Appraising Machine-Readable Records

CHARLES M. DOLLAR

MACHINE-READABLE RECORDS ARE DEFINED as records created for processing by a computer. While this definition encompasses a wide variety of storage media including punched cards, magnetic discs, cassettes, and paper tape, the vast majority of machine-readable records are stored on magnetic computer tape. It is reasonable to anticipate that, over the next decade, on-line storage and retrieval devices with random access capability will replace magnetic tape as the primary storage medium. This suggests the possibility that new computer storage technology will radically alter the appraisal of machine-readable records.

Current appraisal practices of machine-readable records differ in significant ways from those for textual records; and as computer technology progresses these differences will become even more pronounced. Indeed, it is likely that current practices and standards will be obsolete and irrelevant within a decade. It is quite costly to accession and preserve properly a single reel of computer tape. The proliferation of on-line data base management systems will make this process even more expensive, and costs will receive even greater consideration in appraisal decisions. As a result, the consequences of the rationalization (in the British sense) of the records retention process will become more evident, to the discomfort of archivists and researchers.

These possibilities suggest or imply a number of points that merit consideration. The standards and practices now employed in the Machine-Readable Archives Division of the National Archives and Records Service, with attention to the changes likely to occur within the next decade, provide the context for this consideration.[1]

Since 1969 the staff of the National Archives has appraised machine-readable records and thereby contributed to the refinement of certain concepts and criteria that comprise the present "state of the art," as it were. A delineation of the sequence of decisions involved in the appraisal of machine-readable records can convey the current state of the art.

The first decision—whether a file will be appraised—is left largely to agency records officers. A disposition schedule for machine-readable records identifies categories of disposable and non-disposable computer tape files.[2] The former, consisting of processing files that range from initial data input to update transactions, are automatically disposable without regard to subject matter. In most federal agencies at least 60 percent of computer tape files are disposable as "processing files." Most non-disposable files, which must be offered to the National

The author is director of the Machine-Readable Archives Division, National Archives and Records Service. His article is based on his paper delivered on October 5, 1977, at the annual meeting of the Society of American Archivists, in Salt Lake City.

[1] For an earlier study see Meyer Fishbein, "Appraising Information in Machine Language Form," *American Archivist* 35 (January 1972): 35–43.

[2] See General Records Schedule 20, Machine-Readable Records (GSA).

Archives, are so-called master files—the definitive state of a data file in a system at a given time. Although they are defined as non-disposable, master files are accessioned only if they meet a number of stringent criteria. (See Decision Table.) Following is a full discussion of only the more salient of these criteria.

When a tape file is offered to the National Archives, it must be accompanied by adequate technical documentation which at minimum consists of a record layout and a codebook. These are crucial since they provide the key to the exact location of each item of information on the tape and define the value the numeric characters represent. (Thus technical documentation may be seen as a "finding aid" at the item level.) If this essential documentation is missing or can not be reconstructed, the offer is rejected.

If it is intact, the tape is then physically checked for readability by mounting it on a tape drive to be read by a computer. Sometimes minute particles of dust or other material on a tape will prevent a computer from "reading" some portion of the tape. Usually, readability can be restored by passing the tape over a tape cleaner to eliminate these particles. Sometimes a tape can be physically damaged by a crimp or stretching of a portion of a tape so that magnetic signals cannot be read. When any portion of a tape cannot be read because of physical damage, a decision to proceed further depends on both the scope and the magnitude of the damage, as well as the basic value of the file.[3] (These problems are analogous to those involving deteriorating paper, film, or microfilm with poor resolution.) If the tape is readable, a record count and a partial printout are produced, as well as a duplicate copy of the tape for security storage, while the archival qualities of the file are evaluated.

As is the case for records in any form, the major archival consideration is the legal, evidential, and informational value of records. At the state level, it appears that some machine-readable records are of permanent legal and evidential value.[4] For example, in the State of Illinois, nineteenth-century land sales records, which clearly are of legal value, are being converted into machine-readable form. At the federal level, few computer-readable records impinge on legal rights or document significant agency decisions and programs accomplishments. Consequently, informational value is usually the basic concern. The concept of informational value refers to the residual value of records after agency needs and individual rights have been satisfied. To put it another way, the value of the records is such that the information can be analyzed in ways and for purposes other than those for which the agency originally collected them.

Generally, 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. Similarly, census information at the household level is more valuable than a summary at the enumeration district level. The rule is that while you can never disaggregate summarized data (down from group

[3] The decision will vary, depending upon record length, block size, and the pattern of error distribution. The same error in every block would be handled differently from random errors. If only a few blocks are unreadable, the value of the file is not seriously diminished. On the other hand, if more than 5 percent of the blocks are unreadable, in most instances the file would be rejected.

[4] Records with informational value are being created in a number of states. For example, in Vermont longitudinal data on wildlife population size, habitat, and migration patterns dating back to the 1930s is in machine-readable form.

data to individual data), you can always aggregate micro-level data to the desired summary level. Thus, unaggregated micro-level data has the greatest potential for further computer processing.

A file's potential for linkage with other data is another consideration of informational value.[5] Usually, records arranged at the lowest reporting unit (individual person or individual business firm) have considerable linkage potential. Common attributes (if they share similar codes) such as place of residence, occupation, sex, age, and the like permit the linkage of groups with similar attributes. Personal identifiers such as name and social security number permit even more sophisticated data linkage.

The evaluation of a file's potential for further processing and data linkage is merged with an assessment of the importance of its subject matter. This is approached, just as in the case of textual records, in terms of the interests and concerns both of current researchers and those working fifty years from now. Obviously, this kind of evaluation requires an understanding of a wide variety of research trends. It also involves considerable luck, since accurate prediction of future research trends is at best an educated guess. It is necessary, then, to turn to established researchers or special interest groups for guidance. For example, the National Archives requested a subcommittee of the Social Science Research Council to review some preliminary retention guidelines for certain Bureau of the Census tape files.

If a tape survives the test of informational value, attention is turned to data validation. This involves a manual comparison of the codebook and record layout specifications with the partial printout made as evaluation began. 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.

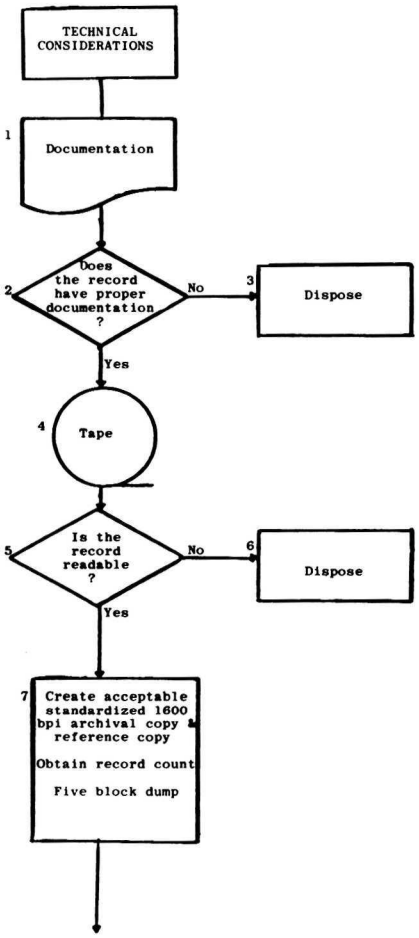
Data validation also involves consideration of the reliability and validity of the data. Since records of informational value probably will be used in ways and for purposes other than for which the agency collected the data, careful attention is paid to possible biases. This is particularly important with data collected for regulatory purposes. Frequently, data validation can reveal the existence of data imputation—cases in which estimates have been substituted for missing responses or incorrect figures. Unfortunately, in many instances there is no simple or inexpensive way to identify specific data imputation, even though the overall process can be documented.

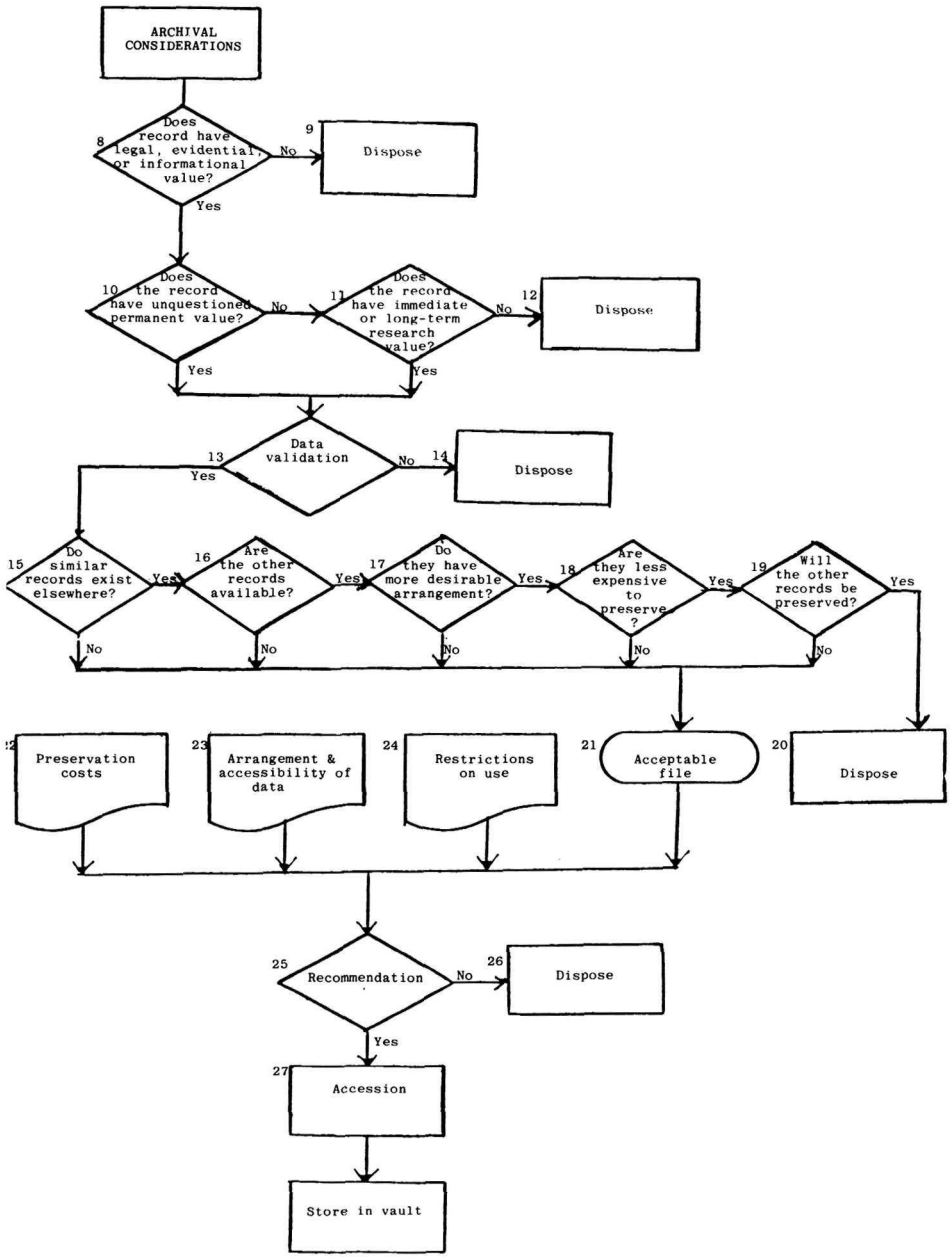
Even when a tape file satisfies all of the appraisal criteria discussed above, arrangement and accessibility of the data and estimated preservation costs must be weighed against its value before its accession into the National Archives is recommended. Arrangement refers to the internal data structure of the file, while accessibility refers to whether or not the file is software dependent or is in some non-standard character code.

From the viewpoint of the National Archives, a file dependent on any system or package is termed "software dependent"; that is, it can only be processed in a computing environment that supports the system. Increasingly, at all levels of government and in the private sector, data base management systems, informa-

[5] In a review essay, Myron Guttman notes that accumulating life histories of individuals is not the only way to study social mobility. He points out that linking groups with shared characteristics is a viable alternative. See "The Future of Record Linkage in History," *Journal of Family History* 2 (Summer 1977): 155.

APPRAISAL OF ADP RECORDS





tion retrieval systems, and special purpose statistical software packages are being used. Since the policy of the National Archives is to preserve files in a software independent state, this means that conversion is necessary. Such processing can be very time consuming and expensive.

While arrangement can involve a number of problems, some of which are quite technical, the most common problem occurs when, for ease of processing, related records of variable length have been formatted as fixed-length records. This is done by padding the shorter records with zeroes, padding which 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.

The mounting costs of accessioning and preserving machine-readable records cannot be ignored in appraisal decisions. Thus far the experience of the National Archives is that it costs approximately $360 in staff time, computer time, and supplies to accession a single reel of tape and prepare it for dissemination when it is software independent, is in a standard code, and requires no data compaction.[6] While data compaction is not very costly for a single reel of tape, it can become quite expensive when hundreds of reels of tape are involved. It costs even more to convert data to standard character codes or unpack it from a data base management system or statistical analysis package. Currently, such conversion costs run between $400 and $600 per reel of tape. The long-term preservation costs using existing storage technology over the next twenty-five years would be about $5 per year for each reel of tape.[7]

Using these estimates, the approximately 1,500 reels of tape now in our holdings represent an investment of approximately $1 million, one-half of which consists of conversion and projected preservation costs. Since these 1,500 reels of tape comprise about one-half of 1 percent of the total volume of tapes we estimate are of sufficient value to warrant acceptance by the National Archives, it is clear that an enormous sum of money—probably on the order of $200 million—would be involved in accessioning, converting, and preserving the projected 300,000 reels of tape. Conversion and preservation costs are of such a magnitude—approximately $100 million—that they cannot be ignored in determining whether a file will be accessioned into the National Archives. Indeed, given our limited resources, increasing caution and care must be exercised in selecting files to be accessioned.

Since the application of a rapidly growing computer storage and retrieval technology seems inevitable, there are at least two consequences which are potentially very troublesome. The first concerns the development of mini-computers which are relatively inexpensive and powerful enough for many data processing tasks. Word processing is but one such task that now is on the threshold of enormous growth. It is too early to predict the impact word processing will have on the appraisal of machine-readable records. On the other hand, since more traditional data processing tasks will be performed without our having to

[6] It should be noted that this cost includes appraisal, description, and validation of a finding aid at the item level. Also, a fully packed 1600 bpi tape can store the equivalent of about 15 cubic feet of textual records. Thus, the cost corresponds to about $24 per cubic foot of textual records.

[7] This estimate is based on the use of 1600 bpi computer tape and assumes recopying tapes every ten years. When compared with the cost of storing an equivalent quantity of textual records, this is rather inexpensive. Use of a different record mode and storage medium (such as optical laser recording) could greatly reduce preservation costs.

rely on computer programmers and a central computer facility, mini-computers eventually could be as widespread as electrostatic copiers are today. This raises the possibility of an enormous proliferation of machine-readable records in a highly decentralized environment. The absence of any central control would make it more difficult to ensure the proper disposition of machine-readable records.

A far more serious problem will be the increase in data base management systems. Such systems provide an on-line or interactive computer environment in which users may retrieve and analyze data while working at a desk-top typewriter-style terminal with a video display screen. A significant aspect of an on-line computer environment is that random access capability eliminates the necessity for the sequential or serial storage and retrieval common to magnetic tape files. In an on-line computer environment a user can retrieve information from a variety of storage areas instantaneously without needing to know where the data is physically located on a disk or disks. Because sequential storage and retrieval no longer constrain the arrangement of data, the computer itself decides where subsets of a data file will be stored, with the criterion being maximum utilization of the available storage capability.

Since the concepts of processing files and master files, crucial in the development of disposition instructions for machine-readable records, are based on the paradigm of magnetic tape files, it is possible that these concepts may no longer be viable in an on-line computer environment. This possibility becomes even more likely when data is merged from a variety of different files or created as a subset of a larger data file. Typically, an analyst would select the data elements needed for a particular study and then instruct the computer to create a temporary data set composed of these elements. His subsequent statistical analysis might be displayed on the video screen with results copied by hand or printed out by the computer. Once the analysis is completed, the temporary data set would be erased. The only evidence of the process would be the analyst's brief report which might include a copy of the summary of the statistics. And even its existence would depend upon the analyst's work habits. In this scenario there are no processing files in the usual sense of the term, and the automatic erasure of the temporary data set means there would be no master file. Furthermore, the problems of provenance could become insoluble, since there would be no record of data transactions to reveal the sources of data.

Earlier it was noted that preservation will become increasingly expensive in the future. The reason for this is rather simple. A 300 megabyte disk (the equivalent of about 7 reels of tape written at 1600 bpi) may contain a variety of data organized by a computer to maximize storage space on the disk. It is possible that one or two very valuable files may be dispersed throughout the disk in such a way that it would be most economical to accession all of the data, the garbage as well as the treasure, because the expense of having a computer assemble subsets of data that comprise the one or two valuable files may be prohibitive. An even greater problem is the cost of making these files accessible. Essentially, there are only two alternatives: to accession the files as they are embedded (formatted) in a data base management system along with the data base system itself, or to unpack the files so that they no longer are software dependent. Either alternative will be very expensive, although the latter may well be less costly in the long run since updating it to new technology will be relatively easy and inexpensive.

Even a modest increase in preservation costs could have the unfortunate consequence of requiring even more solidly demonstrable research potential as a justification for accessioning. Of course, the more closely a file relates to current research the easier it is to demonstrate its research potential. This could lead into the cul-de-sac where increased preservation costs result in less and less machine-readable material being preserved for fewer and fewer researchers. Given this situation, it is conceivable that social-science researchers fifty years from now will be in an era of data poverty because so little useful data will have been preserved. The irony could be that a great volume of data may in fact have been preserved. This rather bleak outlook for the social science researchers and the archivists of a half century from now could be compounded even further by the fact that the proportion of machine-readable to textual information is steadily increasing at both the state and federal levels.

Of course, this is not a problem confined to machine-readable records since these developments will affect all records. Some of the long range consequences of rationalizing the record disposition process begun in the 1940s are beginning to emerge, albeit not very clearly. The systematized process of records retention, along with severe economic constraints, may result in preserving fewer and fewer records of less and less long-range research value without regard to storage medium!

This rather gloomy assessment of the future may not set well with many archivists and researchers. Certainly, there is a positive side that should be noted. There is abundant evidence that changing computer technology eventually leads to greatly reduced computing costs. It is possible that, in the long run, computer storage technology will make it feasible to ignore preservation costs in appraising machine-readable records. This would be of no little consolation to a growing number of archivists who will appraise machine-readable records.