmit data from great distances (thereby obviating in principle the need for being a local repository), current economic and political realities have constrained efficient use of the modern computer technology. These realities suggest that distributed data centers at the local level will continue to play a major role in the transmission of information.

It is worthwhile briefly to describe the important role these local data libraries have played in the last fifteen years, and to suggest how they will continue to participate in social inquiry. A number of the centers were established before their national governments established machine-readable archives divisions within the national archives. As a result they became de facto repositories for federally-produced data. For many studies obtained from outside their institutions, no adequate archival facility existed elsewhere, and the centers became by default permanent repositories. In other cases, although these centers had not been designated repositories for federally-produced machine-readable data, federal agencies turned to them for assistance in retrieving data files which the agencies produced but could no longer retrieve. The data centers also acted as a transfer agent and depository for data files produced by foreign governmental agencies and research institutes. For other studies which could theoretically be obtained elsewhere, some data centers assumed archival responsibility on the grounds that the supplier could not adequately preserve and maintain the valuable resource, or that it was economical in time and money for the university data center to maintain an on-site copy of the data.

There are of course other reasons for preserving data and for supporting a system of distributed data centers within a national and international context. A growing literature (cf. Clubb, Hofferbert, Miller, Rokkan, Boruch and Wortman, Nesvold) makes cogent arguments for supporting the

national data centers, data libraries, and data laboratories for social scientific research and teaching. The underlying philosophy of preservation and access holds that transfer of the data collections from private research organizations, governmental agencies, and foundations to these data centers has greatly magnified the return on the original private and public investment in data gathering, and has encouraged and facilitated social scientific inquiry. The data archive participates in the processes of innovation, dissemination of scientific results, and information transfer. It acts as a scientific laboratory which encourages the sharing of data, multidisciplinary exploitation of evidence, and "multiple and complex analytical applications" (Hofferbert and Clubb, p. 383). The data center makes a pedagogical contribution by allowing the student to participate directly in empirical scientific inquiry, developing problem-solving modes of behavior like those of students in the natural sciences. Less obviously, the data archive plays a role as an agent for administrative and technical assessment of information transfer activities and mechanisms. It offers administrators and researchers the opportunity to assess, in a rigorous and analytical fashion, the technical, administrative, economic, and policy issues related to standards of data quality, documentation, access, and diffusion. Collegial behavior is facilitated. Common access encourages standardization and "commonality of research among widely separated scholars" (Miller, p. 411). Finally, a democratic society such as ours is committed to public access to the products of research and to knowledge-producing modes of behavior, and it is through the data centers that this access is facilitated.

Rockwell argues that in the 1980s these data centers will assume greater importance for the social sciences than they have in the past, because of such factors as the increasing cost of gathering data and conducting surveys. We can expect that in the 1980s social scientists will capitalize on resources such as those found in these local data centers. For example, the increasing number of time series of replicated data is becoming a major resource for cohort and panel analysis. Rockwell suggests that it is unlikely that we will see many new surveys mounted during the coming decade and that "from the perspective of social indicators research, the resources of these data centers are important precisely to preserve long time series of data." He gives the following reasons: "More generally, cumulative social science research demands ample opportunity to return to the same data bases for repeated inquiries. The journals increasingly reflect the field's recognition of the importance of good social measurement: standardized measures available on a repeated basis in a time series data base."

Some of the data centers now face critical problems preserving data on magnetic tape, the principal medium of long-term storage, because of changes in computer technology, aging of the collection, and magnetic tape deterioration. The Data and Program Library Service at the University of Wisconsin, for example, has found that a growing number of tapes, even ones guaranteed for 15 or 20 years, are developing non-recoverable read-write errors. These errors are seriously affecting the quality of the data preserved.

The problems of physical deterioration are not limited to elderly reels of tape. DPLS has encountered quality control and deterioration problems with tapes purchased two and three years ago from a manufacturer with whom DPLS has dealt for some years. In addition, DPLS has found, as have others, that the computing center's tape drives have influenced the condition of DPLS magnetic tapes, and have been responsible for parity error problems.

In response to these and other problems, including changes in computer technology, DPLS instituted a minimal tape maintenance program three years ago to convert data sets written on older tapes. When DPLS began calculating the costs of a full-scale tape maintenance program, it became obvious that the magnitude of its collection, the staff time required to rectify the problems, and the computing and capital equipment expenditures required were beyond DPLS's limited resources. It was at this point that the DPLS staff began an investigation into current and potential mass storage media developments and decided to conduct a survey of data centers to ascertain how their staffs are handling data stored on magnetic tape. DPLS thought that the literature might provide some insights into developing its own program of tape maintenance and would also provide the data library and archival community with information on the status of data preservation.

The following discussion is a report on technical problems related to preserving machine-readable records, and on the findings from the small survey. Although some may believe that data stored on magnetic media can be treated like a book left on a shelf to gather dust, that in fact is not the case. Problems discussed in Part II of this paper suggest the need for accelerating the development of long-term archival storage media now in the experimental stage, particularly because of the increasing generation of statistical data. The problems facing the North American data libraries and archives, and how their staffs are coping with the preservation of data stored on magnetic tape, are taken up in Part III. The survey results suggest that data library and archive staffs recognize that preserving their valuable resources is necessary to ensure continuing support to the research and teaching community. All have either developed formal tape maintenance programs or are aware of the need to develop better practices for preserving their collections.

II. Technical Problems Associated with Preserving Machine-Readable Records

Among the major problems faced by the librarian and archivist who deal with computerized records is that current magnetic storage media[1] and most of the mass storage devices[2] now in various stages of development fail to meet archival storage requirements--that is to say, the preservation of digital data for a very long period.

Problems associated with permanent preservation of data include the physical size of data, machine independence and media standardization, reliability of the storage medium, the medium's sensitivity to environmental conditions, lifetime maintenance and cost of the medium, accessibility of the information, and cost of duplication. Volz, Dollar, and Geller elaborate on these problems from the archivist's perspective, and their comments bear repeating. To convey the problem of physical size of data, Volz presents this example:

> A typical book contains about 10 to the 7th bits. (I interpret him to mean bytes or characters rather than bits.--Author) Encyclopedia Britannica contains about 10 to the 9th bits. Such volumes can be readily stored in today's technology. For example, Britannica would roughly fit on a single IBM 2220 disk pack. However, the problems are not storage of a single volume of text, but rather large collections of such volumes.

The DPLS collection contains many such "volumes." For example, one data file in the collection contains approximately 545 million characters. And although most files do not approach this size, the DPLS collection contains more than 6000 data files. In the last year, data files which fill two or more 2400-foot magnetic tape reels have become the norm. We expect that with the 1980 U.S. Census of Population and Housing, average data files will be stored on multiple reels of tape. Increasingly as scholars turn to administrative records for research, data collections will require adequate storage devices.

With the exception of magnetic tape, which offers compatibility when written on different tape drives if the same density and character codes are used and if utility software is available to translate the character codes (Dollar, p. 29), all other magnetic storage media (to my knowledge) are machine-dependent and non-standard. (For example, cassette tapes produced by the Sykes Corporation cannot be used on IBM equipment.) Without machine independence and media standardization, archival storage becomes a very nearly unsolvable problem in compatibility. Machine dependence also affects preservation of data in another way: system files written on one machine cannot be transferred to another computer system (e.g., SPSS system files produced and transferred between computers of the same manufacturer may not be readable).

Preservation is also affected by machine obsolescence. The rapidly changing computer technology has resulted in removal of equipment used for the initial creation and copying of the data. Thus, the data archives created in the 1960s, when magnetic tape was read and written on 7-track tape drives, have found that their computing centers have replaced their equipment, and that their data cannot be read on the new equipment. The result is reduced access to their collections and increased preservation costs, because all their data must be converted to meet the specification of the new equipment.

Another archival concern is how long the media retain a reliable image of the data. Volz notes that "due to the relatively short span of time over which really large mass memory devices have been in use, only limited empirical data is available." Dollar comments (p. 29):

> Permanent preservation of digital data requires storing the records in a mode in which under normal conditions the recording signal will not degrade and the medium will not deteriorate to the point that data recovery is impossible... This means a non-erasable mass storage capability which is not vulnerable to irreparable loss of archival records through human carelessness or system malfunctions.

Geller, manager of the Magnetic Media Group at the National Bureau of Standards, elaborates (pp. 37-38): "Experimental evidence has shown that failures to extract the information from magnetic media are almost always attributable to the physical deterioration of the media rather than to the deterioration of the data." Although there are now estimates of an archival lifetime of 10 years for magnetic tape, there is really no way to simulate the reliability of a magnetic tape as a storage medium for a long period of time. Most archival data, our records indicate, are accessed infrequently (every few years) or not at all. Thus, without a regular and frequent maintenance program, tape deterioration is not apparent until the data are requested, accessed, and copied. Archives in existence since the 1960s face aging

problems associated with the quality of the magnetic storage medium. Tapes produced before 1972 cannot be reused because of the deterioration resulting from the poorer magnetic tape. Archives whose holdings date from before 1972 may need to replace large parts of their magnetic tape collections.[3]

Environmental and handling conditions affect the lifetime of the magnetically encoded data and necessitate expensive environmental controls to prevent adverse forces and "debilitating humidity and temperature conditions" (Dollar, p. 29) from affecting the recording signal and from impairing the storage medium. The magnetic tape on which all data libraries store their collections is well known to be susceptible to environmental conditions. For example, a 2400-foot tape will "try to change its length by approximately one foot for every 10 degree change in temperature or 10 percent change in humidity," Volz states. He continues:

> Friction of the tape wound upon a reel tends to prevent these changes in length from taking place, resulting in high pressures on the tape and perhaps some permanent changes to the tape. Occasionally some slippage may occur resulting in flaking of the oxide from the surface of the tape, which not only may itself lose information, but creates debris which will interfere with the reading of other bits.

Lifetime maintenance and media costs are considerable. To ensure accessibility to the stored information, tapes must be duplicated so that the archive always retains a reliable image of the data (i.e., so that unrecoverable errors on one file do not result in irretrievable loss of data and a file's integrity is assured by maintaining a second copy. For an archive of record, costs of maintenance and preservation can be significant: if the original data file has gone through several data processing activ ties (updates, corrections) over time, all the file iterations must be maintained. Data files stored on magnetic tape must be "rolled over" (i.e., copied) at least once every two or three years. This entails a considerable allocation of resources for an archive. New magnetic tape must be purchased and computer time must be bought for copying; staff time must be available for carrying out the maintenance program--preparing the software, documenting the procedures, evaluating results of magnetic tape quality, and completing the administrative records to document output onto the new storage medium. To the extent that documenting and administrative record-keeping can be automated, human resource savings can be significant, since it is the record-keeping activities which are labor-intensive.

What this discussion suggests is that the social scientific research activity requires adequate funding to maintain necessary supporting facilities. The laboratory of the social scientist requires modern and reliable mass storage equipment for long-term preservation of the materials used for scientific discovery. Maintenance, while perhaps more visible in a natural sciences laboratory or a traditional library (where there are devices for controlling humidity, facilities for rebinding books, and programs for the security and physical protection of the collection), is a necessary condition for social scientific activities.

III. The Survey

Between January and March 1980, DPLS conducted a mailout-mailback survey on tape maintenance activities in data centers (libraries and archives) located in North America. Three sources of information were used to identify these centers. The list of data centers provided in SS Data: A Newsletter of Archival Acquisitions was supplemented by a review of all the catalogues of data holdings at DPLS and of the DPLS administrative correspondence files. With the exception of one survey research data archive (which was identified in SS Data), survey research institutes were systematically excluded, as were governmental archives (e.g., the U.S. National Archives and Records Service and the Public Archives of Canada), national and international repositories such as the Inter-university Consortium for Political and Social Research and the Roper Public Opinion Research Center, and federal agencies, such as the U.S. Bureau of the Census and Statistics Canada, which disseminate data to the social research community.

ICPSR member institutions which provide access to ICPSR data through a departmental faculty member were also excluded because most of those departments would not qualify as a data center or library/archive in any rigorous way, particularly because control over their materials is lacking and because they play a minimal or nonexistent role in information dissemination about data for other than the Consortium's holdings.[4] (This statement is of course offered without any hard evidence, and needs verification.)

The questionnaire was sent to 37 organizations, of which 34 had responded by the end of March 1980. After reviewing the completed questionnaires, four data centers were deleted from the final sample. Either most of the items in the questionnaire were not relevant to their organization, or their holdings were so specialized that the information we sought could not be utilized in our analysis, or they were not a university-affiliated organization. Only one university-affiliated data center did not respond. The final sample on which our analysis is based is 30 university data libraries and archives.[5] Because we cannot say with any assurance that our original list constituted the universe of data libraries and archives in North America, our review of tape maintenance activities offers no tests of statistical significance. Rather, our intention here is to describe current data library maintenance activities and to present a profile of these activities in a select group of data centers.

We need to probe more deeply into the state of these data organizations to understand how they are structured, what activities they carry out, and their influence on social scientific activity at their institutions. These are all important questions for which we have little or no information. But certainly these questions are worth pursuing, for they add another dimension to what we know of how organizations charged with information transfer participate in the knowledge flow process. The very high response rate and the enthusiasm with which people responded is evidence that these staffs do want more information about the problems of their colleagues and how they are coping with current economic and political realities.

A. A Profile of North American Data Centers

In an effort to reduce respondents' reporting burden, questions about their organizational structure, activities other than tape maintenance, funding,

and collection were kept to a minimum. We wanted to know when the center was established and the estimated size of the data and tape libraries. We posited that an early establishment date and a large collection would lead to inadequate levels of funding for purchase of magnetic tapes and for maintenance activities. It would lead also to dissatisfaction with the quality of the maintenance program. We were interested in knowing whether there were any differences in maintenance activities if a data center were an independent department or affiliated with another department, library, research organization, or computing center. We wanted to know from where the data collection was derived, that is, its original sources; its estimated growth in data files and magnetic tapes over the next five years; how the data were used; and whether the staff had noted any changes in the number of requests and in types of files requested in the last two years.

On the basis of our services at DPLS, we have noted an increasing tendency toward use of government-produced data and toward larger and more complex files requiring at least several reels of magnetic tape. Until rather recently, DPLS served as a research support facility, and undergraduate class projects have constituted no more than 15 to 20 percent of our use.

We wondered how different or similar the situation was at other data centers. We thought that if staffs were noting changes in the number of data files and types of data being requested, this could signal the growing complexity of the data being used by members of their institutions, and of increasing demands being placed on the library staff. Neither the questions nor the responses permit us to infer what is happening at the local level, although we can make some educated guesses.

Concerning the computer facility available to the data center, we wanted to know what computer is primarily used for most activities. We wanted to know how the data center stores its data and the current storage mode on magnetic tape. We then turned our attention to whether the organization was encountering or anticipated tape storage problems and whether the staff had investigated any ways other than magnetic tape for transfer and long-term storage.

Our last set of questions concerned the burden of changes made to the computing center, and the adequacy of financing to preserve the integrity of the collection and carry out maintenance activities.

Figure 1 shows that 17 of 27 data centers, or 63 percent, were established between 1966 and 1972.[6] These years correspond to a period when universities and external funding agencies provided increased financial support to the social sciences. Between 1973 and 1977, we see a decline in the number of data centers being established; but in 1977 we once again see an increase. Figure 2 shows that more than half of the data centers (N=17) have collections of between 100 and 699 data files. The size of the data collection appears to have little relationship to when the center was founded. See Table 1. Why is unknown; it may have something to do with the size of the user community, as well as the resources available for collection building. Figure 3 shows the size of the magnetic tape collection. Here we see that 18 of 30 centers have fewer than 399 magnetic tapes.

Figure 1. Date of establishment.

1963	1	*
1964	1	*
1965		
1966	1	*
1967	3	***
1968	4	****
1969		
1970	2	**
1971	3	***
1972	4	****
1973	2	**
1974	1	*
1975	1	*
1976	1	*
1977	3	***
	27	

Figure 3. Number of reels.

-100-199	9	*********
200-399	9	*********
400-599	4	****
600-799	2	**
800-999		
1000+	6	******
	30	

Figure 2. Estimated size of data collection.

100-299	6	******
300-499	6	******
500-699	5	*****
700-899		
900-1099	2	**
1100-1299	2	**
1300-1499	1	*
1500-1699	2	**
1700-1899		
1900-2099		
2100-2299	1	*
2300-2499	1	*
2500-2699		
2700+	3	***
	29	

Table 1. Date established by size of collection.

		Size of collection (number of reels)			
		100-899	900-2999	3000+	
Date established	1973-	6	2	0	8
	1968-72	6	4	2	12
	-1967	4	2	1	7
		16	8	3	27

When we look more closely at the relationship between the number of data files and number of magnetic reels, we see some connection, but at the same time we see that storage conditions vary. Several data centers utilize modern storage technology to pack large amounts of data on a small number of tapes, while others store their data at much lower densities. See Table 2.

We asked the staffs to estimate the growth in the number of data files and magnetic tapes per year over the next few years. For those centers which supplied this information, estimated growth in the number of files per year was the following: 32 percent (N=7) estimated between 10 and 30 files; 64 percent (N=14), 31 to 75 files (a large spread); and four percent (N=1), 150 or more. Estimated growth in the number of magnetic tapes, as expected with the advance in storage technology, was 43 percent (N=9), between two and 25; 48 percent (N=10), between 30 and 80; and nine percent (N=2), between 100 and 125.

The next set of questions dealt with the sources of the data in the collection, how the collection was used, and whether there have been changes in the types of requests made to the staff. By far the largest source of data is the Inter-university Consortium for Political and Social Research (ICPSR). Of 29 data centers reporting sources of data, 45 percent (N=13) report having up to 59 percent of the collection from ICPSR, while 55 percent report between 60 and 100 percent ICPSR materials. The average was about 65 percent. Surprisingly, acquisitions from the federal government are very low; 79 percent (N=23) report between 0 and 25 percent of their holdings from federal sources. Because social scientists are increasingly consuming federally-produced data, we expected a greater percentage of the centers' collections to be from the government. Of course, it is quite possible that the centers are obtaining federal data from ICPSR and then reporting ICPSR rather than the government as the supplier.

Not surprisingly, the private sector accounts for an insignificant percentage of collections: 78 percent of the centers report between 0 and 5 percent. The center's own institution, the Roper Center, and other distributors make up only a small fraction of the remaining suppliers: 83 percent report no more than 15 percent from their own institutions; 83 percent have between 0 and five percent from the Roper Center; and 79 percent get from 0 to 10 percent from other sources, primarily international and intergovernmental.

Data centers have typically been the product of research activity at an academic institution. As new generations of graduate students trained in quantitative methods and data handling enter the teaching profession, quantitative methods and the use of the computer are introduced into the classrooms. Considering that data handling was the purview of sophisticated graduate students during the middle and late sixties, we should expect a third generation of former graduate students now to be faculty members and a data center to be responsive to their teaching needs. We therefore expect that instructional use of the data center, as a laboratory for scientific activity, will constitute a significant part of its over-all use. Table 3 shows the use of the collection for teaching, research, and other (primarily policy) activities. Here we see that research is indeed the principal reason for the use of the data center (mean=65 percent), but that instructional use does represent a significant activity (mean=33 percent).

We wondered whether staffs had noted any changes in the number of requests and types of data files requested during the last two years. We asked whether there were any increases in the number of requests, whether the files were structurally more complex, and whether files were requiring more than one or two reels

Table 2. Size of data collection (number of data files) by size of magnetic tape collection (number of tapes).

Size of data collection	Number of tapes: 100-399	400-999	1000+	
1700+	2	1	2	5
700-1699	7	0	1	8
100-699	8	5	2	15
	17	6	5	28

Table 3. Percentage of collection used for teaching and research.

Instruction %	N		Research %	N		Other %	N	
0-20	13	(45%)	10-40	6	(21%)	0	26	(90%)
25-50	11	(38%)	50-80	16	(55%)	5	1	(3%)
60-90	5	(17%)	90-100	7	(24%)	33	1	(3%)
	29			29		100	1	(3%)
							29	
mean = 33%			mean = 65%					

Table 4. Changes in the types of requests and data files over past two years.

Increase in number of requests	5	(17%)
Files more complex	4	(13%)
Files require more than 1 or 2 reels of magnetic tape	2	(7%)
Increase in file requests and data files more complex	1	(3%)
Increase in file requests and files require more reels	1	(3%)
Increase in requests and files more complex and require more reels	9	(30%)
Files more complex and require more reels	3	(10%)
No changes noted	2	(7%)
Not ascertained	3	(10%)
	30	

of magnetic tape. The increase in the number of requests and growing complexity of the data were the two largest single categories of changes noted during the last two years; 30 percent of the respondents noted all three changes.

The next series of questions reports on media storage of the collection, current and anticipated storage problems, and investigation of other storage media. As Table 5 indicates, magnetic tape is the medium of storage for the the data centers, with 29 of 30 centers storing between 95 and 100 percent of their data on magnetic tape. The current storage modes appear to be EBCDIC, nine channel, 1600 BPI, although there are still data centers (almost a third) which store their data in seven channel, even parity, 556 BPI, and an increasing number of centers which are moving to EBCDIC, nine channel, 6250 BPI.[7]

More than half the data centers said that they had no tape storage problem now (N=16, 53 percent), while 47 percent (N=14) reported problems. When asked what kinds of storage problems they were encountering, almost half gave lack of space as the principal one. Table 6 indicates a pattern to the tape storage problem: too many tapes, lack of space (usually associated with on-site storage rather than off-site), leading to off-site (or remote) storage as a necessity.[8] Data centers located in computing centers and affiliated with libraries indicated no problems, whereas those which were independent or affiliated with research organizations appear to be encountering storage problems.

In response to the question about whether they anticipate a storage problem in the future, 62 percent (N=18) responded yes, and 39 percent (N=11), no. Of the 18 who anticipate problems, the need for storage (whether on- or off-site), storage costs, and the large collection were cited as major problems.[9]

We wondered whether any data centers had investigated ways other than magnetic tape for transfer and long-term storage of their data collection. 37 percent (N=11) had, whereas 63 percent (N=19) had not. For those who had investigated other media, off-line disk was cited by five, video disk by two, microfilm by one, and computer by two.

Finally, we were interested in knowing whether those who had noted a tape storage problem had also investigated other media for long-term preservation. As our Table 7 indicates, 57 percent (N=8) of the 14 responding that there were tape storage problems had not done any investigating, while 31 percent of those indicating no tape storage problem had investigated other ways of storing data.

The last set of questions explores the impact of changes in computer technology, of increased requests for data files, and of the adequacy of funding for tape purchase and maintenance activities. We asked whether the computing center had made or planned to make changes which have affected or would affect the way in which the data center stored its data. Almost half (47 percent) noted that the computing center had made changes but the changes did not affect the way the data were stored; however 33 percent did note that seven channel tape drives were being phased out and that the data center must convert its data collection. In response to the question about whether changes in the types of requests being made for data had placed a burden on the library in terms of available budgetary resources to maintain the physical integrity of the collection, 73 percent (N=22) responded no, while 27 percent

Table 5. Media storage of the collection.

Cards			Tapes			Disk		
%	N		%	N		%	N	
0	22	(73%)	75	1	(3%)	0	25	(83%)
1	2	(7%)	95	5	(17%)	1	1	(3%)
2	3	(10%)	98	2	(7%)	3	1	(3%)
5	2	(7%)	99	3	(10%)	5	2	(7%)
25	1	(3%)	100	19	(63%)	40	1	(3%)
	30			30			30	

Table 6. Tape storage problems.

	First problem	Second problem
Too many tapes	3 (21%)	
Storage charges	1 (7%)	1 (33%)
Not enough space	6 (43%)	
Lack of environmental controls		1 (33%)
Unused files combined with active files	1 (7%)	
Backup charges prohibitive		1 (33%)
Remote storage necessary	2 (14%)	
Age	1 (7%)	
	14	3

Table 7. Whether tape storage problem exists by whether data center has investigated other storage media

		Investigated other storage media		
		Yes	No	
Tape storage problems	Yes	6 (43%)	8 (57%)	14
	No	5 (31%)	11 (69%)	16
		11	19	30

(N=8) said yes. While 90 percent (N=26) claimed adequate funding for tape, only 66 percent (N=19) said they could afford a tape maintenance program. For the 10 centers which said that funds were inadequate, major reasons cited were that there was not enough staff (30 percent) or that both funding and staffing were inadequate (30 percent).[10] The data centers affiliated with a computing center and a research institute have no difficulty in supporting a tape maintenance program, while more than half the centers which are independent departments or affiliated with a teaching department cite inadequate funds to maintain such a program.

Our last question on financing asked where financial support came from. The data library budget is the source for maintenance activities for 41 percent (N=12); computing centers account for 17 percent (N=5); a mix of library and computing center for 10 percent (N=3); the data library budget and ad hoc requests for maintenance funding, 10 percent (N=3). The remaining percentage was divided among ad hoc requests, other, no support provided, and a mix of data library, computing center, and ad hoc funding.

B. Tape Maintenance Activities

In this section we examine the quality of tape maintenance activities, degree of satisfaction with the data center's program, and whether there is any difference in the quality of activities between those who are satisfied and those who are dissatisfied with their program. Our concern here is with the set of activities to preserve data on magnetic tape. According to the literature, tape maintenance involves creating back-ups, controlling the movement of the magnetic medium from abrupt environmental changes, maintaining environmental controls (temperature and humidity) in the storage area(s), monitoring these controls periodically to observe changes, having access to an off-site facility for storage, and maintaining a record keeping system for effective control and administration of the tape library.

What we observe in Tables 8 and 9 is that data centers can clearly be given high marks for protecting their data by maintaining back-ups of every data file and controlling the movement of the medium; but their monitoring of environmental controls, providing off-site storage for the data, and maintaining complete evaluation histories of the magnetic tapes are not as good as they should be. Considering that data centers are transferring data from supplier to data center and data center to computing facility, fully 66 percent are not letting their magnetic tapes sit for at least 24 hours before mounting them. This can result in too much stress on the medium, cracking, and data destruction. While 70 percent say they have environmental controls in the storage area(s), only 33 percent say they monitor the controls. Unless data centers can guarantee full protection (against loss) of their master and back-up copies, off-site storage of at least one copy is a requisite for preservation. Yet only 57 percent (N=17) say they have access to off-site storage. Sixty percent of the data centers say they maintain adequate procedures for recording status of each magnetic tape; yet further examination of their responses indicates that this is not so: fully one-third do not appear to be recording the results of their periodic review of their archival and working tapes. Although 77 percent state that periodic review is carried out, cleaning and testing is carried out only by 55 percent, and certification and precision rewinding by around 23 percent; however, a number of the data centers (particu-

Table 8. Tape maintenance activities.

"Many data and tape libraries have a set of activities to preserve their data on magnetic tape. Check as appropriate those carried out by your library or archive."

	Yes	No	Total
a. Maintain back-ups of every data file	29	1	30
b. Control movement of medium	24	6	30
c. Magnetic tape sits for 24 hours	10	20	30
d. Environmental controls in storage area(s)	21	9	30
e. Monitoring of environmental controls	10	20	30
f. Off-site facility for storage	17	13	30
g. Record history of status of each mag tape	18	12	30
g1. Age	14	4	18
g2. Manufacturer	8	10	18
g3. Size	12	6	18
g4. Certification	9	9	18
g5. Evaluation history	6	12	18
g6. Other (tape contents)	6	12	18
h. Periodic review of archival and working tapes	23	7	30
h1. Cleaning	12	10	22*
h2. Testing (evaluation)	12	10	22
h3. Certification	4	18	22
h4. Precision rewinding	5	17	22
h5. Other (roll-over, reading)	17	5	22

*1 not ascertained.

Table 9. Contents of the record keeping system.

"Do you have a record keeping system (either manual or automated)? Check as appropriate."

	Yes	No	Total
a. Identifies each reel	30	0	30
b. Identifies reel's contents	30	0	30
c. Provides location (and movement) of the reel	22	8	30
d. Describes status	17	13	30

Table 10. Periodic tape cleaning.

Yearly	5
Every two years	4
Every 5 years or more	3
Not at all/very seldom	9
Other (ad hoc basis)	6
NA	3
	30

larly the IBM community) are using special software to scan their tapes before the data are used. As Table 10 shows, only 30 percent (N=9) are regularly cleaning their tapes (yearly or every two years is what is recommended). Some 47 percent do not clean their tape collection at all or on an ad hoc basis. Only three data centers indicated that they neither had developed nor had access to special software to evaluate the physical integrity of their magnetic tapes; thus, maintenance responsibilities (or the lack thereof) are not explained by inaccessibility of software. And although almost half say they must pay for cleaning and evaluating services supplied by their computing services, only a few data centers, as we described earlier, have indicated a funding problem. Rather, the explanation probably lies in the availability of staff to carry out maintenance activities. Almost two-thirds of the sample stated that the library staff is responsible for tape maintenance. A more in-depth analysis of the level of responsibilities and demands placed on the data center staffs would give us more information about this aspect of the tape maintenance problem.

The next two questions deal with the level of satisfaction with the data center's tape maintenance activities. In response to the question, "Are you satisfied with the things you do to protect your collection?" 53 percent (N=16) said "yes," and 47 percent (N=14) said "no." Probing further into the "no" responses, we asked, "If not satisfied, would you do any of the following?" Clearly, respondents are aware that they need to improve present practices of tape maintenance: tape quality must be monitored on a regular basis. Somewhat less than half responded that they must upgrade their record keeping practices. Only a small percentage attend to the need to establish environmental controls. It may very well be that they believe that they have less direct control over site environmental conditions and that, therefore, any attempts to influence the quality of these conditions would be fruitless. On the other hand, they may feel that environmental controls are already satisfactory and that this is not an issue in their tape maintenance practices.

We wondered whether there were any differences between those respondents who said they were satisfied with their present practices and those who said they were not, with respect to the activities each group is carrying out. Table 12 looks at all respondents who said they carry out maintenance activities. Comparing satisfied data center staffs to the dissatisfied ones (both conducting good maintenance practices), we see little difference in the absolute numbers in each group except in two areas: more satisfied than dissatisfied staff members control the movement of magnetic tape and record the status of each magnetic tape in the collection. In sum, it might be suggested that the degree of satisfaction with one's maintanance practices lies in the quality of record keeping.

C. Needs

The last question in our survey asked respondents whether a document on minimal standards for tape maintenance of an archival data collection would be useful to them. With only two exceptions, the response was positive. We also asked them what they would like to see in such a document. The responses are described in Table 13. Indeed, the greatest interest lies in report forms for record keeping in the tape library and a bibliography of the state-of-the-art research on archival storage (23 of 28 respondents). Next are procedures for protecting the magnetic tapes that undergo environmental changes, and inventory control procedures (17 and 18 of 28, respectively). The responses

are consistent with behavior reported by the data centers' staffs and known to DPLS: record keeping is always a lower priority in an organization which has user services as its primary goal. Record keeping is neglected because it takes time and is transparent to the user. Staff is usually inadequate to support quality record keeping (which also includes inventory control).

Table 11. Satisfaction with tape maintenance activities.

"If not satisfied with present maintenance practices, would you do any of the following?"

	Yes	No	Total
Maintain back-ups of master files	1	12	13*
Monitor tape quality on regular basis (evaluation, certification)	8	5	13
Develop complete records on status of every tape in collection	6	7	13
Establish environmental controls	2	11	13
Establish off-site facility for tape storage	5	8	13

*1 not ascertained.

Table 12. Satisfaction/dissatisfaction with tape maintenance practices by those who conduct tape maintenance activities.

	Satisfied	Dissatisfied	Total
Let mag tape sit for 24+ hours	4	4	8
Control movement of mag tape	13	10	23
Establish environmental controls	10	9	19
Establish off-site facility	8	8	16
Conduct periodic monitoring	5	4	9
Record tape history and status	10	7	17
Carry out periodic cleaning	11	10	21

Table 13. Contents of a document on tape maintenance.

	Yes	No	Total
Procedures for protecting mag tapes which undergo environmental changes	17	11	28
Procedures for maintaining environmental controls in storage facility	13	15	28
Report forms for managing the tape library	23	5	28
Inventory control	18	10	28
Bibliography of state-of-the-art research on archival storage	23	5	28

Also, many believe that record keeping takes one away from the data, which are the raison d'etre of the center. The reality, however, is that without good record keeping practices, good user services cannot be provided and the collection is placed in jeopardy.

IV. Concluding Remarks

During the coming decade, precisely at a time when managers and administrators have come to recognize the importance of organizations which preserve, maintain, and disseminate statistical and other data, university-affiliated data centers will be faced with limited funds to maintain their collections. Obviously the economic realities call for creative technical and administrative solutions to the costly problem of data preservation. One solution is the development of storage devices which ensure long-term and stable preservation, to reduce the cost of yearly or bi-yearly file roll-over. Devices are now in the experimental stage, and prototypes offer hope that effective media will be available at reasonable cost within ten years. Another solution is better administrative practices to reduce the labor-intensive activity of record keeping. The computer and data base management software offer an opportunity to become more efficient and cost-effective--that is, to employ labor-saving devices for maintaining records of the data and tape libraries, inventory control, and retrieval and updating of information for periodic review of the status of the data and storage medium. Nevertheless, both the new storage devices and use of data base management systems are initially costly items: tape drives which permit writing of data at a density of 6250 BPI may involve outlays of anywhere between $125,000 and $150,000--perhaps beyond the means of all but the largest computing centers.

Development of the data bases for record keeping systems will involve a sophisticated programming staff familiar with data base management; for although most computing centers already provide some type of data base management software, it is usually not designed with administrative record keeping in mind. It requires software interfaces, necessitates substantial data base investment and data entry personnel, and incurs continuing operational and maintenance costs. Unless the administrative data base can be designed with multiple users in mind, developmental costs will have to be borne by the data center itself. Unless the data center has unlimited free computing and programming assistance, use of labor-saving devices such as a data base management system will not occur. The best strategy for a university-affiliated data center with limited funds is to investigate the potential user market for such automated administrative record keeping systems and to convince this user community of the need to develop good administrative practices to maintain their data.

In the meantime, however, data centers would be well advised to upgrade their present practices of tape maintenance to preserve access to their collections. Inadequate attention is being given to the importance of environmental controls and the need for monitoring these controls on a regular basis. Better protection for the collection through off-site storage of the master files is required. There appears to be too much reliance on the data center's computer center for carrying out basic maintenance. Computing centers are primarily involved in throughput operations and not preservation activities. Tape cleaning and evaluation need to be performed regularly and must be followed by adequate record keeping of the evaluation. Since the majority of the data

centers do not hold large collections, periodic cleaning and evaluating of their tape libraries should not prove too time-consuming to be carried out within the constraints of their budgets. Since all data centers expect growth in their collections, and particularly since they have probably underestimated this growth, they would be well advised to activate a program of good tape maintenance.

Because the focus of this survey was narrow, the data do not provide us with insights into the organizational problems of the data centers, staff allocations, demands for services, and the current budgetary situation, all of which probably influence the quality of the maintenance practices. The increasing reliance on statistical evidence for research, policy, and program planning, and the influence of libraries generally in the information transfer process, suggest that further examination of the data center would be useful. This small survey of tape maintenance practices should be followed by a more extensive survey of the data centers, to reveal more fully how they facilitate the flow of information and contribute to intellectual inquiry. Current national funding priorities promote too centralized, too structured, and too hierarchical use of data repositories. This policy risks paralysis of the larger system and denies the pluralistic nature of information needs and services. Local data centers are important contributors in a pluralistic system. Their efforts in the areas of dissemination and maintenance of valuable archival data resources need to be fostered.

Endnotes

1. I mean conventional mass storage media, such as magnetic tape and disc; existing new storage systems, such as the IBM Photostore and SDC TBM II; extensions of current magnetic tape technology, such as the Calcomp Automatic Tape Library, IBM 3850, CDC 385000 System, Precision Instruments (OMEX) System 190.

2. Potential mass storage developments (based on other technologies) include the OMEX Vidicon System, video disk, direct digital film-based storage, holographic storage, electron beam memories. Volz comments that "there will be a number of new mass storage devices to reach the marketplace over the next two to three years. However, the immediate concern of the developers of these devices is to achieve a high recording density of large system capacity without extensive consideration for longevity of the media. This means that while some of the techniques do have some potential for archival purposes, many of the first applications are likely to be for large volumes of data which are non-archival, that is, data which can be safely discarded after a few years. Once adequate recording densities and access time are achieved, attention will be more focused toward the archival properties of the media.... Existing mass storage devices are not truly adequate for the archival (sic) of large collections of data and the most imminent new technologies will probably also not be acceptable. A really good solution for large data collection archival (sic) is still a number of years down the road and good higher level software support is still further away." For a description of holographic storage, see Maugh.

The Public Archives of Canada have been investigating the technology of recording data on special discs by exposure to focussed laser light--the video disk.

Locke states that this "technology has recently reached the point of sufficient technical maturity such that it should be seriously considered as a basis for the storage of archival materials." He goes on to say that "laser recording provides the only economical basis for large-scale, machine-readable storage. In addition, data recorded by laser are expected to exhibit longer lifetimes and better security than data stored by any other information storage process. What is more, archival materials converted to digitally coded laser records can be preserved forever without any degradation whatsoever by the simple process of periodic replication protected by error-correction coding."

In conversations with the author, Harold Naugler, director of the Machine Readable Archives of Canada, noted that it would probably be some years before production versions of these recording devices are on the market and have been tested. Volz comments that the "hardware to perform both recording and playback is projected to be in the vicinity of $200,000 for a trillion bit storage. However, devices that only read are expected to be available for just a few thousand dollars. Commercial marketing of the device is probably two years away." The high cost of the device certainly puts it beyond the capital equipment acquisition budget of every data archive and probably most computer centers. These devices also must have a write capability to be of any utility to the archive which has as a major function the dissemination of its collection.

3. Another problem may be recording technique. Prior to 1600- and 6250-BPI with phase and block-coded recording, the recording techniques for data did not have the capability of correcting for errors caused by minor flows in the magnetic survace of the tape. In addition, newer tapes have a much smoother surface than those manufactured in the late 1960s. As a result, there is less wear on the read-write heads, less wear on the tape, and less likelihood of debris accumulating on the tape, according to Volz.

4. As will be described later, a few of the identified data centers disseminate only ICPSR data. It also turns out that disqualifying ICPSR member institutions in our sample resulted in eliminating a few data centers that could have participated in our survey.

5. Twelve are classified as independent departments or organizations; three, affiliated with a teaching department; ten, affiliated with a research organization; three, affiliated with a computer center; and two, affiliated with a library.

6. The histogram excludes one data center established in 1941 and two centers which could not supply this information.

7. At this point it is useful to describe the computer hardware at these data centers: IBM accounts for 47% (N=14), Amdahl, 7% (N=2), CDC, 27% (N=8), DEC 10, 3% (N=1), DEX-VAX, 3% (N=1), ITEL AS6, 7% (N=2), Xerox, 3% (N=1), and Univac, 3% (N=1).

8. The question about the type of tape storage problem was left open-ended; as a result, totals exceed number of centers which responded that they had problems.

9. Almost 30% of the data centers noted that they are storing their master files on-site, and 44% both on- and off-site. These high statistics are cause for some degree of concern for long-term data preservation.

10. Respondents were asked to check any of the following which applied: collection too large, financial support inadequate, or not enough staff.

References

Boruch, R.F. and Wortman, P.M. "An Illustrative Project on Secondary Analysis." Secondary Analysis. San Francisco: Jossey-Bass, Inc., 1978, 89-110.

Dollar, C.M. "Problems of Magnetic Recording in Archival Storage." Digest of Papers. Spring Compcon 1977, the 14th IEEEC Society International Conference, San Francisco, 28 February-3 March 1977, 28-30.

Geller, S.B. "Layaway, Standby and Reactivation Procedures for Computer Magnetic Media." (N.D.)

Hofferbert, R.I. and Clubb, J.M. "Introduction." American Behavioral Scientist, 19(4), 381-386.

Locke, J.W. "Videodisc Pilot Project Progress Report: Phase I (8Sep78 to 31Mar79)." Prepared July 1979 for the Public Archives of Canada (mimeo).

Maugh, T.H. "Holographic File: An Industry on the Verge of Birth." Science, 201, 431-432.

Miller, W.E. "The Less Obvious Functions of Archiving Survey Research Data." American Behavioral Scientist, 19(4), 409-418.

Nesvold, B.A. "Instructional Applications of Data Archive Resources." American Behavioral Scientist, 19(4), 455-466.

Robbin, A. "Technical Guidelines for Preparing and Documenting Statistical Data." In Boruch, R.F., Wortman, P.M., and Cordray, D. (Eds.), Secondary Analysis; Policy and Practice in Applied Social Research. San Francisco: Jossey-Bass, Inc. (Forthcoming)

Rockwell, R. Personal Communication, 20 December 1979.

Rokkan, S. "Data Services in Western Europe: Reflections on Variations in the Conditions of Academic Institution-Building." American Behavioral Scientist, 19(4), 443-454.

Volz, R.V. "Computer Based Mass Storage Technology." Prepared for the Conference on Archival Management of Machine-Readable Records, Ann Arbor, Michigan, 7-10 February 1979.