The Potentials and Problems of Computers

Robert V. Kemper

What is the *appropriate* use of computer technology for preserving the anthropological record? Some specialists agree with Jerry Pournelle, a well-known computer expert who has been writing since the mid-1980s that "real soon now" some combination of off-the-shelf computer hardware and software will solve our problems of salvaging, storing, and retrieving all manner of unpublished and published materials. Other experts, especially archivists who have been subjected to such unkept promises for more than two decades, are much less sanguine about the role of computer technology in dealing with the ever-expanding mass of anthropological information. Given the diversity of opinions and the passions computer technologies evoke, we need to examine their strengths and weaknesses for the preservation and dissemination of anthropological records. Moreover, as the number of anthropologists using microcomputers in the field, in the laboratory, and in their offices increases toward a hundred percent of the profession, we must also consider how to preserve information generated on and stored in diverse electronic media. This challenge may ultimately be more serious than that involved in preserving and accessing paper-based records.

Anthropologists concerned about using their materials for decades to come may be less worried about the preservation of an original paper document, photograph, audiotape, or computer diskette than in preserving the *information* contained therein. This strikes me as the central issue dividing archivists and historians of anthropology from the broader anthropological profession. The first group, for good reasons, wishes to preserve materials in their pristine form insofar as practicable and prepare appropriate finders' aids for them. The second, and much larger group, for other equally good reasons, is more concerned with access to vital information, whatever its physical form, and assurances of its reliability and completeness.

The major areas in which computers can help to preserve the anthropological record include cataloguing collections, building information systems, and retrieving data and documents.

Cataloguing Collections

Currently available computer technologies can help to organize collections more efficiently than can be done through manual systems. Many museums and archives are studying and implementing (sometimes

expensive) computer systems to provide better access to and control over their collections. Although the cost of acquiring and maintaining these computer systems should be amortized over the life of the objects included in the computer data-base management system, most institutions and funding agencies require that we justify the purchase and maintenance of such systems over much shorter periods, sometimes only a single budget year. Such requirements pose problems for long-term preservation.

Case Study #1

In 1984, I was faced with the challenge of dealing with what I have come to call the Isabel T. Kelly Ethnographic Archive, which I had inherited without foreknowledge, and which required three vehicles to move from Kelly's home in Mexico City to Dallas. After learning how to do the basic curatorial work of removing rusty paper clips and putting documents in appropriate acid-free folders within acid-free cardboard boxes and placing photographic slides and materials in polyester sleeves within light-tight plastic boxes, I confronted the problem of how to catalog the materials and make them available to interested scholars. My graduate-student assistants and I experimented with microfilm and microfiche, but soon abandoned those possibilities because of the poor reproduction of field journals written in pencil and the impossibility of reading large maps in microform. I also decided that a catalog of the materials had to be created and that such a catalog should reflect, insofar as possible, the original groupings of the materials implicit in their arrangement in Kelly's home. I used a computer data-base program (DayFlo) that I happened to own. With a small grant from my university and with considerable personal investment of time and money, we managed to create a good working catalog of the field notes, maps, photographic materials (slides, prints, negatives, and even a silent film), correspondence, and other related documents (Kemper and Marcucci 1989).

We proceeded on our own with little help from the archivists in the university's Special Collections library because they knew less about computers and data-base management software than I did. Our only real problem arose when, after having finished coding nearly all of the items in the collection, we discovered that the software's algorithm for sorting the code number assigned to each item did not work as we expected. We had not realized, and nowhere did the software documentation mention, that alphanumeric strings required leading zeros to sort properly. This required an additional pass through the database to add zeros where necessary.

Having a computer-based catalog for the Kelly Archive has made much easier the work of interested persons who have come to Dallas to consult the materials or have inquired about a particular item. Unfortunately, the software company responsible for the DayFlo program has not been as successful in marketing its product in the past seven years as have other companies in the same field (such as Ashton-Tate [now owned by Borland] with its dBase system; Borland with its own Paradox system; and Questor with its Argus system). Acclaimed by some reviewers as the

"Database Product of the Year" in 1984 for its significant innovations (especially for its ability to permit up to 32,000 characters per field and for its built-in word processor and report generator), DayFlo nevertheless failed in the marketplace. Thus, although the computerized catalog of the Kelly Archive works well, other scholars interested in the Kelly materials have not been licensed users of the DayFlo software. As a result, I had to make paper copies of the catalog in order to avoid copyright problems that would arise if I distributed the software program itself along with the data-base file containing the catalog.

None of the scholars interested in the Kelly materials has as yet chosen to purchase the DayFlo software (for about $100) and then learn how to use it so as to have rapid random access to the items catalogued. While I could have acquired a license from the software company to distribute the program with the catalog, my limited resources did not permit this. My alternative has been to spend the time and money to move the catalog of the Kelly Archive from DayFlo to Paradox (which did not exist in 1984 but is now relatively widespread in the marketplace) and create a simple ASCII (American Standard Code for Information Interchange) version of the catalog. This ASCII version can be used with virtually any word processing software.

The lessons from this case study for archivists and others dealing with existing (and future) collections of anthropological materials are clear.

(1) Develop catalogs that depend on data-base management software whose market share is already well-established among potential users (i.e., anthropologists rather than the staffs of libraries or museums), because when (not if) these programs become obsolete, software companies will provide a "migration path" from the old programs to new programs.

(2) Develop an ASCII version of the catalog so that users who have only simple word processing software can gain rapid random access to the items catalogued.

(3) Prepare paper copies (on acid-free paper) of the catalog, including indexes for all significant key words likely to be of interest to potential users.

(4) In both computer-based and paper versions, include an explanation of the field names and codes used in the catalog.

Building Information Systems

Writing at a time when personal computer technology was primitive by today's standards, Van Houten stated without hesitation that "At present, no magnetic storage medium meets archival standards for permanence and durability" (1985:73). We still know too little about the permanence and durability of computer diskettes, compact disks, and laser disks. Even though well-established companies (such as 3M, Sony and Phillips) claim that information stored on the latest version of CD-ROM (Compact Disc - Read Only Memory) should survive for fifty to a hundred years, the

debate about the useful life of computer-based storage media continues. Without waiting for the experts' final verdict, anthropologists are already using computers to build complex information systems. Therefore, it is important to examine the major issues involved in building information systems from extant anthropological records.

Data Conversion

Assuming that one is willing to convert paper documents, photographs, drawings, and other anthropological materials into an electronic format so that the information can be made accessible to other scholars of the present and of future generations, it still must be determined how to perform the task. Given current computer technologies, one can depend on direct keyboard data entry for textual materials or use high-speed scanners to copy most kinds of textual and graphic materials up to 34" x 22" in size. Larger format documents (such as site maps or archaeological profile drawings) need to be photo-reduced to be scanned, which may mean a loss in resolution.

Keyboard Data Entry

Perhaps the slowest but surest way to create an electronic version of original paper documents is to have them typed directly into a computer via a word processor or database management system. This is labor intensive and therefore can be quite expensive. In almost all cases, keyboard data entry changes the physical appearance of the information being preserved.

Case Study #2

One of my goals in developing the Kelly Archive was to find a way for her extensive unpublished field materials to be made available to interested persons. The first project involved Kelly's fieldwork among the Coast Miwok of northern California. Mary Collier and Sylvia Thalman, representing the Miwok Archaeological Preserve of Marin (MAPOM), came to Dallas and studied the Miwok materials with the idea of publishing all of Kelly's field notes. We photocopied all of the materials, and they returned to California. They made a second trip later, and we remained in touch by phone and mail. After settling on a computer system and a word processing package, and using volunteer labor, they managed to convert all of Kelly's written Miwok materials into computer files, which followed the HRAF coding system imposed on the original field notebooks by Kelly herself. These files were then indexed and blended with Kelly's photographs and drawings (along with a brief biography of Kelly and a statement about the Kelly Archive) to produce a handsome large-format volume (Collier and Thalman 1991). As a result, the information on the Coast Miwok in the Kelly Archive is now available in a form unlike that of the original field notebooks, while these are preserved in Dallas. Eventually, the computer version of the Kelly materials on the Coast Miwok will be placed on a CD-ROM along with the rest of the Kelly Archive, which will be made available to interested scholars.

Scanning: The Weak Link in the System?

The scanning phase of data conversion involves a critical conceptual problem: the choice of *image-based* versus *text-based* technologies. The resolution used in scanning should be the minimum necessary to produce legible results when the material is output on screen or sent to a printer. Generally, resolutions of 300 dots-per-inch (dpi) are sufficient for the current generation of laser printers, and are more than sufficient for screen displays that do not output more than 100 dpi. The newest generation of scanners offers resolutions up to 600 dpi and up to 600 dpi and up to 256 "gray levels" or even full-color capabilities. The use of these higher resolutions will result in a significant increase in the document size for images but will make no difference for text materials converted into ASCII characters.

Image-Based Systems

In a sense, image-based scanning systems provide electronic "photocopies" of page after page of the original materials. The end-user can access these electronic copies through a display monitor and through an output device such as a laser printer.

A single 8.5" x 11" page, regardless of the mix of text and images, when represented by a scanned bit-mapped image at a resolution of 300 dots per inch typical of current laser printer output, results in a file of about one million bytes. Large page sizes and higher resolutions increase the file size of the scanned image accordingly. On the other hand, a page of typed text might contain only two kilobytes in a standard ASCII text file.

Scanning systems based on image-only conversion methods would not be feasible were it not for recent advances in compression-decompression software and hardware. Use of such schemes (e.g., Group III and Group IV Fax standards) reduces the size of a full page to as little as fifty kilobytes, depending on the mix of text and graphics on the page.

The advantage of image-based conversion systems is that the arrangement of text and graphics on the original page can be preserved — although a full-page monitor (at least 17" screen) is usually needed in order to view the entire page at a size that is legible. Image-based displays permit the end-user concerned with proper citation of an original document to determine the exact location of text and graphics from page to page. Furthermore, materials not readily converted into text (such as photographs, charts, graphs, maps, and even marginal notes written by hand) can be reproduced through image-based systems.

The major disadvantages of image-based conversion systems are: the relatively large size of image files compared to text files; the difficulty of indexing and retrieving image files; and their inability to manipulate directly the text within the image files (for example, extract a section and insert it into a separate computer file for word processing or data analysis).

Case Study #3

During 1988-1989, I was co-principal investigator for a pilot study to investigate the feasibility of CD-ROM technology for archaeological reports (Wendorf et al. 1990). This project involved a considerable amount of image-based scanning of published reports in contract archaeology. A major finding of the pilot study was that the greatest costs associated with scanning are human labor, not equipment. For instance, in order to take advantage of the automatic document feeder attached to the scanner, it was necessary to work with paper records of a consistent size and configuration (i.e., portrait or landscape mode). We also had to learn to use scanner settings for contrast and gray levels to achieve the best image within the lowest file size. In the case of bound documents, it proved necessary to disbind them and subsequently arrange for their rebinding.

Optical Character Recognition Systems

The still-developing field of optical character recognition (OCR) involves converting printed text to ASCII characters for on-screen display and printer output. This provides a 1:1 correspondence, within the limits imposed by the standard ASCII character set. Some scanners can recognize character attributes, such as "bold" or "italics," and retain this information for use by a supported word processing program. If the scanner cannot deal with character attributes, and if it is desirable to retain such attributes in the final copy, then additional keyboard data entry will be needed. Also, while some OCR programs can deal with Western foreign languages, none currently can handle the variety of phonetic transcriptions used by many anthropologists working in non-Western regions. (An international committee is currently working on a superset of ASCII that would include designated values for a number of non-Western languages, but even this expanded ASCII character set will be inadequate for some anthropological materials.) Beyond the limitations of the current ASCII character set or some future enhanced ASCII superset, we need a common document architecture that can also handle graphical elements in combination with complex text attributes.

A principal conceptual problem of using OCR-based systems to convert paper documents into electronic files is that most current OCR systems must separate text from graphics in order to perform their character recognition task. The result is the creation of a text file and a series of separate graphics files that must be linked together to provide the end-user with some semblance of the original document.

There is, of course, no need to perform OCR work on anthropological materials already in computerized formats. But until the millions of pages of backlog are cleared away, optical character recognition systems will likely play an important role in the building of anthropological information systems. Keyboard data entry is still competitive with scanning (especially with OCR-based systems) in the early 1990s. As the speed and accuracy of scanners continue to improve, and as the cost for high-end systems continues to drop, however, it is probable that scanning will soon become more cost effective for converting original

anthropological materials into a format appropriate for building information systems and for making such information widely available.

Data Storage Systems

A decade ago, a project to create a transportable archive of anthropological records would have focused on microfilm and microfiche technologies. Those technologies were image-oriented but involved separate cataloguing for each film or fiche image. Moreover, access was limited to sequential searches with microfilm, while microfiche offered the possibility of semi-random searches (that is, one could move to any image on the fiche rather easily).

In the 1990s, a wide range of mass storage systems is available for microcomputers. These include hard disks (ranging in size beyond one gigabyte, with fast access times and random search capabilities); cassette tape backup systems (ranging in size beyond 120 megabytes, with relatively slow access times and limited to sequential searching); digital audiotape systems (more than two gigabytes in size, but which suffer from relatively slow access times and are limited to sequential searching); magneto-optical disks, including Write Once, Read Many times (WORM) drives with multi-gigabyte capacity and random access capabilities, but with relatively slow access times; and various types of laser disks, including CD-ROM, CD-I (Compact Disk-Interactive), DVI (Digital Video Interactive), and multimedia CDs, all of which offer capacities of around 650 megabytes, random access capabilities, and moderate to slow access times. Recently, Kodak introduced the Photo CD system for storing 35mm slides. With 120 Slides stored in five different resolutions on each CD, this system promises to be an inexpensive way to protect and access slides, including old ones whose emulsions are likely to deteriorate much sooner than will the CDs onto which they are digitized.

Retrieving Information: Data versus Documents

Anthropologists combine sequential and random access methods as they peruse written reports, look through sets of photographs, or examine correspondence files. We are usually limited to sequential access methods when viewing films and videotapes or listening to audio recordings. With computer technologies, we can combine random and sequential access methods with virtually all materials. However, the key to developing successful information systems is the ability to retrieve data of interest to a wide variety of end-users. As Zoellick has stated, "An essential distinction must be made between retrieving documents and retrieving data" (1987: 63).

It is important that information placed in computer-based archives, including CD-ROM systems, include document retrieval capabilities. It is more difficult to include full-scale data retrieval, especially for information stored as bit-mapped images rather than as ASCII text. According to Zoellick:

> Document retrieval systems can be distinguished by whether they rely primarily on searching or on browsing. Searching begins with a term or set of terms that we believe occur in the documents of interest on a particular subject. . . . In browsing, we open the database to a particular document and then read what is there. . . . Another useful way of distinguishing between browsing and searching is to view electronic document retrieval as a matter of locating some particular content at a particular place in the document collection. The content is the what of the retrieval operation; the place is the where. Searching moves from what to where. We know the terms associated with what we want, and the system finds where they occur. Browsing moves in the opposite direction, from where to what. We open the database to a location and what is there (1987: 65ff.).

Text-retrieval "engines" process data-tags, build indexes, compress data, and even influence the user interface. Image-retrieval "engines" do the same for bit-mapped graphics files. A few commercial products provide both text and image retrieval within a single environment, but none yet provides the intuitive (and sometimes very inefficient) approach to information used by human readers. Current efforts to use "artificial intelligence" and "expert systems" to develop superior information retrieval methodologies are just reaching the marketplace.

In a provocative essay about the "dark side" of document-image processing, Christopher Locke argues that currently available indexing schemes are fundamentally flawed.

> The core issues here are epistemological: How do we know something? What constitutes knowledge? . . . [D]ocument-image processing presupposes that what is important in a document is already known. You type "This document is about x" into the index field. But that's today. What about tomorrow? You may want to turn to your information resources to mine a completely different type of ore. . . . If there is no retrieval hook, responses to queries will not return any documents. If searchers asking intelligent questions are not informed, events will follow their own course with no one the wiser. . . . You will always be caught short if you can't quickly restructure or reindex information in light of more recent intelligence. . . . Information resources, and especially document collections, are not just mountains of facts you already know but research bases to explore for clues about what you don't yet understand. More than that, they are the foundation on which new knowledge will be built by accretion (1991: 202, 204).

On the other hand, we always have the option of examining every electronic document just as we might peruse written documents. Indexing and retrieval software can help as we search and browse — just as the HRAF Outline of Cultural Materials codes can assist anthropologists in cross-cultural comparative analyses. But, as with the OCM (which can be considered a very simple keyword coding system), the end-user must be sensitive to the possibilities for discovering relationships among data not recognized by the original fieldworker or by the indexer.

Data Display

It would be desirable to reproduce exactly the original format of anthropological records for display on a microcomputer monitor. Even if this were feasible, few anthropologists presently own the requisite equipment (i.e., large monitors with high resolution adapters; at a minimum, a 15" screen with 1,024 x 768 Super-VGA color adapters with one megabyte of video memory), nor are they likely to afford such equipment in the future.

An alternate mode of displaying the information contained in anthropological records involves combining text mode for the character-based data and graphics mode for images such as line-art and half-tone and color illustrations. A related approach would offer the end-user the option of toggling between text and images.

The main conceptual problem of information display involves the ways in which we present information in the established format of "hard copy" documents (such as field journals, photographs, and correspondence) versus how we present information in the computer environment. For instance, most academic journals to which anthropologists submit manuscripts require that tables, charts, and other illustrations be separated from the main body of text. Then, in the final page proofs, the author sees — for the first time — the proposed physical layout of the article as it will seen by readers. Now that it is possible to "tag" images so that they can be linked interactively to specific texts, it is no longer necessary to integrate physically text and images on the same electronic "page." This connection is usually accomplished through what are known as "hypertext" functions in the software program.

The visual quality of computer displays is almost always inferior to the printed page and will likely remain so, except for very expensive large-screen, high-resolution monitors. The resolution of offset printing may run from 300 dpi to 2,400 dpi, whereas the resolution of monitors varies from about 50 dpi to 100 dpi. The computer display is less capable of revealing important details, but compensates for this deficiency by permitting zooming and other manipulation of screen images and text.

A significant issue regarding data display is the "aspect ratio" at which images are shown on screen. Monitors with CGA (Color Graphics Adapter) or EGA (Enhanced Graphics Adapter) output do not give a "true" 1:1 aspect ratio of the original printed image. However, VGA and Super VGA standards, which offer a 1:1 aspect ratio, are now widespread and relatively inexpensive. Thus, VGA or Super VGA monitors and adapter cards must be used in order to ensure that illustrations (such as artifact drawings) displayed on screen match those found in hard-copy documents.

Data Manipulation

Not only do we want to retrieve information; we also want to manipulate it to our own purposes. For example, we might wish to extract data from the text, tables, or illustrations and then convert them into a form

acceptable to our word processing, spreadsheet, data-base, or statistical analysis packages. For the last two decades or so, the usual way to accomplish these tasks of data manipulation involved hand copying or photocopying followed by manual data entry before analysis could begin. At present, computer keyboard data entry or scanning offer better strategies for data manipulation. Even where scanning is limited to image-oriented software, subsequent users with OCR software can extract text from the computer images and then move the secondary text files into their own word processing programs.

Case Study #4

During the 1960s and 1970s, George M. Foster laboriously hand-copied and photocopied the baptismal and marriage records in the parochial archives in the community of Tzintzuntzan, Michoacán, Mexico. These records are typical of the local archives now examined and analyzed by anthropologists. Over several years, Foster brought back to Berkeley, California an impressive set of field notes and documents based on the parochial archives. However, the thousands of records have been used mainly to check on demographic and genealogical information gathered from other sources (including the local civil register and informants' statements). Around 1980, Foster gave me a set of the carbon copies of these parochial and civil register records, but I paled at the cost involved in converting them into computer records suitable for analysis through mainframe computer programs. Even with the emergence of microcomputers, the cost of processing the information into data-base programs proved to be prohibitive.

In 1990, in the face of significant price cuts in microcomputers, I hit upon a more effective solution. I took two relatively inexpensive IBM PC-compatible laptop computers to the village. I hired and trained two local young women to do data entry directly from the parochial archives into a database program (the Spanish version of Microsoft Works, version 2.0). By the end of 1993, they had entered more than 35,000 individual baptismal records (dating back to the 1780s) into the laptop computers, and had nearly completed the related marriage and death records.

Not only are we preserving the information in the parochial archives, but we are doing so in a way that makes it much easier to study demographic and genealogical trends. In addition, I am creating a comprehensive index (computerized and hard-copy) for the use of the village priest and his assistants, who are often asked to provide copies of records to local residents.

Problems with Electronic Records

Increasingly, anthropologists are creating their original records — field notes, correspondence, grant proposals, teaching materials, etc. — on microcomputers rather than on paper. Most of us have suffered one or more hard disk "crashes" or have endured the failure of some hardware or software component at a critical moment. As a result, we have learned to back up our files on diskettes or tapes and to produce hard copies.

Unfortunately, few anthropologists have developed appropriate preservation strategies for their machine-based materials, including text files (e.g. e-mail messages) on the INTERNET and specialized graphics files on the world-wide web. At a minimum, we should print copies of important computer files on acid-free paper and store these documents in archival-quality folders and storage boxes. Also, the finders' aids for these archived documents should exist outside of the computer environment. For materials dependent on unusual programs or programming languages, it may be necessary to provide information about those programs and keep backup copies of the program diskettes and manuals in the archives. We must avoid the possibility that future scholars (or community members) cannot interpret our data because of hardware or software obsolescence.

A moment's reflection about the kinds of materials in our personal files shows that we need to develop some system of "triage" so that the most fragile and vital information gets preserved first, while other materials of lesser importance (or for which duplicates are known to exist elsewhere) are conserved later, if at all. Although we are likely to give first priority to original field data of all kinds, future historians of anthropology will also be interested in materials related to teaching, proposals both funded and unfunded, and diverse professional and organizational activities.

As a first step in determining priorities for preservation, a computer data-base management system that would identify the kinds of information to be preserved should be created. Then, questionnaires could be sent to relevant organizations, institutions, and individuals to find out what they are doing to preserve their records. A preliminary questionnaire of this kind was distributed by CoPAR in 1994, although processing the returns and establishing a satisfactory "lexicon" for dealing with the preservation of anthropological records remain challenging tasks.

A second step should be to require that plans for archiving research findings be included in grant proposals and in final reports to funding agencies. Just as contract archaeology projects have strict requirements for curation, the same could be applied to all anthropological research projects. Of course, it would be necessary to allocate funding to cover the appropriate costs of long-term curation (for example, photocopying of field materials onto acid-free paper as well as computerized cataloguing or storage).

A third step involves determining appropriate access to anthropological materials. What kinds of data should be preserved and then made available to scholars and to the public? How long should certain kinds of data be withheld from scrutiny? How do we provide adequate protection to informants? And who decides the answers to these questions — anthropologists, funding agencies, government bureaucrats, or representatives of the people studied? In many respects, computer-based data files can be made more secure, especially with respect to casual browsing, than their paper copy equivalents.

A final step involves the format in which anthropological materials are to be preserved. This brings us full circle to the issues raised at the beginning of this paper. Few anthropologists are likely to be willing to give up originals of their materials, and many would object (for diverse reasons) to depositing copies of their records in any archives beyond their immediate control. Perhaps the solution will be to create electronic catalogs of anthropological materials, which might — after the death of the individual anthropologist or after the passing of a certain number of years, depending on the kinds of data involved — then be accessioned into designated archives. The creation of such a database might be the first appropriate use of computer technology in preserving the anthropological record, but it will certainly not be the last.

Summary

- Computers will play increasingly significant roles in cataloguing collections of anthropological materials.

- Although converting paper-based documents to electronic media can be complex and costly, enhanced data manipulation capabilities often justify the effort involved in building anthropological information systems.

- The useful life of electronic storage media is still open to debate, especially for the extreme environments in which anthropologists often do fieldwork.

- Information created in electronic media should also be stored in alternative materials, including acid-free paper versions destined for archival preservation.