

January 1983

Appraisal Strategies for Machine-Readable Case Files

Ross J. Cameron

National Archives and Records Service

Follow this and additional works at: <https://digitalcommons.kennesaw.edu/provenance>



Part of the [Archival Science Commons](#)

Recommended Citation

Cameron, Ross J., "Appraisal Strategies for Machine-Readable Case Files," *Provenance, Journal of the Society of Georgia Archivists* 1 no. 1 (1983) .

Available at: <https://digitalcommons.kennesaw.edu/provenance/vol1/iss1/6>

This Article is brought to you for free and open access by DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Provenance, Journal of the Society of Georgia Archivists by an authorized editor of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.

APPRAISAL STRATEGIES FOR MACHINE-READABLE CASE FILES*

Ross J. Cameron

The increasing use of computers in both administrative record keeping and in social science research challenges archivists to reevaluate previous appraisals of many types of records. The creation of computerized data bases which contain information from legal, criminal, medical, welfare, and other investigatory, regulatory, client, and personnel files has brought case files into this group of records which must be reappraised. Although they may include the same subject matter as textual case files, machine-readable case files may be appraised differently because of the media on which they are recorded and the amount of information they contain. This new media greatly reduces some of the problems in preserving and using textual case files--for example, large volume, slow and tedious access, and protection of privacy.

The appraisal of machine-readable case files involves essentially the same considerations as that of other machine-readable records. In reviewing these general considerations, their particular relevance to case files where appropriate will be discussed.¹ Though these comments are primarily based on experience in the Machine-Readable Archives Branch, National Archives and Records Service, the author does not intend that they apply only to federal records or even only to government records. They should apply to any machine-readable case files.

*The opinions expressed in this paper are those of the author and should not be construed as official policies, procedures, nor recommendations of the National Archives and Records Service.

Several primarily technical considerations must be made first in the appraisal of all machine-readable records. The appraiser must determine that the file is the final edition, or master file. A raw input data file or working copy should be disposed of unless it is the only, or most thoroughly edited, version of the file extant. In rare instances two versions of a file might be kept if the appraiser discovers that editing of the file included the deletion or alteration of data elements or units of observation for reasons other than the correction of keying errors, data inconsistencies, and the accidental duplication of cases. Changes reflecting subjective judgments which might bias the data or its interpretation should be carefully examined and noted in the appraisal.

Adequate technical documentation is an essential part of the file. At the very least the documentation must include a record layout, which indicates the location of each piece of information in the record, and a codebook, which explains the value or meaning of coded information. Sample forms on which the information was first recorded and reports on the uses made of the data and the conclusions drawn from it are also important documentation. Operator's and user's manuals, which explain the processes of creating and using the file, are useful for the appraisal, though they might not be included in the documentation package.

Another crucial technical consideration is the readability of the data. If the physical condition of the tape is such that a portion of the data is neither readable nor recoverable, the appraiser must decide whether the extent of and possible bias from the damage is sufficient to invalidate the usefulness of the file. This judgment is based on the general archival decision on the legal, evidential, and informational value of the file.

If these technical considerations are satisfied, the archivist then makes the primary archival evaluation as to whether the legal, evidential, and informational value of the file merits its permanent retention. For machine-readable records the informational value is usually the most important of these. This is determined by the subject matter and quality of the data elements in the file; the extent of its coverage, and its potential for further analysis or reanalysis. As for all records regardless of

media, the significance of the subject matter is judged on the basis of current and predicted future research trends.

Several interrelated trends in research have led to the initial or increased research use of many case files and other records. Social and economic history have grown considerably in recent years. Numerous sub-fields such as welfare history, labor history, the history of crime, and the history of physical and mental health have also developed. And interest in historical approaches among sociologists, economists, and other social scientists have expanded the research community. The growth in the use of quantification and statistical analysis among historians and other social scientists has significance for both textual and machine-readable case files. Machine-readable case files provide an ideal source for prosopography, or collective biography. If the file contains members of the group a researcher wishes to study, the personal characteristics are already collected and ready to be analyzed to provide a group profile.

Many case files are very important sources for the growing study of non-elite history, or history from the bottom up. Most non-elites do not leave records; or, if they do, their descendants do not retain them nor deposit them in appropriate repositories. Therefore, most historical records that survive are from or about elites. Since the study of history is necessarily based on surviving records, it thus has been biased because it overwhelmingly reflects the ideas and activities--the lives--of elites. Just as the records of prominent or wealthy people are more likely to be preserved, so are those of prominent or large businesses and institutions.

Direct and indirect government involvement in the daily lives of non-elites has expanded greatly in the last half century with the growth of regulatory and social welfare programs. This has resulted in the creation of large volumes of records with information on the otherwise unrecorded characteristics and activities of this segment of the population and of the business and social world. Some of the present imbalance in the records of elites and non-elites can be rectified by the retention of case files. For example, the Machine-Readable Archives

Branch has accessioned the case files of the Equal Employment Opportunity Commission. In order to monitor compliance with the Civil Rights Act of 1964, the commission maintains files of demographic and sociographic information on the employees and trainees of private employers, joint labor-management apprenticeship programs, employer-conducted apprenticeship programs, state and local governments, and public elementary and secondary schools, and members of labor union locals. Other records on many of the smaller businesses and institutions included in these data bases will probably not be retained for future research.

In addition to the subject matter of the data the archivist must also appraise its quality. Quality is judged on three criteria--reliability, validity, and accuracy. Reliability refers to whether all persons using the same procedures would arrive at the same value for the data element. That is, would everyone count or assign the same numeric value or code? Validity refers to the appropriateness of the procedures, or operational definition, of the data element. That is, does this data element truly represent the concept being studied?² And accuracy refers to whether the data has been keyed correctly. That is, is the value within the specified range of values for that data element, and is it logical in relation to other data elements in that record? The investigatory and regulatory nature of many case files makes the reliability and validity of subjective judgments and other data very important. For example, a drug user information system might be disposed of because the subjects were persons arrested for other crimes but suspected of being drug users by the arresting officer with or without any evidence.

The extent of the file's coverage, or its universe of observations, is another important archival consideration. The chronological and geographical coverage usually should be sufficient to provide representative coverage of the subject matter. If the machine-readable copy is a sample from textual records, then the validity of the sampling procedure must be evaluated. If records for only a brief, insignificant time period exist, then the file should be rejected. Limited geographic coverage

may indicate which repositories would be appropriate for retaining the data, or it might result in the file's destruction. For example, a Wiretap Commission file was created to analyze the success of wiretaps in prosecuting suspected organized crime figures. However, data from metropolitan New York City where a very large portion of the cases occurred was not received. This lack of coverage could result in the file's disposal.

One of the most important considerations on the informational value of the file is its potential for further analysis. This further analysis may be of two types. The first is internal--can the information within the file be analyzed in ways beyond those which the creator and users performed? Machine-readable case files may be very susceptible to this type of further analysis because the creators are often primarily interested in individual cases rather than in profiles of all subjects. Demographic and sociographic information recorded for routine identification purposes may seldom, if ever, have been analyzed in conjunction with the subject matter data. This is particularly true if the information is made machine-readable for tracking or housekeeping purposes rather than for research. For example, a Housing and Urban Development file on rehabilitation loans and grants was created in machine-readable form to simplify tracking of loans, grants, and loan repayments. But, it also contains routine demographic and sociographic data which is not analyzed by the agency and, thus, offers potential for further analysis. Tracking systems may offer other opportunities for study, especially time-series analysis.

The second type of further analysis is external. This refers to the file's potential for linkage with other data files, particularly ones not available to the creator or previous users. Direct linkage is possible if each file contains personal identifiers such as name, social security number, or some other common identification number. Greater potential for direct linkage exists with files of related subject matter. Indirect linkage, or cohort analysis, is possible if the files contain common demographic or sociographic data elements such as age, sex, occupation, education, or geographic location. In

this way case files might be linked with census or survey files resulting in potential for new analysis.

The fact that a file has been thoroughly analyzed and cannot be linked with other files does not necessarily mean that it is disposable. If the initial analysis resulted in significant or controversial findings, other researchers in the field may wish to reexamine the data for themselves in order to evaluate or confirm the original conclusions. Members of many disciplines and sub-disciplines have called for data archives for this express purpose.

Machine-readable case files, thus, offer much potential for informational value. They may also lead to a growing concern for the possible legal and evidential value in machine-readable records. As more agencies and institutions come to depend on computers for storing and using large volume file systems, case files may exist in machine-readable form only. The investigatory and regulatory nature of case files would thus make them important sources of legal and evidential value. For example, the Immigration and Naturalization Service is in the process of creating a large statistical reporting system which will be the primary source for information on aliens and deportees, eventually replacing textual records.

In addition to technical considerations and the legal, evidential, and informational value of the file, the appraiser must also take into account whether the information exists in another media or mode. If so, the archivist must answer a series of questions about the other copy. Is it available? Will it be preserved? Does it have a more useful arrangement? Will it be more or less expensive to preserve? Will it be easier or less expensive for researchers to use? And finally, does it contain more or less information?

These questions are very important in the appraisal of machine-readable case files since there are often associated textual case files. Because of the large volume of most textual case files, machine-readable copies offer advantages in terms of space and other preservation and reference costs. Even more important, it is much easier to delete personal identifiers and to provide

disclosure-free copies of machine-readable case files to researchers for as long as privacy has to be maintained. The great concern over the protection of individual privacy makes this aspect very important. It is also much easier to extract cases with desired common characteristics from a machine-readable system.

Sometimes, however, machine-readable versions of the tracking or housekeeping type mentioned earlier may also present serious problems. First, they may not contain enough information from the textual case files to merit their retention in place of the textual records. Nevertheless, they may be useful as indexes for the selection of cases with desired characteristics. Second, they may be online systems containing only current data on active cases. If previous data and nonactive cases are not transferred to a history file, this system would not merit retention. As more agencies convert to machine-readable systems for large files, these latter problems should diminish.

In review, machine-readable case files offer some advantages over textual ones: (1) they are already prepared for the prosopographer to generate subject profiles and perform statistical analysis; (2) they have great potential for further analysis, especially linkage with other files whether directly or through cohort analysis; (3) it is much easier to provide researchers with valuable files and still protect the privacy of the individual; (4) last, but certainly not least, is the savings in storage space and other preservation and reference costs

NOTES

¹The author has generally followed the decision table for machine-readable records in Charles M. Dollar, "Appraising Machine-Readable Records," *American Archivist* (October 1978): 426-427.

²Hubert M. Blalock, Jr., *Social Statistics*, 2d ed. (New York: McGraw-Hill Book Company, 1972), 12-13.