# Archival Principles and Records of the New Technology

TRUDY HUSKAMP PETERSON

**Abstract:** Basic archival principles apply to records created by computer technologies. The concepts of informational and evidential value as criteria for retention and the concepts of provenance and original order as keys to arrangement will continue to provide a framework for archival analysis. Description techniques will have to be augmented to accommodate more refined descriptors. Although these basic principles remain applicable, the computer technology forces archivists to address anew the problems of obsolescence and change, privacy and public use, mixed records systems, changing research demands, and the nature of the archival profession itself.

*About the author: Trudy Huskamp Peterson is a staff member of the National Archives and Records Service, where she has held a variety of positions. Her undergraduate degree is from Iowa State University and her advanced degrees are from the University of Iowa. The views expressed in this article are solely hers and do not represent the official position of the National Archives and Records Service.*

*This article is a slightly revised version of the core session paper presented at the 46th annual meeting of the Society of American Archivists, 21 October 1982, Boston, Massachusetts.*

TO THE MAN WHO HAS A HAMMER, the whole world looks like a nail; to the person who has a computer, the whole world looks like input. Or so it seems. We are awash in computer jargon and inundated by propaganda about the paperless office. For archivists this electronic future is a bit frightening; but because it is heralded by hype, it is also a bit boring. Many archivists ignore the new technology, hoping to retire before it invades their archives. They may be right on an individual basis; collectively, as archivists, we cannot ignore the monster in our midst.

For the archivist, the most significant question involving machine-readable records is whether they can be managed in accordance with basic archival principles. Archival theory gives archivists a framework for analysis, a way to understand the world of records. If this theory will work for machine-readable records—as we know it works for textual, audio, video, and photographic records—then the records of the new technology will not crack the foundation of the archival world. If, however, basic archival principles cannot be applied, fundamental adjustments in archival analysis must occur.

This article will focus on the application of basic archival principles to records created by the new technology. It will not look at the application of new technology to the management of archives—that is another issue. It will argue that basic archival principles apply to records created by the new technology, although in most instances the principles have to be amplified somewhat.

First, we must define what the basic archival principles are. For this discussion we will consider, under the rubric of appraisal, the concepts of informational and evidential value as the criteria for retention. Under the rubric of arrangement, we will consider the concepts of provenance and original order and look at the definitions of a series, file, and record. Under the rubric of description, we will consider the level of description necessary to meet the user's needs. We will not consider preservation in specific, although it will be mentioned in various sections, because preservation is increasingly the domain of specialists, not the domain—except in a managerial sense—of archivists.

Second, we must define what we mean by the new technology and how it is used in record keeping. In the term "computer systems" we now include not only the familiar piece of equipment that has a keyboard like a typewriter for an input device and that stores the typing on tapes or disks, but also the video and optical disk technologies that give a look-alike representation of the original item but generate that representation by means of a computer. Throughout this article the look-alike image, which is analogous to the original item, will be called a facsimile output.

The task the early computers did best was to count and sort. Consequently, the computer is perhaps most extensively used in housekeeping functions: payroll, personnel, procurement and supply, property management, financial management, mailing and distribution, and so forth. As the computer has evolved, it has been adapted to a wide variety of other uses. It stores and manipulates programmatic statistical database systems and is ideal for maintaining bibliographic databases. It is used in tracking systems (such as keeping track of the status of a lawsuit for a law firm or the stage of the award of a contract or grant), for indexing (often indexing records maintained in paper or film format), for scanning and recording (whether from a NASA satellite viewing the Earth and giving us the weather report we see on television each night or within an office scanning a

page of text), for printing and publishing, and for simulation modeling. Increasingly it is used as a whole-text storage system.

In spite of the varied uses, the computer has been very imperfectly integrated into institutional life. One characteristic of current record keeping systems is their mixed formats: for example, the index is computerized but the records are stored in paper; the outgoing correspondence is stored in the computer but the incoming is stored in paper; the outgoing is in the computer but the incoming is a mix of computer storage and microfilm of paper; the index is on-line but the storage is film. The point is that the systems we know today are likely to be transition stages in a major shift of record storage methods, and the characteristics of the mature system are still a matter of conjecture.

Amazing as the computers are, their function in an institution is simple: they are purchased to assist the institution in conducting its business. As business tools, they produce records, the business form of the document. The classic definition of the document is an item with three parts: a base, an impression upon the base, and information. By insisting that the information be fixed in some way on a base, the definition excludes purely oral communications, such as the conversation in the office, the chat on the street corner, or the unrecorded telephone call. Implicit in the definition is communication, the idea that by fixing the information it can be understood by someone who had no personal contact with the originator, or, similarly, that at some later time the originator can use the document to recall information to memory. The crux of the definition is the act of fixing the information—not the type of base, nor the type of impression, nor the character of the information, nor the length of time it is fixed.

Bearing that definition in mind, it is clear that format makes no difference to the fundamental nature of a document or a record. So, too, format makes no difference to archival principles, for these principles relate solely to the fact that the product of the activity is a record. Further, the principles relate primarily to the selection, preservation, and use of the information in the records and only secondarily to the base and the means of recording employed.

Let us look first, then, at appraisal of records in the new formats. If we are honest, we admit that all appraisal has two sides, one intellectual and one practical. On the intellectual side, we ask questions about the information: evidential and informational values and what the user can learn from and do with the records. On the practical side, we ask questions about whether we can store the records, how much it will cost to preserve them, how often the records are to be transferred, whether sampling is practical, and whether a researcher will ever use them. The intellectual side of appraisal is identical for any record format. The appraiser asks the most fundamental questions about the character of the information in the records, and it does not matter whether the records are photographs or electronic blips. It is on the practical side of appraisal that differences arise from format. Let me point out several, with respect to records of the new technology.

First, with paper records, storage space has been the engine driving the appraisal train. With information stored so compactly in machine-readable format, this engine largely stalls. Instead, however, the cost of preserving machine-readable records, including the updating of the files as newer technologies appear, becomes an important factor.

A second difference is apparent in questions of transfer of the records. In

paper or microform, the transfer is quite simple: at a designated time the file is cut off and physically transferred to the archives. With machine-readable records this is not necessarily so. Sometimes specialized studies are completed and can be transferred in a manner similar to transfer of paper records; but very often new information is simply added to the entire store of records in the machine until there is a large, amorphous, undifferentiated porridge of information in the belly of the computer. If you ask for a yearly transfer, do you want a complete copy of everything in the machine or just that information that has been added since last year? If you want only the added items, what are you going to do with the information when you get it? Will your user want to merge the information from a number of years anyway? Should you merge it? And so on.

A third difference on the practical side of appraisal comes with considerations of information loss. We all know that paper records are lost because records creators throw them away, but it normally takes a certain amount of decision making to haul files from a file drawer and dispatch them to the trash. With machine-readable files, however, the elimination of records may be built right into the system. For example, if the system is used to track the status of a project, when the project goes from the design office to the mock-up room, the words "design office" would be purged from the system and the words "mock-up room" inserted. If you ask for a yearly cutoff of this file, all you will get will be a "snapshot" of the operation at the time of cutoff. The same is true of systems that are designed to allow editing of studies, reports, or legislative bills. Unlike paper files, wherein drafts accumulate, some of these systems simply delete anything except the most current version. If the archivist wants to maintain the records of stages of a project, he must work with the computer programmers to capture it all.

A fourth practical consideration is whether the records are software-dependent. Software is the set of instructions that makes the hardware of the computer work. Every computer needs software, and in that sense all computer records are software-dependent. But in a file defined as software-dependent, the software links with the information in a code. A software-dependent file will print out as gibberish unless it is processed on a computer that has the right software. It is possible to convert a software-dependent file to a software-independent file, but the conversion is both complicated and expensive. The software in a software-dependent system may also be proprietary software, meaning that it can only be obtained from one source, sometimes at great expense, and the software instructions themselves may be such a large amount of information for the computer to hold while it processes the file that only major computer installations can handle the processing. This is even more serious with the proliferation of software-dependent microcomputers that can be decentralized throughout an institution for data storage and manipulation independent of, and perhaps incompatible with, the institution's main computer.

Fifth, in machine-readable records it is vital to have documentation that explains the location of a specific item of information within a record. This is a major problem in the numerical and statistical databases. If you are looking at a screen or a printout that is a string of numbers, they will not mean anything unless you know that, for example, if the fifty-sixth number is a two it means the person in question is female. The archivist must have enough information on how the data is laid out to make the data intelligible. Even if the file has been described to

the archivist as the most fascinating of information, if the documentation of the record layout is not available, the information cannot be interpreted and the archivist must discard the file.

Each of these practical considerations affects the decisions on the retention of the file, but it in no way undermines the primary intellectual determination of evidential and informational value. Let us look at some "worst case" examples.

### The electronic mailbox

This is a system that links all people working in the unit by utilizing display terminals that can hold, send, and store messages. In some ways it is a cross between a letter and a telephone. The critical issue here is not the system but the character of the information generated on it. If the electronic mailbox is used principally to tell people when a meeting is scheduled, announce a fund drive, or send updated instructions on a procedure, the question the archivist must ask is whether there is secondary archival value to be served by capturing this information. If the system is used to draft policy papers, or to discuss collectively the results of experiments and hypothesize about possible explanations, the question the archivist must still ask is whether there is secondary archival value in capturing this information. Once that decision is made, it then becomes a technical problem to determine how to capture the information, and the problem will have to be solved in concert with computer specialists.

### The simultaneous database use

Let us say that an oil company has a database that includes all the basic facts about geological characteristics of rock formations that are potentially oil-bearing. A company geologist uses this database in concert with computerized information on a certain geographic area in Western Utah, displaying on a computer terminal the locations where the computer search has revealed a positive correlation between the characteristics in each database (known as the hits). From this display the geologist selects the most likely location for drilling and tells the company managers of his decision. The question for the archivist is whether there is secondary value in the characteristics database, in the geographic database, and in the exact information that was displayed on the screen when the geologist searched the files. Once again, if the archivist can make that intellectual decision, then it becomes a technical problem to decide how to capture the information.

### Updating databases

Let us say we are working with a state welfare agency, and the agency has all its welfare case files stored in a computerized database. From the database the computer calculates the amount of money to be sent to each welfare recipient each month. The computer must store the most recent address of the welfare recipient so that the check issued by the computer has the correct mailing address. Consequently, when a recipient moves, the old address is wiped out and the new one inserted. If the archivist believes that there will be a future value in retaining information on the geographical mobility of welfare recipients, should the archives retain all the addresses of each person on the rolls? If the archives wants to do so, it then must solve the technical problem of how to capture the information.

In each of these instances, the archivist will need to make a decision, to work with the people running the computer systems, but most importantly to seek the support of the management levels in the institution. Only with strong managerial backing can archivists obtain cooperation in retaining such fugitive information as changing addresses and screen displays of

matches between databases. This is especially true if the programming is expensive or if the retention is inconvenient for the administrative and project personnel.

Turning to the issue of arrangement and records of the new technology, provenance remains the most significant principle for the archivist. Who created the records is the fundamental question, and the answer varies just as it does with paper records. The records may be the general correspondence files of the Department of State; they may be a special project file created by a contractor on behalf of a corporation; they may be a statistical database of all welfare recipients that is maintained by the state center for automated data processing (ADP) on behalf of the state welfare office; or they may be an experimental database created by one scientist and used by all members of the research group. Someone somewhere had to create the records to fulfill an organizational need, and that fact is the clue to provenance. Normally, even if the records are obtained by the archivist from the central ADP service, that unit would not be called the agency of provenance—that is akin to calling a typing pool the agency of record. The key is the person who created or ordered the creation of the information, not the person who typed or scanned it into the machine.

As archivists know, there is a difference between physical and intellectual arrangement, e.g., between the position of an item on a shelf and the position of that item in an inventory. Just as with paper records, machine-readable records may be physically arranged in any order, so long as they are easily retrievable. The focus of archival concern is the intellectual arrangement of materials.

Some archivists have suggested that principles of arrangement are irrelevant for machine-readable records because the data is stored randomly. Upon examination, however, the reverse is more nearly true. Because the records are not human-readable until converted by the machine and because the cost of using the machine to find records is expensive, the location of the records is the most critical factor in using the records. This is true whether the user is trying to locate the one of 600 reels of welfare cases that has a certain case file on it, to find on a videodisk of 10,000 images the one image that has a drawing of the Molly Maguires, or to tabulate and compare the result of a scientific experiment. Let us examine this assertion for each of the five levels of archival arrangement.

Arrangement of records at the repository level depends on adminstrative convenience, not format, and arrangement of the record groups and subgroups is derived directly from provenance. At the series level it becomes more complicated.

The usual archival identifiers for a series are common filing order, subject matter, or physical type. Order and type are distorted by the machine-readable format, e.g., a single videodisk may store a memo, a still photograph, a map, and a book, all of which may or may not be related. With these easy physical characteristics eliminated, archivists must use intellectual characteristics to define the series. One possible definition of a series in machine-readable format is that it is the largest intellectual entity that is created and *recognized as an entity* by the creator of the records. In discussing paper records archivists commonly say that the creator cares about order within a series but not order among series and that it is the job of the archivist to restore order within the series and to create order among series and among subgroups and record groups. Just so, in digitally stored records the creator will want to call up information in groups and will characterize the group in some way (e.g., general cor-

respondence, case files, personnel files). The most inclusive of these intellectual groups is the series. The series is relatively easy to identify when there is facsimile output, for the archivist can look for the same kinds of characteristics that he seeks in paper records. When the output is wholly numerical, however, series identification is more difficult. Some of these numbers (or number–symbol or number–letter–symbol combinations) may represent case file information, but all the information has been converted into numbers. In other files, however, the information is truly numerical, much as the entries in a nineteenth-century ledger are numerical. Numerical output files are initially harder to understand, but once the numbers have been converted to their plain language equivalent—a process roughly like putting headings on columns of a table—it is relatively easy to see what is the body of related information that makes up the series. Often all the records in the numerical-output series have a similar basic layout with similar pieces of information, and the entire body of information can be analyzed as a whole to draw conclusions. This can usually be considered a series. There are, of course, gray areas in identifying numerical-output series; but it is also difficult at times to determine series in textual records.

The file is also quite easy to identify when the information is in facsimile output format. In the machine-readable file, a code is often used to link together the related "pages," just as a file folder holds paper together. Using that code, a case file can be reviewed sequentially, just as it can in paper. Again, like the series, when the information in the file is solely numerical, the definition and identification of the file is sometimes less easy then identifying the facsimile output file. Perhaps the numerical file is easiest to identify by exclusion: it is that group of information that contains more than one record but is less than a series and forms a coherent body of information. (While archivists can intellectually describe a file, the jargon of the computer industry makes it difficult to keep the distinction clear. The industry uses common but confusing terms such as "data set," "data file," and "file," any of which may be identical to a series or a file as an archivist would define it. Perhaps as the computer world evolves the language will become more standardized, but at present words like "file" can confuse as well as clarify.)

Fortunately the identification of the record is easy. The archivist is interested in logical or content-based records, which may differ from the order in which information is stored on the tape or disk (the physical record). If the information is produced in a facsimile format, the visual display will show the logical record identifiable just as it is on paper. With numerical output, a logical record is still easy to identify because the associated code book will always explain where one record ends and another begins. The length of the record in numerical form may be fixed (that is, always the same length) or varied, but the user will always know when a new record begins. Records are still the building blocks of an archives.

In sum, arrangement is still an important consideration when archivists handle records of the new technology. The machine-readable records format emphasizes the intellectual arrangement of the records. Physical arrangement is important in two ways: to enable the machine to locate the physical record within the storage device and to enable the archivist to locate the required tapes or disks, especially when a single file or series is spread over multiple reels or disks for storage purposes.

Of all the archival practices, descrip-

tion is the key to records in the new technological formats. While repository and record group descriptions are the same in whatever format, descriptions of machine-readable records at the series, file, and item level must be augmented. There are two major reasons for this. First, there must be an index, because, for all practical purposes, it is not possible to browse through the records. This index may be to the item (as would be true if a disk stores non-verbal information, e.g., political cartoons), to the file (such as the welfare case file coded by file number), or to the volume and chapter (if, for example, a handbook or manual is on a disk, so that the user can find the section needed). The index may be either to one tape or disk, if all the information is stored there, or to more than one tape or disk, if the information is spread out. Indexes are particularly important if the digitized data is non-verbal such as a photograph, map, or movie. In some cases the index is built right into the software manipulating the machine (it might alphabetize and retrieve by name, for example); in other cases the index is a separate display on the computer screen. Fortunately for the archivist, such indexes are absolutely necessary for the creator of the records, too, and the archives will inherit them along with the records. The archives must thoroughly describe these indexes, however. Both the elements indexed and use of the index itself must be described.

A second reason for expanded description in machine-readable records relates to research use by persons other than the creator of the records. If the records are primarily statistical, full descriptions of the file and series must be made to accommodate the use of the records by themselves or—most importantly—by linking them with another body of records. Linkage is the process of simultaneously using two or more files that have a common characteristic to produce information that cannot be obtained by using one file alone. If a file or series can be linked, its potential research use increases. If, for example, a city police force keeps a computerized record of each arrest including information on the reason for the arrest (possession of drugs, assault, or breaking and entering, for example) and the address of the person arrested, and if a separate computerized file contains information on the average income by ward as found in the 1980 census and on the streets located within each ward, by using the two files together a researcher could determine that arrests for a particular charge occur more frequently in one ward than in another. That is linkage. Linkage can be done with textual records, too; but it is much more laborious and time-consuming. The researcher needs to know exactly what pieces of information are in the files: residence, income, reason for arrest, and so on. This is a much more detailed record description—listing the elements within the item—than is common elsewhere in the archival world, although occasionally a series description for a ledger will list the ledger heading, which is similar to this description. In textual records one can omit the exact heading in a description if one chooses; in computerized records one cannot. That list of elements is normally the vital information that the user wants to know before he purchases or plugs into the file. Without that information, he cannot proceed.

If the archival principles for appraisal, arrangement, and description remain viable, what are the important issues that the new technology forces us to address? Five seem particularly important: obsolescence and change, privacy and public use, mixed records systems, changing research demands, and the nature of the archival profession.

First there is the problem of obsolescence of computers, programs, and

storage media (tapes and disks). Computers are changing so rapidly that it is sometimes hard to find a machine to read a tape or a set of cards made in the 1960s. There is no indication that the rate of change will slow soon; and this means that, if archivists wait twenty or thirty years before accessioning permanently valuable tapes or disks into an archives, they run the risk of not being able to read them or to read them only at great expense. Related to the obsolescence of the machinery is the obsolescence of software, especially the type that makes the information fed into the computer become software-dependent. This includes the increasingly popular database management systems. To make these files available the archives must have the software—and in some cases that exact type of hardware—yet the software is changed and modified and superseded as frequently as is the machinery. Related to both the hardware and software problem is the fragility of tape as a recording medium. The danger of obsolescence leads to the argument that the archives must accession soon after the tape is created or monitor the records creator's computer facility to ensure that as technology advances any permanently valuable tapes are converted to the new format. Taking records in quickly, however, increases the problem of determining the evidential and informational values in the records and greatly increases the probability that access will have to be restricted to some parts of the records. On the other hand, monitoring creates its own set of headaches. We must hope that a widely compatible disk, preferably of a single material and not a sandwich construction, will be developed as a storage medium. Optical disks may be part of the answer, but the rate of change is as rapid as ever, and the industry standardization of the hardware and software is not in sight.

A second issue is the realization that when we begin to hold quite current information in machine-readable form in an archives—information that is easy to link with other information in our possession and to send by telephone lines anywhere in the world in a matter of minutes—we are creating a frightening concentration of information. We must be exceptionally sensitive to our responsibilities to the public as users and to the public as subjects of our records. Modern records force the issue; and while it is possible—although expensive—to produce "public use versions" or "disclosure-free data sets," we will still hold unparalleled information on our fellow citizens. We must make sure that our professional ethics are adequate to the challenge.

A third problem is how to facilitate research on records in mixed formats: textual records with a computerized index, for example. Substantial expense is involved in keeping a major index in active memory in a computer, yet without access to the index no work can be done with the paper records. In an archives, there are many such indexes; and to keep them all in active state in a computer would require a very large and expensive machine. With very small indexes it is possible to print the index on paper and let researchers use it that way, but in most cases this is not feasible. Indexes are only one example of mixed formats: when the outgoing correspondence is stored in the computer and the incoming is on paper, the researchers simply must have access to the computer to link up both sides of the correspondence. Archivists can only hope that computers rapidly become cheap enough so that we can afford to keep numerous machines around, each associated with an index or specialized database in active use.

A fourth major issue is the changing role of the computer in research. We have

reviewed the needs for enhanced description for statistical research; it is vital that archives, data libraries, and data centers come to agreement on the elements to be included in a description of these statistical files. We must make it possible for the researcher who wants to match up police drug arrest records from Virginia and Nebraska to know by reading the archival descriptions whether it is possible to do so. Progress has been made in this area, but we have a long way to go. One assumption that has been made in the past is that the archivist or librarian would interpret the records to the user, either by letter, in person, or over the telephone. Another assumption was that the user would purchase the tape for use on the computer at his institution. The booming market for home computers, however, indicates that those assumptions must soon change. Many informational databases—stock market reports, airline schedules, the *New York Times Index*—are already available to be queried by telephone from the home computer. Surely the demand will come to use an active archival record in the same way. If the records were machine-readable, genealogists could, through a telephone link, check the census, immigration lists, and ship manifests and could thus complete their research without battling lines for microfilm readers. Then will archives put the description of the records on line so the user can review it before he taps into the records? What exactly would appear on the screen to assist the user? How will we as archivists change our description of the records so that the description truly gives the researcher everything he needs to know to operate over a telephone link?

A fifth major issue is the nature and extent of the change that technology will cause in the nature of the archival profession. Some people prophesy that there will be no distinct archival profession, that we will be absorbed into an undifferentiated information management profession. Part of the reasoning here is that information is a commodity to be shared, and the source does not matter. After reflection, however, that seems unlikely. There are distinctions between the way a data center and a data library and an archives handle information, and it is important that in one place the researcher can find the institution's records of enduring value, maintained as the creator created them, clearly identified as to the creator, and maintained whether or not they are actively in use for research. That is the role of the archives. Certainly archivists will be specialists within the larger community of information resource managers, but there will continue to be a distinct need for our skills. It does seem, at least for a while, that there will be a divergence between archivists and manuscript curators, because automated records concerns will come first and most strongly to those who handle records, not those who handle personal papers. If, however, a manuscript curator collects papers of authors, the products of home computers, like the one upon which this article was written, will concern him greatly.

These five problems are just some of the many that archivists must face as we move through this period of rapid technological change. Archivists must realize, however, that we will have little influence on the multi-billion-dollar computer industry. The industry neither knows our needs nor is interested in our problems; current advertisements call computer tapes "archival" if they last more than ten years. This is no slur on the industry. The paper industry has not been particularly interested in archival problems, nor has the photographic industry. With such a small percentage of records ever making its way into an archives, it is not practical to insist that an organiza-

tion ask its members to write on the best paper or use the most expensive computer tape; one cannot make the 99 per cent serve the 1 per cent, especially if it would affect the cost to the 99 per cent. Archivists must adapt to the circumstances created by widespread use of computers, must insist on the importance of the historical records, and must identify and solve our own problems. Our traditional archival principles will continue to serve us well, but we will continue to be the small boat on the ocean of records, fighting to ride each wave on an endless sea of changing record formats.