Database Computerization and Consortium Development for Vertebrate Collections – A Collection Management Perspective

Robert D. Owen

The Museum, Texas Tech University, Lubbock, Texas 79409, U.S.A.

Owen, R. D., 1990. Database computerization and consortium development for vertebrate collections – a collection management perspective. *In:* Herholdt, E. M., ed., *Natural history collections: their management and value,* pp. 105–116. Transvaal Museum Special Publication No. 1, Transvaal Museum, Pretoria.

It is imperative, as collection growth and maintenance become legally, financially, and politically more difficult, to take advantage of advancing technology to consolidate the extensive resources represented in natural history collections. Computerization of collection databases should be viewed strictly as a means of facilitating collection management. Computerization should be oriented toward enabling us more quickly and accurately to answer three simple questions concerning a specimen: 'What is it?', 'Where did it come from?', and 'Where is it now?'. The computer database should reflect the current status of the collection, and does not duplicate the function of the permanent written catalogue. From these points, it follows that not all catalogue or specimen information is appropriate for the computer database. Generally, those information fields that can be searched and sorted meaningfully are useful, and others are wasteful of disk or tape space. Numerous combinations of hardware and software are available for database computerization. An increasingly important criterion for selection is communication – with other institutions, with a central data base, or with one's own ancillary collection files. Therefore, the machine should have modem or hardwire access to telephone or other communication services, and the software should be capable of sending, receiving, and interpreting standard ASCII files. Intermuseum networking or consortium arrangements will amplify the usefulness of collection computerization. Multimuseum databases will allow investigators to determine more accurately and handily what specimens are available for a study, and where they are. It is not yet clear whether a network arrangement or a central database consortium will ultimately prove better. This may vary from one region to another, depending at least in part upon available network technology and host machine compatibilities. In either case, an essential factor will be the willingness of potential member institutions to coordinate with each other concerning their computerization protocols.

OVERVIEW

It is imperative, as collection growth and maintenance become legally, financially, and politically more difficult, to take advantage of advancing technology to consolidate the extensive resources represented in natural history collections. A number of concerns are brought to our attention in connection with the growing number of collections undergoing computerization, and by the imminent possibility of electronic information sharing or exchange. However, we should clearly see that there are no problems or questions confronting us that are fundamentally new. Rather, new emphasis

and urgency are being created concerning old questions and problems that traditionally have been addressed on an intuitive and *ad hoc* basis.

We would all agree that a research collection is only as useful as the information associated with the specimens in it. What may be less generally understood is that there are two essentially different uses for this information – collection management, and scientific research. Furthermore, the particular use for which the information is intended has fundamental implications concerning the responsibilities of the curator (or collection manager) and the user to each other and to the veracity of the information. The purposes of this paper are to emphasize the distinction between the two types of information, and to outline general standards for administration of collection management information.

Information intended for collection management is that which answers one of three questions concerning a specimen: 1) Where did it come from? – the provenance, or collecting data, 2) What is it? – the identity, or taxonomic designation, and 3) Where is it now? – the location in the collection. The third answer may be partly implicit in the taxonomic designation, but also is dependent on whether the specimen is dry or fluid-preserved, where it is in the preparation process, and whether it is on loan. The answers to the three questions listed above are the minimal data necessary and sufficient to guide a qualified user to exactly those specimens in the collection that will be useful to his or her research. The obligation of the collection manager is to provide this information in an accessible and efficient manner to the qualified user, and to make the specimens available to the user for examination.

In keeping with the recognized mission of acquisition and dissemination of scholarly biological information, the staff of mammalian research collections are expected to: 1) preserve samples of the world's mammalian fauna for biological research, (2) ensure that these collections are managed by a professionally trained curator or collection manager, or both, (3) not knowingly install specimens that are accompanied by erroneous information unless that information is clearly marked as such, (4) make available to all qualified investigators the specimens and associated information, and (5) prevent irresponsible access to, or use of, information associated with specimens. It should be emphasized here that, as is traditionally the case, ultimate authority and control of the specimens and associated information reside entirely with the curator and governing body in charge of the collection. The collection manager functions in collaboration with the curator in order to carry out many of the functions of acquisition, registration, maintenance, and organization of the collection with the ultimate goal of providing access to the specimens and associated information for qualified investigators. A qualified investigator is one who has the training or experience to properly evaluate specimen information, to identify that which is novel, and to assess the likelihood of its truth. Of course, either the curator or the collection manager may function as a researcher also; the intent here is to clarify the functions and responsibilities of their roles with respect to the collection and its users.

The curator and collection manager, by virtue of their training, are expected to judge whether there is a potential for irresponsibility. Such a case might, for example, be the unquestioning acceptance of a specimen identification or extralimital collection locality; publication of such a record, if in fact erroneous, constitutes propagation of misinformation and retards, rather than advances, the state of biological knowledge.

The user has the obligation to conduct his or her research in a way that benefits not only the field of scientific knowledge, but also the collections from which he or she has obtained the information. Part of this responsibility is recognition of the potential of inaccuracy of information associated with the specimens, and the need to continually verify, correct, and augment this information. Minimally, the researcher should: 1) formally acknowledge specimens or other information provided by a collection, 2) inform the collection manager of any data errors, nomenclatorial changes, or additional information that may pertain to the specimens used, and 3) make available to the curator reprints (or at least complete citations) of articles that are based upon specimens and information from a collection.

INFORMATION MANAGEMENT

As mentioned above, the increasing trend toward computerization of collection data has seemingly introduced a number of new administrative ques-

tions to be dealt with by the systematics collections community. I would contend that, although new practical problems inevitably will occur, no fundamentally new questions have arisen from the advent of accessible computer technology. No new kinds of information have been created; rather, the distinction between research and management data has been accentuated, and the relative roles of the collection manager, curator, and researcher are being reexamined and becoming more precisely defined.

An integral aspect of the collection manager's activity is the creation, maintenance, and updating of the collection catalogue and of the collection information computer file. It should be clear that the computer file does not replace the written catalogue. The written catalogue is an historical document that may include information concerning changes in the status of a specimen (e.g., nomenclatorial, identification, disposition). The computer catalogue, however, is strictly a collection management tool, carrying only the present status of specimens that are or have been part of the catalogued research collection. As such, it can substantially enhance the collection manager's ability to provide collection information accurately and rapidly to qualified users. The utility of these electronic files will increase dramatically as ancillary collections (e.g., frozen tissues, anatomical preparations, living tissue cultures) become available as cross-indexed files, and as collection computer files from more than one institution become networked so that these ancillary materials (which often are not housed in the same institution as their associated specimen) may be identified and located with considerable facility and with confidence that none have been overlooked.

COMPUTERIZATION

In planning computerization of a museum collection, there are three fundamental steps required before any equipment is bought or any data entered. These are: 1) systems analysis, 2) choice of software, including operating system, and 3) choice of hardware, including peripherals. Systems analysis involves the clarification of what the overall management plan is for the collection, and where computerization fits into that plan. It involves clarification of the expected result of the computerization effort, and how most expediently to achieve that result. It involves planning the file management system and the file structure for best utility, and determining the best method for transferring the data from their present state (e.g., catalogue, specimen tags) to their desired electronic form.

Choice of operating system, file management software, and data retrieval software, may be determined by existing equipment, but the planner will be in a better situation if this is not the case. These choices are best determined by the results of the systems analysis, and should in turn determine the best hardware for the particular collection management system. The hardware should be the last major decision, and should be selected because it is designed for the desired operating system and software, and because it has sufficient CPU speed and size, sufficient disk speed and size, and the correct peripheral equipment to get the job done in an efficient and labour-conservative manner.

SYSTEMS ANALYSIS – FILES AND FILE STRUCTURE

Computerization of collection databases should be viewed strictly as a means of facilitating collection management. The computer database should reflect the current status of the collection, and does not duplicate the function of the permanent written catalogue. From this, it follows that not all catalogue or specimen information is appropriate for the computer database. Generally, those information fields that can be searched and sorted meaningfully are useful, and others are wasteful of disk or tape space.

The database system should have certain attributes for optimum utility in collection management use. First, it must be amenable to periodic expansion, or addition of new records. It should be capable of accepting these records by keyboard entry, as new specimens are catalogued. It should also be able to accept a data set from an external file, such as by tape or other electronic means.

Second, the file should be easy to edit. Two types of editing should both be easily accomplished. When errors in the files are found, these should be simple to locate and correct on a case-by-case basis. Where major changes are needed, such as a species being moved to a different genus or a country changing names, this should

be easy and relatively foolproof with a global change option.

Third, although this is debatable, the database system should accept alteration of the names of the fields or the number of fields represented in the specimen records. This attribute is debatable because we should at least believe that we must make (and have made) the correct decision concerning the appropriate fields when establishing the file structure initially. If we are changing the file structure very much or very often, we have not thought through the purpose of the file, and we certainly are endangering the integrity of the data as well.

The structure of the data is critical to the success of a computerization effort. Again, appropriate data fields are those that can be searched or sorted on. These fields can be thought of as belonging to three general categories, corresponding at least roughly to the three types of information described above as pertinent to collection management: provenance, identity, and location. Provenance includes geographic and date information. Identity includes sex and taxonomic information. Location is determined by the cataloguing and 'nature of specimen' fields, as well as the taxonomic identification of the specimen. The recommended fields are listed here under those general and specific category headings:

- Provenance
  - Geographic
    - Country (Ocean)
    - State (Province, District)
    - County (Municipality)
    - ?Specific locality
    - ?Map grid coordinate
  - Date
    - Year
    - Month
    - Day
- Identity
    - Sex
  - Taxonomic information
    - Taxonomic code (numeric)
    - ?Family
    - Genus
    - Species
    - ?Subspecies
- Location
  - Cataloguing
    - Catalogue number
    - ?Accession number
    - ?Special number (e.g., frozen tissue number)
    - ?Endangered, protected, or type status
    - Collector name
    - Preparator name
    - Preparator number
  - Nature of specimen (see text below for explanation)
    - Skin
    - Skull
    - Postcranial skeleton
    - Fluid-preserved
    - Hide or mounted specimen
    - Other

Fields preceded by a question mark are optional, depending on circumstances and needs of the particular collection.

Geographic information fields that are necessary include three levels of designation. The largest is country, or in the case of Antarctica or international waters, the continent or ocean designation. The intermediate is state, province, or district, and the smallest is county, parish, or municipality. These are sufficient for the collections that are arranged alphabetically by political units. If, on the other hand, specimens are arranged 'geographically' by political units (e.g., eastern hemisphere, south-to-north), then a numeric code, similar to the taxonomic code described below, is needed.

Specific locality is not useful generally in specimen data files, and often is quite space consuming. Map grid coordinates are most useful in collections containing a number of marine specimens. These two information fields should be considered optional, depending on the needs of the particular collection.

The three elements of the date of collection should occupy separate fields, unless the database system in use is one that includes a 'date' field type. It is quite likely for instance, that for reproductive or migratory studies a researcher would want the specimens sorted by month and date, but not by year. The month, in order to be sortable, should be stored as a numeric field. To avoid input errors as well as difficulties in reading reports, the months should be input and output as

three-letter acronyms, and transformed to the numeric field by means of a simple translation program.

The sex of the specimen should be designated by three mnemonic codes. 'F' for female, 'M' for male, and 'U' for unknown are standardly used for collections in English-speaking countries.

Minimal taxonomic information includes a numerical taxonomic code, the genus, and the species name of the specimen. For mammals and other vertebrates, I suggest a six-digit taxonomic code, corresponding to the taxonomic system by which specimens are arranged in your collection. The first two digits indicate the ordinal classification, the second two are the family within the order, and the third two refer to the particular subfamily. The taxonomic code preferably is generated automatically by an on-line dictionary of genera as records are added to the master file. This would of course save some time in data entry. More importantly, it would preclude the errors that would be likely to occur if the data entry person were to look up the appropriate code with each entry. By this simple code, the records can be sorted taxonomically to the subfamily level. The code should extend to the level at which the specimens are taxonomically arranged within the collection. The real criterion is that you should be able to produce a printout that exactly replicates the specimen arrangement in the collection. The taxonomic code typically is not part of a report output. It is primarily for internal use, with reports using the accepted familial and generic names.

Additional taxonomic categories that may be helpful include family and subspecies information. Neither of these is required to replicate the correction arrangement (unless specimens are arranged to subspecies), but they may be useful in answering certain types of information requests. The family name, like the taxonomic code, should be generated automatically from a dictionary, and stored as part of the data file.

The catalogue number certainly is the single most important piece of information associated with the specimen, with or without regard to a computer file. This number ties the specimen to all of its associated information. In museums or collections that maintain a separate accession record, the accession number ties the specimen and its information to the acquisition records. Another number, here referred to as the 'special number', associates the specimens with an ancillary catalogue, such as a frozen tissue catalogue, which in many institutions is maintained separately. In such cases, it is imperative that the tissues and their source specimens can be associated positively and quickly, from either direction.

An additional field of potential value in collection management is the endangered or protected status of the taxon. It is important to be able to flag such specimens for yourself and for potential borrowers, in order to avoid inadvertently violating the numerous national and international regulations that pertain to them. An additional code may be utilized in this field to designate other special or restricted status, such as that of a type specimen.

The collector's name, preparator's name, and preparator's number all should be included as information fields in the file. Names, in order to be sortable, should be last name first, and consistently follow a convention concerning spaces, punctuation, use of initials for given names, and so on.

The 'nature of specimen' information should consist of several fields, each pertaining to a type of specimen that the institution normally would catalogue and install. For example, in mammals, suggested fields would be skin, skull, post-cranial skeleton, alcoholic, hide or mounted specimen, and 'other'. The purpose in maintaining these as separate fields, rather than a single coded field, again relates directly to the fundamental purpose of the computer file-collection management. Each of these fields, rather than carrying codes signifying presence or absence only, will contain information concerning the current status of a particular portion of the specimen. For example, the skin may already be installed (and even on loan), while the skull and the skeleton are still waiting to be cleaned. Again, this list should reflect your institution's procedures. The list should not be overly complicated, but should allow the curator or collection manager to determine the whereabouts of each part of the specimen. A real advantage of this is that it alleviates the need for massive 'hold-up' areas, where entire collections are held until completely processed, for fear of losing specimens.

As recommended above, the 'synthesized' information fields (taxonomic code, family name, and numeric code for month) should be generated during input (or during transfer of temporary data

to the main file) and stored as permanent records with the file. Although they could be generated each time the records are searched, this would save only minimal disk space, and would substantially increase running time for most search-and-sort routines.

SYSTEMS ANALYSIS – DATA CAPTURE (CURRENT AND RETROSPECTIVE)

The plan for data capture should be firmly decided before beginning, especially for retrospective capture, which virtually all collections will require. If the data are already in an electronic form, then the decision (although not necessarily the task) is easy. If the collection is in fact just beginning computerization, this may be the most crucial single decision made. Two important decisions must be made: 1) one pass or multiple passes through the collection, and 2) data capture from the catalogue or from the specimens themselves. The following recommendation is dependent on an adequate commitment from the collection staff and administration to complete a job in a planned and reasonable length of time, rather than 'whenever possible'. This is not meant derogatorily towards those collections that are unable or unwilling to make such a commitment. I would point out, though, that this data capture plan is, in the long run, probably optimal, in terms both of time spent and of producing an error-free data file.

It is recommended that the data be captured in a single pass through the information. Clearly, this will require a longer preliminary period during which the file is only minimally useful. The alternative of making a first pass in which only catalogue number and taxonomic information is entered, will produce a collection inventory, and enable responses to some information requests, at an earlier time. However, a single pass through the information is, in the long run, the less time- consuming method. It is further recommended that the information be captured from the written collection catalogue, rather than directly from the specimen tags. Although a bit counterintuitive, this procedure will ultimately engender a virtually error-free system. The computer database, when finished and corrected as far as possible without checking against the specimens, is sorted in collection order, and then checked against the specimens and their tags. This procedure thus is simultaneously checking for discrepancies between the computer file and the specimen tags, between the specimen tags and the written catalogue, and between the written catalogue and the computer file. Any discrepancy among any of these three repositories of information will be detected in this verification pass. Additional advantages of this procedure are reduction both of specimen handling and of total time expended. The specimens need not be brought, tray by tray, to the computer terminal, and tags turned while trying to enter data. Additionally, the data are more likely to be available in a consistent format in the catalogue than on the specimen tags, again minimizing entry time and error rate, and maximizing compliance with data standards. The printout of the sorted file, in contrast, can easily be carried around the collection for verification against the specimen tags.

Current, or non-retrospective, data capture should be an integral part of the cataloguing process, and again should be accomplished from the written catalogue and verified against the specimen tags before the specimens are installed.

SOFTWARE AND HARDWARE

Numerous combinations of hardware and software are available for database computerization. An increasingly important criterion for selection is communication – with other institutions, with a central database, or with one's own ancillary collection files. This should be kept in mind during any consideration of software and hardware acquisition. Minimally, the machine should have modem or hardwire access to telephone or other communication services, and the software should be capable of sending, receiving, and interpreting ASCII or other standard files.

This discussion will be fairly general. I recommend the article by D. F. Williams (1987) as an excellent and detailed discussion of issues in software and hardware selection. There are, however, several considerations of which I would emphasize the importance, concerning software and hardware selection. The database management system should be capable of direct access to file records, as opposed to sequential access, and should allow indexing of most or all fields. Search routines should include all Booleon logical opera-

tors (and, or, not) and combinations of them. Sorting routines should be flexible, allowing sorts on one or more fields in user-designated order. Output routines should also be flexible.

Data entry and editing are clearly the most costly and time-consuming of tasks associated with computerization. Software that expedites these processes is certainly critical. The data entry software should be such that persons inexperienced in biology or museum work can efficiently perform the task. A user-defined full-screen template, representing an entire record, should be available, to match the entry order to the written record from which the information is being taken. The entry person should not be scanning back and forth to find the next field to be entered. The software should provide optional cursor movement back to previous fields on the screen for correction before transferring the record to the file. Field entries should be automatically retained from one record to the next, thereby expediting entry of series of specimens of the same taxon, from the same locality, by the same collector, etc. The catalogue number should be automatically incremented from one record to the next.

I strongly recommend a two-stage process for data entry. The first stage is as just described, with the records being stored in a temporary file. The second is a verification and editing process performed on this temporary file. A useful routine for this, for instance, is to sort (alphabetically) and list all genera, species, countries, states, and counties found in the file. These lists are then printed, and can be quickly scanned for typographical errors or misspellings, which tend to be obvious in such a listing. These are the most critical variables, as most collections are sorted taxonomically to some specified level, then geographically. It is imperative, therefore, that they be correct so that a properly sorted listing will show each record in its proper place in the collection sequence. Any mistakes detected at this stage are easily correctable with global changes in the temporary data set.

Software should also be available that creates the synthesized fields (e.g., taxonomic code, family name, month code) at the time the temporary file is transferred to be appended to the permanent master file. Backup facilities must also be available, and used whenever updates are made on the master file. Host communications software should be available if the hardware chosen is a microcomputer system.

An important consideration in software selection that is not often discussed is that the log of the user's actions must be easily interpretable. The user must, in other words, be made immediately aware of an unsuccessful file action so that it can be remedied before it is compounded. There is a case known in which the log from an update on a SELGEM file was sufficiently difficult to interpret that the user made several sequential backups and modifications of the file not knowing that it had been destroyed by an unsuccessful edit process. Over 20 000 records were completely lost.

As mentioned above, the much preferable situation is one in which the hardware selection is dictated by software choice, rather than the reverse. Beyond that, it is clear that when storing very large files, and conducting large searches and sorts, the money spent on a large, fast central processing unit, a large, fast disk drive, and a fast, durable printer, is money well spent. I would also emphasize again that communications capabilities are becoming imperative, and a microcomputer-based system should have a telephone modem or dedicated phone line connection to a host mainframe system that in turn has network access to other pertinent hosts.

Although the present trend is toward microcomputer systems, there still are arguments for mainframe implementation of collection computerization. Processor speed, network access, and mail facilities are prominent among these. Equally important is the fact that the arguments against mainframe use are declining as these systems become more versatile and flexible, and installations, particularly in academic environments, have become more oriented towards interactive use and local area networks. As an example, if access to a VAX system is available, I would recommend a hybrid system involving use of a microcomputer for data entry and initial editing. This machine would then be used as a smart terminal to transfer the temporary file to the VAX mainframe, where the synthesized fields would be added, and the temporary file added to the master. Searches and sorts would be done on the mainframe, which will almost certainly have output access to a faster printer than does the in-house microcomputer. Alternatively, smaller output files can be downloaded to the microcomputer for local printing.

SECURITY

We are all concerned with management of collection information in such a way as to facilitate use and to preclude misuse. Growing computerization of collections, including ancillary collections, and the likelihood of database sharing or exchange in the near future will vastly facilitate the use of this information. There is substantial concern among some that it will at the same time facilitate misuse. I would take the position that, by adhering to principles previously established concerning the governance of collections, and by adequate caution exercised by properly trained collection managers, likelihood of misuse should not increase.

Because the specimens and information housed in public institutions (e.g., state universities, state and national museums) are public property, concern often is expressed that widely shared computer files will encourage irresponsible public access to the data. It is clear, though, that the public ownership of the collections does not imply unrestricted access to them, any more than we have unrestricted use of national parks and wildlife preserves, or of police cars. The public has, in some sense, elected not only to acquire and house these specimens and information, but further, to hire appropriately trained professionals to manage them and to protect both the specimens and the information from misuse. The primary issue is management of collection information so as to facilitate use and preclude misuse.

I believe that we have little to worry about concerning information misuse by research professionals associated with the museum community. Experience has shown that this community is too small and well-informed, and the sanctions too severe, for misuse to occur at more than a negligible rate. Again, the advent of computer databases does not present a new set of problems in this regard, but only prompts us to restate our collection management standards and our belief in their appropriateness.

Misuse by persons outside of the museum and specimen-research communities may pose an increasing threat currently. This is not so much due to the advent of computer database as to the need for governmental and private agencies to compile faunal inventories for management plans, environmental assessments, and impact statements. Often these are contracted to the lowest bidding company or agency, and the subsequent pressure to produce a product quickly and cheaply leaves the investigators less than adequately concerned with the veracity of their information. This can very easily result not only in inadequate information, but also misinformation, being represented and treated as fact.

Security against this type and source of misuse can be thought of as comprising two types. Pre-access security is that which prevents potential misusers from obtaining the information. Post-access security is that which either prevents them from communicating unverified data or poorly considered interpretations to other potential misusers; or, less desirable but perhaps effective, imposes post-misuse sanctions on the offender.

Pre-access security

The goal of any information system is to make data readily available to the users. However, measures will necessarily be invoked to protect against unauthorized access. Unauthorized access to the computer system will be guarded by the keyword entry. The software can be designed to allow a limited number of chances to enter the correct keyword and will prohibit access from the remote institution until manually unlocked by a system manager if successive errors are entered. Periodic changes of keywords will further safeguard against unauthorized entry.

All records in a file should be available on a 'read-only' basis. Data will thereby be protected from inadvertent or purposeful deletion or alteration except by a system manager. Additionally, a complete backup file will be maintained on magnetic tape housed separately from the main computer and disk file.

Post-access security

An area of considerable concern regarding data protection involves the irresponsible publication of records by unauthorized persons or agencies. Although a primary purpose of the database is to promote research activity, the curators must retain control governing the responsible publication of data from their institutions. One means of post-access protection is the publication of annual reports by each institution, thereby providing copyright protection for the published information, and

enabling litigation in the event of unauthorized or irresponsible dissemination of collection information.

DATABASE SHARING

Intermuseum networking or consortium arrangements will amplify the usefulness of collection computerization. Multimuseum databases will allow investigators to determine more accurately and handily what specimens are available for a study, and where they are. It is not yet clear whether a network arrangement or a central database consortium will ultimately prove better. This may vary from one region to another, depending at least in part upon available network technology and host machine compatibilities. In either case, an essential factor will be the willingness of potential member institutions to coordinate with each other concerning their computerization protocols.

Network or consortium

A true network involves a number of essentially co-equal host machines and data banks, each having continuous or continuously available communication with one another. Such a network dedicated to intermuseum communications probably is not feasible at the present time because: 1) the diversity of computer systems make communications complex, 2) the inconsistency of formats for recorded information would require a diversity of interpretive software, and 3) the resolution of the above problems and establishment of leased dedicated telephone lines or satellite communications would be cost prohibitive. However, such a network is unnecessary. An alternative, achievable approach is a consortium with centralized host facilities and interactive access by member institutions. The establishment of a computer consortium would be a cost-effective, efficient means of information sharing among museum collections. A very important aspect of this is that membership in the consortium will enable smaller museums to computerize their collections without local access to a mainframe computer or large amounts of disk space. This clearly is advantageous to them, and it also is advantageous to everyone else to have greater access to information about these lesser-known collections.

Consortium development

There are four basic aspects of the development of a consortium: 1) documentation standards must be established, 2) security must be assured for protection of the information, 3) data must be converted for storage in the host system, and 4) communications must be established between members and the host. The first point, documentation standards, was discussed in the preceding section on files and file structure. I believe that any natural history collection staff that is computerizing should do so with networking or consortium membership firmly in mind; thus these documentation guidelines are equally applicable to in-house computerization and consortium planning.

The second point, security, is developed from the basic ideas mentioned in the section on 'pre-access security'. Personnel at each institution who have contributed data to a shared database will have access to all records installed. Personnel at non-member institutions will have access to the data bank only through intervention by a member institution. Persons not affiliated with an academic institution (for example, private industry) will only be able to obtain records subsequent to written approval by the curators in charge of the institution from which records will be searched. The reason for the latter two stipulations is to promote broad-range data access while maintaining curatorial control. Of course, each contributing institution may wish to reserve some specimen records from general availability on the computer system. Completeness of each data file is at the discretion of the contributing institution.

The third aspect of consortium development is a one-time conversion of the member institution's data to the host format. This does not mean that all members must change their data capture procedures or storage formats. It only means that the data are transmitted to the host institution and converted for storage there. Transmission can be in a number of ways, such as sending a tape, over telephone lines, or by existing network connection. Virtually any data retrieval system that can print a data set can write an electronic version of the file in ASCII or another standard code. These can be converted with minimal difficulty to the desired host format.

The fourth aspect, communications, is the area in which a network and a consortium contrast most

strongly, and these considerations require greater elaboration. The hardware and software necessary to implement all organizational elements need only be present in the host computers (regional centres). The member institutions must only be able to initiate communications with a regional centre. Therefore, members need only a terminal, a high-quality modem, and a printer to fully utilize the data-sharing system. Members would then be able to connect to the host computer, make requests, and receive and print output during a single session. Greater flexibility and speed would be achieved by the use of a microcomputer rather than a terminal. Information may be quickly received from the host and temporarily stored on a hard or floppy disk and printed after the connection to the host has been terminated. Most major microcomputer companies have communication packages available to provide efficient connection to host computers.

The presence of microcomputers in smaller member institutions would provide an effective means for museums to computerize their collections without an in-house mainframe computer. Large packets of catalogue information can be stored on either hard or floppy disks and transmitted in full to the host computer via a modem connection. This process would alleviate costly connect-time to the mainframe computer during the critical period of data entry. Member institutions that have not yet undergone computerization because of lack of adequate facilities may be able to efficiently implement automatic data processing of specimen information.

Two mechanisms of output retrieval should be available to requesting members. The first method is to transmit a request and remain on-line until processing is complete and the host computer transmits output back to the remote terminal for printing. This may be an expensive endeavour for a lengthy or complex request. A second, and perhaps more expedient, method is to have output stored in a temporary file. Members may therefore transmit a request to the host computer and request output filing. The member may then break the connection and re-dial after an appropriate time has transpired for completion of the data processing. The member may then download the output file to a printer or disk system. Alternatively, if the member is working on a mainframe with BITNET or other network access, the transaction might be accomplished using this network, with the host simply sending the output file to the member as soon as it is ready.

Pilot programme

A necessary first step for the successful initiation of a museum computer consortium is to have a limited pilot programme. Three or four institutions involved in the computerization process would be selected to participate in a joint data-accumulation effort. The control facility will need to have equipment sufficient for the storage of all computer records, backup copies of the records, and software for searching and sorting the records. This host institution will need to have the following equipment available: 1) a mainframe or mid-range computer, 2) a tape drive for backups and for entering data from the member institutions, 3) large hard disk drives for rapid access to consortium data, 4) communications modems for dialing into mainframe for long-distance access, and 5) a line printer for rapid hardcopy of consortium records. Equipment needs for the member institutions were described above.

Consortium members could compile their records on a standard format tape and mail it to the host computer centre. The records would then be transformed to a standard format and stored on the hard disk system in a common file with data from other pilot programme members. Each institution may then connect to the host computer by modem over conventional telephone lines or by existing network facilities. Searches of records may then be made such that all or part of the computerized information is utilized. The pilot programme will continue until the feasibility and expediency of all hardware and software systems have been demonstrated.

Regional consortium

Subsequent to the pilot programme, a regional consortium of numerous members linked to a single host computer will be initiated. This step in consortium development represents additional institutions that wish to participate in an information sharing system. Hence, broad-based searches, electronic museum loans, and electronic mail may be accomplished.

The growth of a regional centre is not without

limits. The limiting factor for the number of records that can be efficiently stored and searched depends on the size of the hard disk system. An alternative that has recently become economically feasible is the read-only optical disk. These may be imprinted with data from tapes by a contracting company, and sent to the host institution for installation on the disk reader. They offer a substantial saving in disk search and read time, as well as intrinsic protection against alteration or erasure.

National or multinational consortium

Because the space requirements for data storage will exceed the capacity of a regional centre, the establishment of additional centres will be necessary. The new regional centres would require the same equipment facilities as described in the previous section and would function in a fashion similar to the initial regional centre. However, further features are needed to establish the appropriate communication capabilities between regional centres. For example, if a user communicating with one regional centre computer requires data stored at another regional centre, then the request must be transmitted and data retrieved from the alternate host. This system requires an additional modem on all of the mainframe computers (or again, access to a common existing network facility). An additional software package is also required to automatically link to the remote host.

The placement of regional centres is dependent upon the distribution of specimens across the area being served, as well as the capabilities of the institutions to function as regional centres. It is anticipated that about four regional centres ultimately would be adequate to complete the consortium for mammal collections in North America (including Canada and Mexico). At that level, the consortium could have readily accessible records of approximately six million specimens. However, costs for conventional telephone communications increase significantly for international links. Satellite communications are technologically possible and provide rapid access to distant centres, but they, too, are cost prohibitive. Networks such as BITNET are presently accessible in North America, Japan and Europe. Requests and data retrieval may be processed to and from remote centres over these existing network channels. Presently, there is no network system available to institutions on a worldwide basis. At the time that initiation of an international consortium is begun, all of the above communications possibilities will need to be considered for effectiveness and expediency.

Reiteration

Although full utilization of the technology available for true networks is not advocated, a consortium with intermittent communication among regional centre computers is suggested. This will serve the purpose of the research community in a similar fashion, and will be more cost- effective than other alternatives presently available. Consortium development is presently feasible and has the advantages of: 1) facilitating specimen-based research, 2) enabling broad-range searches and cross-referencing of specimen information, including collections and even specimens divided among more than one institution, 3) enhancing intermuseum loan processing, 4) reducing institutional workload, and 5) enabling small museums to participate more fully in the benefits of computerization, and thereby be more fully utilized by researchers.

ACKNOWLEDGEMENTS

I believe that a few ideas presented here may be original, but I emphasize that most are not. I have drawn liberally from the articles listed in the References, and particularly from discussions with several associates: Ronald K. Chesser, Robert B. Tucker, Janis Files, and Suzanne B. McLaren. Robert J. Baker, Clyde Jones, and Gary Edson have been continually supportive of our computerization efforts in The Museum, Texas Tech University. June Logan did the word-processing, and Joaquin Arroyo-Cabrales, Richard R. Monk and Suzanne B. McLaren provided critical reviews of the manuscript. Elizabeth M. Herholdt and C. K. Brain of the Transvaal Museum arranged for my

participation in this Symposium, and a Fellowship from the Foundation for Research Development of the Council for Scientific and Industrial Research, Pretoria, provided support.

REFERENCES

CATO, P. S. and FOLSE, L. J., 1985. A microcomputer/mainframe hybrid system for computerizing specimen data. *Curator* **28**: 105–116.

CHESSER, R. K. and OWEN, R. D., 1987. Computerized information consortium. *In:* GENOWAYS, H. H., JONES, C. and ROSSOLIMO, O. L., eds, *Mammal collection management,* pp. 145–154. Texas Tech University Press, Lubbock.

COMMITTEE ON INFORMATION RETRIEVAL, 1985. *Survey report on computerized information retrieval in mammal collections of North America.* American Society of Mammalogists.

COMMITTEE ON INFORMATION RETRIEVAL, 1988. Guidelines for usage of computer-based collection data. *Journal of Mammalogy* **69**: 217–218.

FOLSE, L. J. and CATO, P. S., 1985. Software needs for collection management. *Curator* **28**: 97–104.

FOLSE, L. J., CATO, P. S. and SCHMIDLY, D. J., 1987. Hybrid computer system at Texas A & M University. *In:* GENOWAYS, H. H., JONES, C. and ROSSOLIMO, O. L., eds, *Mammal collection management,* pp. 135–143. Texas Tech University Press, Lubbock.

LEBLANC, S., 1987. The first rule. *Spectra* **14**(1): 1–3.

MCLAREN, S. B., 1984. Application of automatic data processing in recent mammal collections. *Scientific Software Quarterly* Summer 1984: 7–13.

McLAREN, S. B., GENOWAYS, H. H. and SCHLITTER, D. A., 1987. The computer as a collection management tool. *In:* GENOWAYS, H. H., JONES, C. and ROSSOLIMO, O. L., eds, *Mammal collection management,* pp. 97–110. Texas Tech University Press, Lubbock.

SARASAN, L. and NEUNER, A. M., 1983. *Museum collections and computers.* Association of Systematics Collections, Lawrence, Kansas.

WILLIAMS, D. F., 1987. Computer selection criteria. *In:* GENOWAYS, H. H., JONES, C. and ROSSOLIMO, O. L., eds, *Mammal collection management,* pp. 77–95. Texas Tech University Press, Lubbock.

WILLIAMS, D. W., 1987. *A guide to museum computing.* American Association for State and Local History, Nashville, Tennessee.

WILSON, D. E., SABO, B. A. and BLAIR, G., 1987. Automated data processing procedures at the U.S. National Museum of Natural History. *In:* GENOWAYS, H. H., JONES, C. and ROSSOLIMO, O. L., eds, *Mammal collection management,* pp. 111–128. Texas Tech University Press, Lubbock.

WOODWARD, S. M. and EGER, J. L., 1987. Microcomputer system at Royal Ontario Museum. *In:* GENOWAYS, H. H., JONES, C. and ROSSOLIMO, O. L., eds, *Mammal collection management,* pp. 129–133. Texas Tech University Press, Lubbock.

Present address of author:

R. D. Owen
Department of Biology
University of Missouri–Kansas City
Kansas City
Missouri 64110
U.S.A.