Problems and procedures for preservation and dissemination of computer-readable data

Dollar, Charles M.

Charles M. Dollar

Problems and Procedures for Preservation and Dissemination of Computer-Readable Data

Introduction

Since much of the social science data, especially in the United States, dealing with the 1960s and 1970s is computer based, the preservation and accessibility computer-readable data demands the immediate attention of social science researchers. If there is to be significant social science policy research in the future on the 1960s, 1970s, and 1980s the preservation and assessibility of the computer-readable data so vital to this research must receive adequate attention now.

In this context, therefore, this paper treats the problems of preservation, storage, and access to computer-readable processes-produced data from the vantage point of the National Archives and Records Service. Part one sketches an historical overview of the experience of the National Archives and Records Service in this area. Included is a discussion of the application of computer processing information technology to the creation, use, and storage of Federal records since 1951, and how the National Archives and Records Service has responded to this new technology. Part Two examines certain problems, policies, and procedures in the preservation and access of computer-readable data.

Computer Technology

The development of computers in the United States really began during World War II when a crude electromechanical computer called MARK I was developed to generate firing tables for the U.S. Navy. Toward the end of the war the Electronic Numerical Integrator and Computer (ENIAC), which was the first electronic general purpose computer, went into operation at the University of Pennsylvania. After World War II the U.S. Army and the Manhattan Engineering District (which had the responsibility for developing the atomic bomb) supported the development of computers. Also, the Office of Naval Research formed a computer section to study the impact of computers on science and technology. During this same period the National Bureau of Standards developed a computer called SEAC for Standard Eastern Computer. However, the first commercial development of a general purpose computer was the result primarily of the work of J. Pressley Eckert and John Mauchly who had developed ENIAC at the University of Pennsylvania. They fomed a company called the Eckert-Mauchly Computer Corporation (later acquired by Sperry-Rand) and secured a contract from the Bureau of the Census to produce a general purpose computer[1]. This computer, which was called UNIVAC for Universal Automatic Computer was installed and used to process the census returns of 1950.

Even though by today's standards UNIVAC is quite crude, at that time its usefulness in storing and manipulating enormous amounts of information was quickly recognized. In 1952 there were five computers in the Federal government. During the next decade this number grew to 1,006. By 1977 the total number of general purpose digital computers being used by the Federal government exceeded 9,648. It is estimated that the annual cost of computer hardware alone is between six and ten billion dollars.

More important than the number of computers is the amount of information created and processed by them. Currently, Federal agencies use between 8,000,000 and 9,000,000 reels of computer tape (2400 feet long) to store and process information[2]. Furthermore, approximately 40 per cent of the information processed by Federal agencies is computer based and is increasing each year.

Almost every Federal agency now uses computers to generate, store, and manage information. While much of this information is routine bookkeeping of little interest to social science researchers, a significant amount does provide basic research materials. For example, between 1963 and 1972 the Department of Defense supported a number of computerized information systems that analyzed a wide variety of data regarding the United States war effort in South Vietnam. Included in these systems are data used to measure the effectiveness of the army's pacification program; to provide detailed information on terrorist activities; to identify patterns of North Vietnamese and Viet Cong activity and to serve as an intelligence base for military decisions; and to assess the ecological, physiological, economic, social, and military effects of the herbicide program[3]. More than thirty files comprise this unusual data

[1] Shelton, Bill, and Duncan, Joseph W., Revolution in U. S. Government Statistics, 1926 –1976 (draft manuscript from the Executive Office of the President, Office of Management and Budget), Ch. 4.

[2] Magnetic tape is the primary storage medium for computer-readable records. Punch cards and paper tape seldom are used for storage. Disc storage devices are used principally for records that require virtually on-line access.

[3] For more information see Catalog of Machine-Readable Records in the National Archives of the United States (1977).

base. Other Federal agencies that create substantial computer-based information include the Internal Revenue Service, the Department of State, the Bureau of the Census, the Bureau of Labor Statistics, the Department of Agriculture, the Department of Health, Education, and Welfare, the Environmental Protection Agency, and the Energy Research Development Administration, to name only a few.

One interesting aspect of how Federal agencies use computers is management of information systems. In the Department of State a system now in operation annually receives and stores by computer more than 500,000 cables from U. S. embassies and consulates. The messages are indexed by computer and assigned a reference number. Since they are arranged in the order they are received, easy access to a particular cable or cables on a specific subject is possible only through an on-line computerized index. Equally interesting is the computerized information system the Watergate Special Prosecution Force used in conducting its investigation of the so-called Watergate affair. This system contains abstracts of testimony before the Senate Select Committee (the Ervin Committee), the Grand Jury, and office interviews. Some seven data items such as name, date, type of incident and the like were flagged so that investigators could retrieve abstracts meeting certain criteria. For instance, all of the abstracts of an individual's testimony or all of the testimony in which a single person or a combination of people was named could be retrieved. This system could be refined even more by restricting retrieval to a certain time period or a certain event.

The National Archives and Computer Technology

This brief sketch of the growth of the application of computer information processing technology to the creation, use, and storage of records within Federal agencies is incomplete without a review of the National Archives' activity in this area. Until the early 1960s the National Archives was doing very little. Indeed, the prevailing view was that IBM punch cards, then the primary storage medium, were non-record and redundant since tabulations derived from them existed in printed form. This accounts for a decision in July 1936 to dispose of some eight million cards which contained information from the Census of 1930.

Happily, this view underwent considerable modification in the 1960s. In 1963 the Social Science Research Concil, concerned about the increasing quantity of social science data in machine-readable form, appointed a Committee on the Preservation and Use of Economic Data. This group registered its concern about the preservation of economic data stored on punch cards and magnetic tapes in Federal agencies. At about the same time several people at the National Archives became quite concerned about the status of machine-readable records in Federal agencies. One of them, Meyer Fishbein, urged that key persons from the National Archives meet with the SSRC committee[4].

One result of the discussions was an examination of the possibility of the National Archives establishing a data center to handle the growing volume of machine-readable records. This undertaking coincided with a survey of statistical data on punch cards and magnetic tapes in 13 selected Federal agencies conducted jointly by the Bureau of the Budget and the National Archives. The study, which was completed in 1964, concluded that it was impractical to consider establishment of a Federal data center. Instead, a far more urgent problem requiring immediate attention was that of bringing the disposition practices of Federal agencies regarding machine-readable records under Federal regulation. The survey of the 13 selected agencies demonstrated that current policies and practices did not ensure retention of valuable data either for government use or scholarly research. Despite this less than enthusiastic response of the National Archives to the concept of a federal data center, the Social Science Research Council approached the Bureau of the Budget immediately about establishing a Federal Data Center. This proposal soon became entwined in the issue of computers and the invasion of privacy and was dropped by the Bureau of the Budget.

However, the National Archives' concern about proper disposition practices for machine-readable records did not abate. Archivist of the United States, Robert Bahmer, established a committee to undertake a detailed study of machine-readable records in Federal agencies. The Committee's report in 1968 included a number of recommendations, one of which called for the establishment of a special organizational unit to deal with machine-readable records. In 1969 the new Archivist of the United States, James B. Rhoads, created a data archives branch or staff to complete a survey of magnetic tapes in Federal agencies, developed an inventory of all such files, and tapes containing information of possible long term value.

Since 1969 the Data Archives Staff, now called the Machine-Readable Archives Division, has accomplished much. A major inventory of machine-readable tapes has been completed and updated. Some 2,000 reals of magnetic tape containing records of long term value have been accessioned into the National Archives[5]. A reference service that includes copying tapes and providing special purpose extracts from multi-reel files has grown to the point that more than 1,200 tapes were copied for a fee in this fiscal year. A special tape storage vault with controlled environmental conditions was completed. General Records Schedule Number 20, which identifies categories of disposable and non-disposable records, was developed and substantially revised last year. Federal regulations regarding proper storage and maintenance of machine-readable records have been greatly strengthened. And in April of 1977 the Archivist of the United States authorized the Machine-Readable Archives Division to receive machine-readable records which are of high current interest but whose long-term value is uncertain. The vehicle for accomplishing this mission is the Cen-

[4] Data Archives Staff Report, August 1970, Washington/DC 1970, p. 1.
[5] Catalog of Machine-Readable Records in the National Archives of the United States.

ter for Machine-Readable Records. Discussions with several Federal agencies that produce machine-readable records of interest to social science researchers are now underway that could result in the Center serving as a kind of clearing house for information regarding which Federal agency has particular files. Furthermore, the Center will, where possible, seek from Federal agencies machine-readable records that are of interest to the social science research community and make them available. In the long run this could be one of the more significant activities of the Center for Machine-Readable Records.

While much still remains to be done, it is clear that the National Archives has established a program whereby a wide variety of computer-readable records of the Federal government will be available to researchers in the present and the future. Certainly this is no mean accomplishment, and it documents the commitment of the National Archives to the preservation and dissemination of computer-readable records.

Appraisal of Computer-Readable Records

Since the National Archives lacks the resources to preserve *all* of the computer-readable records that Federal agencies produce, the process of selecting those computer tape files to be preserved is important. The experience of the National Archives for the last twenty-five years or so in dealing with textual or printed records for Federal agencies is that only about three per cent of them eventually are preserved. If this rule of thumb is applied to computer-readable records then the national Archives faces the prospect of preserving between 240,000 and 270,000 reels of computer tapes as of 1977[6].

The key problem here is how to identify this three per cent or 240,000 to 270,000 reels. The Machine-Readable Archives Division has developed a records schedule that identifies categories of disposable and non-disposable computer tape files[7]. Through this schedule the Archivist of the United States has delegated authority to Federal agencies to destroy all computer-readable records that are classified as disposable by the record schedule. In most instances, „processing files“ which

6 This percentage may be too low. Lionell Bell has suggested that more computer-readable records today will be preserved vis-a-vis preservation of comparable textual records because the computer greatly extends the power of both administrators and users to handle large volumes of data. See The Archival Implications of Machine-Readable Records, VII International Congress of Archives (Washington/DC, September 27–October 1, 1976).

7 The official title of this schedule is „General Records Schedule 20. Machine-Readable Records“.

range from data input to update transactions, are disposable without regard to subject matter. Non-disposable records which must be offered to the National Archives are „master files" that constitute the definitive state of a data file in a system at a given time. General Records Schedule Number 20 identifies several classes of disposable „processing files" and non-disposable „master" files. Generally, routine housekeeping files are „master files" that are disposable because of the trivial information they contain. An example of a disposable housekeeping file would be an inventory of army vehicles or airplane parts. On the other hand a „master file" containing information relevant to social science research in its broadest meaning would be preserved. Statistical and scientific „master files" must be offered to the National Archives.

The fact that a file is not disposable under General Records Schedule Number 20 does not automatically mean the National Archives will preserve it. A number of stringent criteria have seen developed that must be met before a file is accessioned. These criteria are summarized in the appraisal decision table for ADP records in Figure 1.[8]

The process of applying these criteria goes something like this. After the tape reaches the National Archives, we deal initially with what we call technical considerations, the first of which is the adequacy of the documentation. Federal regulations specify that tape files transferred to the National Archives must be accompanied by documentation that includes agency prepared technical memoranda relating specifically to the file, a record layout, a codebook, publications derived from the file, and two completed General Services Administration tape inventory forms . If absolutely essential documentation such as a codebook or a record layout is missing and can not be reconstructed, the appraisal terminates and the tape is returned to the agency for disposal.

Usually, the essential documentation is intact and the tape is then physically checked for readability, which is the second phase of the technical considerations. This means mounting the tape on a drive and reading it. Sometimes temporary read errors are encountered that can be eliminated by passing the tape over a tape cleaner or mounting the tape on another drive. When permanent read errors are encountered, the decision about readability depends both upon the scope and magnitude of the errors as well as the basic value of the file. For example, a few unreadable blocks do not seriously diminish the technical quality of the file. On the other hand, if more than five percent of the blocks are unreadable, then the file would be considered unreadable[9]. However, no hard and fast rule is applied since even an incomplete file of major substantive importance still could be very valuable.

[8] Figure 1 is reproduced at pp. 470–471.

[9] This varies, depending upon record length, block size, and the pattern of error distribution. The same error in every block would be handled differently than random errors. Also, the importance of the records themselves would be another consideration. One way to handle permanent errors is to delete the block from the file and print it out for inclusion in the introduction to documentation that is prepared.

At the same time the readability of the file is determined, we also obtain a record count and a five block dump and create a standardized 1600 bpi archival copy and a reference copy. (See step 7 in the decision table.) These two copies are placed in our vault while the archival qualities of the file are evaluated.

The major archival consideration is the legal, evidential, or informational value of records (see step 9). Few computer-readable records of Federal agencies impinge on the legal rights of citizens and the Federal government. The evidential value of records refers to documenting significant agency policy decisions and the agency's mission accomplishments. Although computer-readable records seldom provide this documentation, the manner in which data is collected and used can reveal a great deal about an agency's perception of its missions.

The concept of informational value refers to the residual value of records after agency needs have been satisfied. Or to put it another way, the value of such records is that the information they contain can be analyzed in ways and for purposes other than those for which the agency originally collected the information. A trite but cogent example is the U. S. decennial census which is conducted every ten years (every five years beginning in 1985) in order to count the population and to summarize trends about the social and economic well-being of citizens. For all practical purposes, once the Bureau of the Census publishes its multi-volume report, the information has little value for agency purposes. However, to social scientists and genealogists, and other researchers, the value of these records lies in the information they contain about particular persons, groups, situations, events, and the like.

The informational value of computer-readable records is proportional to their level of aggregation. For example, a summary of census data at the enumeration district level is far more valuable to researchers than a summary of county level census data. Correspondingly, census information at the household level is more valuable than a summary at the enumeration district level. This rule is that while you can never disaggregate summarized data (reduce grouped data to individual data) you can always aggregate individual level data to the desired summary level.

A file's potential for linkage with other data is another consideration of informational value in determining whether the National Archives will preserve it. Usually, records arranged at the lowest reporting unit (individual person or individual business firm) have considerable linkage potential. Common attributes such as place of residence, occupation, age, and sex (if they share similar codes) permit the linkage of groups with similar attributes.

An assessment of the importance of the subject matter the records cover is as important as an evaluation of a file's potential for further processing and data linkage. Subject matter importance is defined in terms both of the interests and concerns of researchers today as well as the anticipated interests and concerns of researchers fifty years from now. Obviously, this kind of evaluation requires an understanding of a wide veriety of research trends. Also, it involves considerable „luck" since the accurate prediction of future social science research trends (and by definition the data required to support this research) is at best an educated guess.

If the decisions in steps 8, 10 and 11 of the decision table are affirmative, then attention turns to data validation. This step involves a manual comparison of the codebook and record layout specifications with the five block dump mentioned earlier. If this comparison reveals any inconsistencies, values not noted in the codebook, or missing data they are noted and included in the written appraisal report[10].

Data validation also involves consideration of the reliability and validity of the data. Since records of informational value probably will be used in ways and for purposes other than those for which the agency collected the data, careful attention is paid to possible biases. This is particularly true of data collected for regulatory purposes. Frequently, data validation will reveal the existence of data imputation where estimates have been substituted for missing responses or incorrent figures. Generally, data imputation occurs when consistency checks are made during edit runs prior to creation of a master file. Unfortunately, there is no simple or inexpensive way to identify where specific data imputation has occured, although the overall process itself can be documented. An even more complex data imputation problem that data validation may uncover is the extent to which estimates have been made in working with disparate data to accord with appropriate definitions, to fill gaps in coverage, and to obtain statistical comparability among geographical units over time. Given the way this kind of data imputation occurs in, say, the Bureau of Economic Analysis of the United States Department of Commerce in preparing estimates of personal income for local areas, there is no audit trail, so to speak, that can be followed. Only the broad outlines can be discerned as inquiries are made about how the file was created.

It is important to note that the process of data validation involves no corrections or cleaning of the data. The responsibility of an archivist is to record for potential users all of the deficiencies and limitations learned about a particular file while in the process of assessing its long term value. It is the user's responsibility to determine the extent to which the reliability and validity of the data in question meet his research goals.

Even though at this point in the decision table (step # 22) a file may be acceptable, arrangement and accessibility of the data along with estimated preservation costs must be weighed before recommending accession into the National Archives.

[10] Typically, this requires 10 to 15 hours for a file with few problems and includes preparing an introduction to the documentation. This introduction contains an evaluation of the technical quality of the file. Incorrect data codes and missing values are noted. If permanent read errors were encountered, then a printout of that block(s) is included along with a statement as to the probable cause. There are automated data verification programs, but the amount of time required to prepare the input data when the codebook is not machine-readable can be extensive. While manual data validation can be tedious, it does have the virtue of giving archivists a „feel" for the data. For a useful discussion of checking for errors see Roistacher, Richard A., A General Consistency Check for Machine-Readable Data, in: Sociological Methods and Research, Vol. 4, No. 3 (February 1976), pp. 301–320, and Hammer, Michael, Error Detection in Data Base Systems, in: Proceedings of the National Computer Conference 1976, Vol. 45, pp. 795–801.

Arrangement refers to the internal data structure of the file while accessibility refers to whether or not the file is software dependent or is in some non-standard character code (such as XS—3 that the Bureau of the Census uses).

Increasingly, Federal agencies are employing data base management systems such as NIPS and System 2000 and special purpose statistical analysis packages such as SPSS and OSIRIS, to name only a few. From the viewpoint of the National Archives a file embedded in any of these systems or packages is software dependent in that it can only be processed in a computing environment that supports the system. This is not a major problem with regard to SPSS or Osiris; however, it is with regard to NIPS, which IBM no longer supports, and System 2000, which is proprietary. Since the policy of the National Archives is to preserve software independent files, this means that conversation is necessary. Our experience is that such conversion is expensive. For instance, the military files dealing with South Vietnam (discussed earlier) were embedded in a NIPS format then transferred. Since only those users who had access to a computer system that had NIPS could process the files, we concluded that the value of the files was sufficient to justify an expenditure on the order of $ 400 per reel.

The problem with arrangement is less critical in that usually for ease of processing variable length records are formatted as fixed length records. The result is considerable padding with zeroes that expands the physical size of the file. Since storage space is at a premium, data must be compacted as much as possible in order to reduce the number of reels in storage. This compaction also can work to the advantage of users who may have to pay for only one reel of tape rather than two reels.

The cost of preserving computer-readable records with existing technology is such that it can not be ignored in determining whether or not a file will be assessioned into the National Archives. Therefore, we must exercise even greater care in the future when selecting files to be assessioned. Naturally, this will encourage a conservatism that can result in preserving only files for which high research interest is demonstrable. Even though it is too early to predict its full impact on the National Archives, I do think that in the long run it will not be possible to accession the wide variety of information implicit in the concept of process-produced data. It is possible, given this constraint of preservation costs, that fifty years from now the computer-readable holdings of the National Archives will serve only a handful of researchers whose interests coincide with the narrow scope of our holdings. In effect, therefore, the exercise of such care and caution in selecting records to be assessioned could yield a very undesirable outcome; the preservation of files that may be at best of marginal value fifty years from now and the destruction of files that fifty years from now would be considered very valuable. It is conceivable that serendipitous data bases compatible with the questions of researchers in 2027 will be scarce, given the care and caution with which the National Archives must operate in this area.

Presently, the National Archives preserves computer-readable records on magnetic tape even though it is not accepted as a permanent archives storage medium. Computer tape on which magnetic signals have been recorded is a highly fragile storage medium that is vulnerable to powerful magnetic fields and can deteriorate rapidly in an unstable environment[11]. Tapes written at 1600 bpi and stored under optimum conditions should be reliable for ten years. However, in order to achieve these optimum conditions it is necessary to eliminate the major causes of deterioration of computer tape. These are exposure to radical changes in environmental conditions, dirty tapes, and fluctuations in tape tension in the rewind mode.

The National Archives maintains a fireproof vault in which some 20,000 reels of tape can be stored at a temperature of about 68° Fahrenheit and 50 per cent humidity[12]. A hygrothermograph in the vault monitors continuously the temperature and humidity. A backup cool air unit is activated if the temperature in the vault becomes excessive.

Dirty tapes occur when debris is deposited during the manufacturing process or when they are passed over dirty read/write heads. This problem is dealt with in part by insisting that computer operators clean the read/write heads regularly and write only on new certified tapes that have been cleaned. The latter is particularly important since the manufacturing process tends to leave debris on the tape surface that can cause temporary read/write errors. Therefore, every tape is passed over a tape evaluator/cleaner before being used. Preservation tapes are passed over a tape cleaner again after being written on to remove any dirt deposited by the read/write heads.

Since tension in the stacked (or rewound) tapes layers is a critical factor in long-term storage, balanced tension while rewinding is important[13]. Unfortunately, the tape drives now available to the National Archives lack this feature. The „read backward" mode could be used, but it would double the computer cost. A relatively inexpensive solution is at hand since the tape cleaner referred to above rewinds under controlled tension.

These optimum storage conditions are supplemented through writing an archival

[11] For an excellent discussion of tape vulnerability to magnetic fields see Geller, Sidney, Erasing Myths About Magnetic Media, in: Datamation (March 1976), pp. 65–70.

[12] A 2400 foot reel of magnetic tape tends to change total length about one foot for every 20° Fahrenheit change in temperature or every ten percent change in relative humidity. These effects are independent and can occur concurrently.

[13] Too much tension causes the tape to assume a permanent stretch or curvature which can damage the tape surface on the first read attempt or can cause uneven winding on the second rewind. Too little tension allows some layers to separate at some point in the stack and in some cases the layers will slip or fold into permanent creases that can not be read. Thus, it is very important that the tape drive not introduce sharp variation in tension as the tape is rewound.

master copy tape and a reference copy tape at 1600 bpi in EBCDIC. In addition, each copy is verified as error free. Both the master copy and the reference are placed in plastic cannisters and stored upright in tape racks. The reference tape copy is removed from the vault as needed to process reference requests. The master tape copy leaves the vault only for a periodic check of readability.

A three percent statistical sample of tapes is tested each year for readability. This test consists of mounting the tapes on a tape drive and reading them. It is likely that most read problems will be temporary read errors which can be corrected by passing the tape over a tape cleaner. If a tape which is part of a multi-reel file has more than five temporary errors the entire file is checked. Tapes with a significant number of temporary or one or more permanent errors are copied onto new tapes and verified as error free. Plans now call for copying every tape onto a new tape after ten years of storage. These procedures cost less than rewinding tapes every six months. Also, the use of backup tapes (master copy and reference copy) provides an added dimension of protection.

Even though the National Archives and Records Service does not recognize magnetic tape as a permanent archival storage medium research, development now underway suggests that the time is near when permanent storage medium for computer-readable records will be available. While no firm position has been taken, several criteria have been developed that could be used in assessing the capability of mass storage technology to meet the Archives' needs for permanent storage medium.

An acceptable permanent storage medium for computer-readable records in the National Archives should meet several criteria. Absolutely essential is a storage mode in which under typical conditions, the recording signal will not degrade over time. This requires a non-erasible or read only capability. To achieve this criteria the recording signal must affect an unalterable change in the storage medium. A second criterion is a non-volatile storage medium that could be treated much the same as ordinary library books. Maintenance costs, such as that associated with a controlled environment and periodic refreshing of the storage medium, would not be necessary. A third criterion is immediate verification that an error free copy has been made. The importance of DRAW (Direct Read After Write) is significant when a large volume of records is processed and it is too costly or otherwise impractical to verify an error-free copy later. While it is expected that most computer-readable records in archives storage may not be used often, it is most desirable to have a storage medium that can be read repeatedly with little or no degradation in either the signal or the medium. A fourth criterion, therefore, is that the read capability must not involve physical contact with the storage medium like that of tapes passing over read/write heads. High packing density at a low cost is the fifth criterion. Closely related to high packing density is a high data transfer rate, which is the sixth and last criterion. A data transfer rate on the order of one megabyte per second is necessary for efficient processing of voluminous files (some of the multi-reel files in the National Archives contain as many as eighteen 1600 bpi reels). The significance of a high data transfer rate becomes increasingly critical as the volume of computer-readable records grows.

Currently, magnetic recording is the principal means of storing computer-readable records. For more than twentyfive years magnetic technology has dominated the storage of digital data[14]. Magnetic tape offers low cost, direct read after write, erasability, and a high density, especially 6250 bpi tape. Magnetic disks, on the other hand, permit random access, along with some of the other features of magnetic tapes. Nevertheless, magnetic recording has serious deficiencies in archival storage, not withstanding a number of significant developments in the last year or so[15]. Low maintenance, non-erasability, and no physical contact with the medium are almost impossible to achieve magnetically. The one exception to this is magnetic bubbles which do not involve physical contact with the medium while reading or writing.

Recent developments in non-magnetic storage media research suggest that optical recording offers the greatest potential for meeting archival storage requirements for machine-readable records. While several optical recording devises and approaches are still in the development stage, the technology generally consists of a powerful light beam that can be modulated to produce a stream of bright and dark spots representing the 1's and 0's of digital data, a lens or objective to focus the beam, and a sensitive storage medium. A laser which can operate at about 10^6 pulses per sesond with the power of one watt is sufficient. Optic lenses that can focus a light beam down to a dimension of less than one micron of surface area are available. This corresponds to 6 x 108^8 bits per square inch ot the equivalent of one fully packed 2400 foot real of magnetic tape written at 6250 bpi. Storage media include photographic emulsions and film with a thin reflecting coating of metal[16].

[14] Freeman, David N., The New Mass Storage Systems, INFO 76 Conference, Chicago 1976.

[15] Dollar, Charles, Problems of Magnetic Recording in Archival Storage, in: Digest of Papers, Spring COMPCON 77, IEEE Annual Meeting, San Francisco 1977, pp. 28–80.

[16] Chen, Di, and Zook, David J., An Overview of Optical Data Storage Technology, in: Proceedings of the IEEE, Vol. 63, No. 8 (August 1975), pp. 1207–1212; Gaylord, Thomas K., Optical Memories: Filling the Storage Gap, in: Optical Spectra (June 1974), pp. 29–34; A Report on the Archiving of Binary Computer Data on Microfilm, Commissioned by the United Kingdom Central Computer Agency, in: Informational International, Inc. (February 1977); Kaczorowski, Edward M., Optical Mass Storage, Digest of Papers, Spring COMPCON 77, San Francisco 1977, pp. 33–35; Kenney, G., et al., An Experimental Optical Disc Data Record, in: Technical Note, No. 144, Philips Laboratories (July 1977).

The National Archives and Records Service preserves computer-readable records in order to make them accessible to users – now and in the future. Accordingly, the policy of the National Archives and Records Service is to provide the widest practical access to computer-readable records that is consistent with statutory and regulatory constraints and available financial resources.

A unit within the Machine-Readable Archives Division has the responsibility for providing reference service that include preparation of a catalog that describes our holdings in non-technical terms. The 1977 edition of the *Catalog of Machine-Readable Records in the National Archives of the United States* lists more than 125 entries that range from a public opinion survey on population growth and the future to a metropolitan Washington area transportation study. Identification of these entries is keyed to the concept of record groups rather than subject matter. As the volume of our holdings increases we expect to construct a machine-readable subject index to each entry or file in order to facilitate user access to data relating to specific subject areas. Our current practice of describing entries probably will be adequate for several more years or until the volume of our holdings makes it unwieldy.

Two general types of reference service are provided: copies and extracts. Copy reference work includes card-to-card, card-to-tape, tape-to-tape, and tape-to-printout, although tape-to-tape copying makes up about 90 percent of the work. A policy of providing data to as many users as practical in a format that conforms to the computer hardware requirements where the data will be processed means that a user may request a seven or nine track tape (BCD or EBCDIC), with or without internal labels, and written at a density of 556, 800, or 1600 bpi. A complete documentation package for each file is available in electrostatic copies or 16 mm microfilm. Approximately 75 percent of our reference work consists of tape copying.

Generally, extract work is done for a user when only selected portions of a multi-reel file are needed. A special purpose program is written according to user specifications and an output tape with a modified documentation package is provided to users.

The Machine-Readable Archives Division does not provide special software service. Indeed, our preference is to have tape files software independent so that many more users will have access to a file. I should hasten to add that most of our files are formatted so that they can be accessed by SPSS or OSIRIS after being prepared for data input. Sometimes we accept as reference tools special software programs written for certain files. We will make these programs available to users but assume no responsibility for maintaining them. Our limited resources simply do not permit involvment in all of the problems of supporting special purpose software.

Users must pay for the services received. A tape-to-tape copy without regard to density, character set, or labels costs $65 per output reel plus documentation. The

Figure 1:
Appraisal of ADP Records

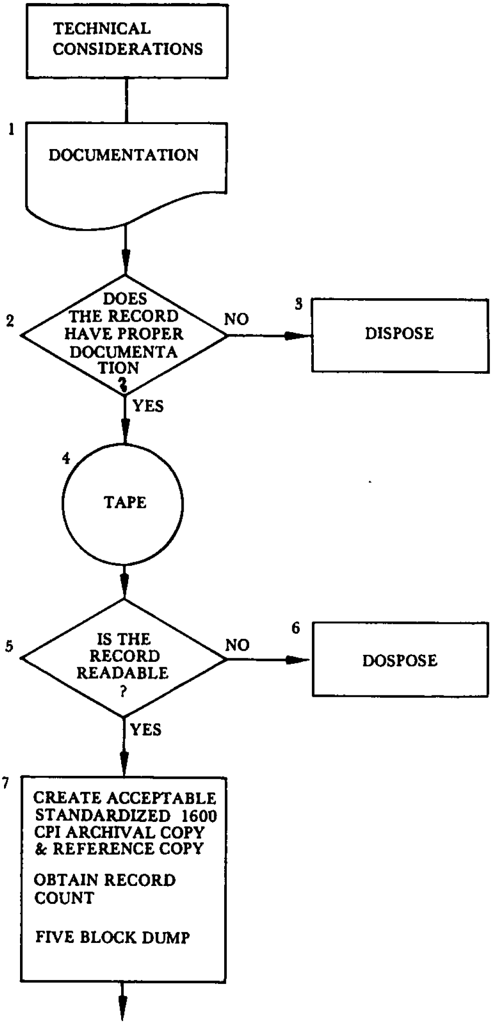
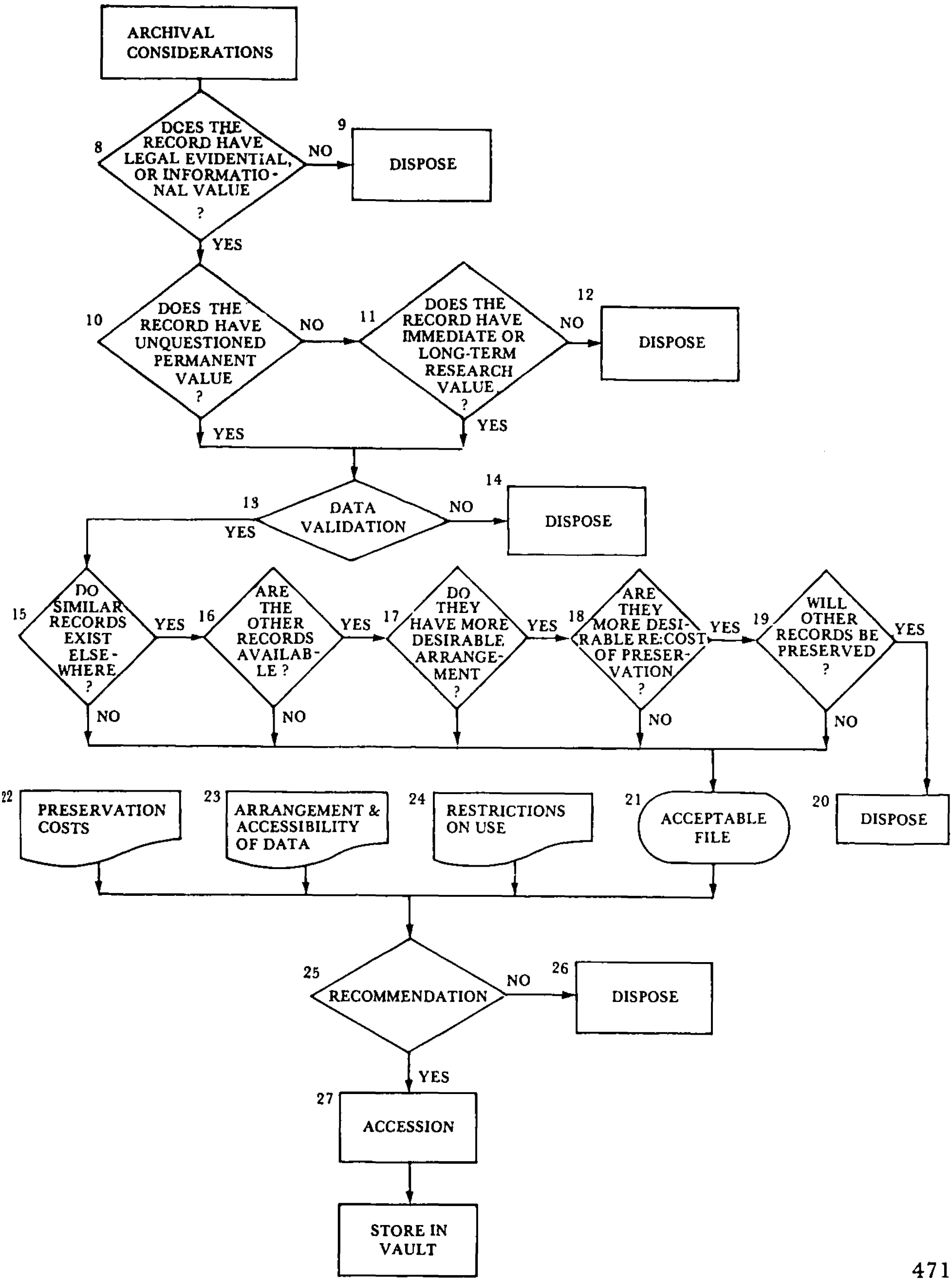


Figure 1:
continued



fee charged for extract work is $65 for each output reel plus $150 per hour of computer processing time used in extracting the desired information. If the desired output is a printout the only fee charged is $150 per hour of computer processing time. This is the lowest fee charged by any Federal agency that is required to recover costs incurred in disseminating computer-readable data.

Inevitably, when disemination of computer networking is advanced as the future mode of disseminating computer-readable records. There is no doubt that the technology now exists for computer networking. Some 160 universities, military installations, and the like in the U.S. and Europe are now part of the ARPA Network. A number of computerized data services provide online transcontinental computing service. The primary problem from our perspective is that data transmission rates are so low that it would be prohibitively expensive to users who desire several reels of tape. Thus far, our experience is that most users of the computer-readable records in the National Archives are satisfied with the delivery of tape by the United States Mail Service. Of course, another approach would be for users to be part of a computer network which could access a subject index to computer-readable records in the National Archives. If the documentation were computer-readable and on-line with the index, a user could peruse the two, select the items of information he wanted, and then key in a request which could be processed and the output returned by U. S. mail. Our experience has been that most of our users are not sophisticated enough in this area to benefit from such access. As users become more competent in on-line information retrieval and the cost of computer networking decreases, the National Archives and Records Service will play a role in computer networking. In the meantime, we will continue to use magnetic tapes and the United States Mail to meet the data needs of our users. At the same time and as our resources permit it, we will initiate action to make it less difficult and costly for the National Archives to become a viable part of such a network.

Conclusion

This paper has described now the National Archives and Records Service preserves and disseminates computer-readable records. A point worth repeating is that careful attention is given to the matter of identifying those files that merit preservation. The National Archives believes that considerable progress has been made in this area, especially with General Records Schedule 20. Nevertheless, both the preservation and dissemination of computer-readable records in the future will be influenced largely by technological developments. In this regard, the National Archives' experience thus far in dealing with computer-readable records is only a beginning. The experience of other national archives will provide a better picture of the magnitude of our collective problems and solutions.