

# REQUIREMENTS ANALYSIS

## PART II

### *SYSTEM REQUIREMENTS*

COLLECTIONS INFORMATION SYSTEM  
RE-ENGINEERING PROJECT

MUSEUM OF VERTEBRATE ZOOLOGY  
UNIVERSITY OF CALIFORNIA  
BERKELEY

*Stanley D. Blum  
Barbara R. Stein  
James H. Beach*

**DRAFT**  
10/20/95

## V. System Requirements

### A. Introduction

A system requirements analysis takes the system design another step closer toward technical specification. System requirements describe, in a more technical way, the characteristics, components, and capacities of a system that will satisfy the functional requirements (described above).

System requirements in the past have typically been described in more detail than we intend to do here. Two factors mitigate our need for a more detailed analysis. First, the Museum of Vertebrate Zoology is not a “free standing” institution that must take complete responsibility for its entire information technology infrastructure, but rather is part of a larger university environment, with an existing and evolving infrastructure. This effectively removes certain decisions from this analysis as it is clearly advantageous for the Museum to use the existing infrastructure.

Second, the trends in information technology are clearly toward “openness” and interoperability. From network cards, to scanners, to database management systems, information technology products are being developed as interchangeable parts that permit, or even require, complete systems to be assembled from components manufactured by different vendors. While true “plug and play” interoperability may not be a reality yet, it would be almost difficult today to repeat the mistakes of the past, in which collection management systems were acquired as complete, but “closed” packages, incapable of working with any other products. As market forces turn system components into commodities, it becomes more and more unlikely that any single component choice can lock an organization into a “dead end” that requires all previous investment to be discarded. If some requirement of a systems analysis is overlooked or vastly underestimated, it is usually possible to correct the oversight by “scaling-up” one or a few components, and the cost to do so is less prohibitive now than it was in the past. The one component area where this is not true is the physical network infrastructure, but the MVZ does not have the capacity to change that infrastructure and must work within the constraints of what already exists.

The purpose of this system requirements analysis is to provide a sound basis for a specification of the system architecture, an architecture that will both satisfy immediate requirements and will be scaleable and expandable to accommodate future needs. The analysis begins with a review of the categories of information processing software that will be required to meet the needs of the Museum. This is followed by a qualitative and quantitative review of the data that comprise the Museum’s information management “problem”, a review of needs that derive from users of the system (i.e., the tasks they will perform, their locations, and the times they will need access to the system), and a review of the requirements for system and data security. The systems requirements analysis closes with a review of flexibility requirements, and a review of existing constraints on the system architecture.

### B. General information processing requirements

#### 1. Database management capabilities

The Museum holds nearly 600,000 specimens that are described by a relatively consistent set of variables. Although the values change for each specimen, the categories that appear on

cards, tags, and labels exhibit a high degree of consistency, particularly within a collection. This type of information can be managed best by a database management system (DBMS) rather than a document management or full text retrieval system. A DBMS is the only type of software capable of providing the ability to evaluate and operate on data within their proper context; e.g., to retrieve locality records that fall within a range of elevations, or to distinguish the person who collected a specimen from the person who identified it.

**a) *The appropriate data model***

Most DBMSs are designed to operate on data that conform to a particular structural data model. The most common data models are: hierarchic, network, relational, and object-oriented. Of these, the relational model is the most appropriate for MVZ and its collections information because it offers the greatest flexibility and support for unplanned queries, a common need in scientific data management, and because relational products dominate the market place.

**b) *Complexity of information/data structures***

Information models recently developed for natural history museums show that collections information is far more complex than has been represented in early collections databases. We estimate that between 60 and 100 tables will be required to correctly represent MVZ collections information. The relationships between tables will include a fair number of many-to-many and recursive relationships. This degree of complexity approaches or surpasses the limits of currently available DBMSs in the desktop class, but is well within the limits of DBMSs in the workstation class.

**2. *Application development, query, and reporting tools***

Database management systems of the type described above are now typically designed to function as database servers. While they provide strong data management and security features, they provide essentially no native facilities for making the information they contain easily accessible to unsophisticated users. Instead, they engage in "conversations" with external applications, which then perform the "presentation" function. The MVZ, therefore, will need to acquire an application development tool that facilitates the creation of easy-to-use applications. This kind of tool allows programmers to create desktop applications that include the complete variety of standard interface "widgets", such as forms, tabular views, menus, list-boxes, buttons, pop-up windows, etc., without having to spend a lot of time doing low-level coding. The programmers are then free to focus on facilitating the flow of information between the users and the database. The applications produced by the programmers function essentially as translators, allowing users to interact with the database without having to learn the arcane language of the database.

Custom applications developed by programmers allow users to enter, edit, and delete the contents of the database, but typically provide only the most commonly needed views of the data. A separate *ad hoc* query and reporting tool is usually necessary to provide users the degree of control they desire in extracting and formatting data from the database. Again, the query and reporting tool enables users to tell the database precisely what they need without having to learn the database's language.

### **3. Text and document management**

The functional requirements described above do not reveal a large, near-term, need for document management and large scale “information retrieval” of the kind typically found in law firms, government agencies, libraries, etc. Full scale document management involves storing a large number of documents, enabling them be retrieved according to subject content, and displayed with embedded figures and without a loss of formatting. It is possible that in the long-term, field notes and Museum correspondence may be scanned, and the strategy for managing document images may include “document management” in this traditional sense. It is also possible that the database management system may provide adequate capabilities for managing document images. At present, full-featured document management should be viewed only as a potential long-term need.

In the near-term, the MVZ’s need for text management lies within the realm of “information publishing”; i.e., creating and maintaining multi-media HTML “documents” for use in the MVZ Web-server. These will be original documents, and the greater need is not for managing a large number of them, but for making them accessible via the World Wide Web.

### **4. Geographic information processing capabilities**

Virtually all collection objects have associated with them data that describe their location of origin. Spatial data thus comprise a significant component of collections data. The Museum has a need for geographic information processing capabilities in two areas: the determination of geo-coordinate data from textual locality descriptions, and the generation of maps and other analytical products from collections information and other geographic data.

Fewer than two percent of MVZ specimens are associated with geo-coordinate data. The vast majority of MVZ collections data are therefore effectively unavailable to analysis by a GIS below the level of secondary or tertiary political entities. Translating textual descriptions of location into geo-coordinates has proven to be an extremely difficult task to automate. It is not even possible to predict that an automated solution will be available within the next ten years. The Museum must therefore undertake a significant effort to geo-code collecting localities before MVZ collections data can be made maximally useful. The time required for such a project can be reduced significantly with a tool that enables a user to digitize the latitude-longitude of a locality from a digital map.

The degree of power and sophistication needed for the second function, analysis and display, remains to be determined. It is possible that MVZ may develop collaborative relationships with one or more organizations possessing extensive GIS capabilities. In such a relationship, MVZ would provide collections data and biological expertise, while the collaborating organization would provide complementary analytical capabilities and expertise. The collections information system, therefore, must have the capability to “serve” information to geographic analysis and mapping systems that are either local or remote.

### **5. Image processing capabilities**

Requirements for capturing, manipulating, and displaying images arise in several areas. Vertebrate systematists are not generally optimistic about the ability of mass-produced specimen images to capture taxonomically significant information. The one exception noted thus far concerns the skulls of mammal type specimens. The MVZ follows the traditions in mammalogy and ornithology and does not lend type specimens from these collections.

External researchers must examine MVZ mammal and bird types on site. It is possible that a series of 1-5 black and white photographs of each type skull, made available over the Internet, could supplement published information and enable external researchers to determine whether additional on-site examination is required. The Museum holds 345 primary mammal types. Thus, a project to digitize skull photographs is well within the realm of feasibility.

A second use of imaging technology would be to make the photographic, technical and fine art collections available over the Internet. The historical photographs are particularly important as habitats are rapidly disappearing. Although digitizing these collections is a long-range goal, as with any visual resource collection, adding a digital image to the basic descriptive record improves the effectiveness of the catalog.

The last potentially high-volume need for images concerns the field notebooks and correspondence. Most field notebooks are written by hand, thus it is unlikely that the text could ever be encoded in a machine readable form. It is more likely that notebook pages will simply be scanned as binary images, attached to relevant data objects, such as specimens, localities, expeditions, authors, etc., and made available for viewing. Scanning the entire collection of more than 670 notebook volumes is likely to generate more than 100,000 images. A carefully designed pilot project will necessary to establish a timeline for implementation. Completion of this project is outside the five-year scope of this development effort.

Small-scale imaging needs will arise in the development of WWW-documents. The use of images within this medium generally makes it much more engaging and thus more effective as a means to disseminate information. Each document will probably have a relatively narrow focus, such as a hyper-text essay on the founding of the MVZ, and is not likely to require a large number of images.

In each of the contexts described above, the need for imaging technology in the Museum is limited primarily to the basic capture, storage, and delivery functions. The need for complex image manipulation and modification is presently not assessed as significant, or at least not beyond the capabilities provided by standard desktop (Mac- and PC-based) imaging software, such as Adobe Photoshop. The objective in such image manipulation will be to minimize the file size of each image, without appreciably degrading image quality.

Modest needs for image processing may also arise within the realm of GIS and mapping. Museum staff are unlikely to undertake tasks as complex as satellite image or aerial photography analysis, but they will want to combine collections data with base maps and other geo-referenced data, and to visualize the resulting maps. The GIS and/or mapping software should handle most of such "image processing" needs, but scanners and printers may be called upon to serve both GIS/mapping and standard image input-output needs.

## **6. Audio**

The need for digital audio stems from the ornithological research work being conducted by MVZ staff. Audio signal processing requirements include: analog to digital conversion (using existing cassette and reel-to-reel tapes as the source); mass storage of digital audio files; digital audio playback (via existing hi-fi equipment); and the generation of sonograms

from a variety of spectrographic analyses. All of these capabilities (except mass storage) currently exist within the MVZ in the form of a sound analysis workstation -- an Intel 486 personal computer with an audio digitizing card and signal processing software. Only a small portion of the audio tape collection has been digitized, and it remains to be determined whether or not the research or collections management functions will generate a requirement to save the digitized samples. Digital audio files are quite large (see below), but digitized versions could both preserve the collection without a significant loss in quality, and make the collection more accessible to remote users.

## **7. Video**

The need for full-motion digital video arises from the Hildebrand film collection of animal locomotion. At present, only brief segments of these films have been converted to video tape, but the Museum does intend to convert the entire collection to this format as a means to make the material more accessible. The need for digital video versions of these films is unclear at present. Although digital video versions of the films could serve as a more permanent archive, the attendant reduction in image quality makes the digital video a poor "backup" medium. Digital video versions of the films could effectively make them more accessible to remote users, but the demand for the material is presently unknown.

### **C. Data Types, Volumes, and Usage Patterns**

The information processing functions described above create or use data of different types, which in turn create hardware and software requirements for input and output. In addition, the rates at which data are created, read, updated and deleted (data dynamics or usage patterns) impinge upon a variety of system design issues, such as the appropriate storage media for a particular data type, the design of backup protocols, and the design of physical data structures to optimize system performance.

#### **1. Alphanumeric data**

Virtually all of the collections information that has been automated thus far are of the alphanumeric type; i.e., are composed of text, numbers, and punctuation. These data are the Museum's most important information resource and managing them effectively is the most important charge for the new system. The treatment of alphanumeric data is, therefore, more extensive than for other data types.

**Input:** Keyboards are projected to be the dominant input device for new text data. The next most important data source will be data files created outside the scope of MVZ information management; e.g., "authority files" of taxonomic names from another organization, or new specimen and locality records that have already been computerized by the collector. Finally, it is possible that existing paper-based information can be entered via optical character recognition. Type-writer quality, paper-based inventories do exist for the Lantern-Slide collection, the Grinnell-Miller Library reference books, and the Museum's publication list. All of these collections, however, have been assigned a relatively low priority for automation.

**Output:** Output devices (or modes) needed for text data will consist of "standard" display monitors, printers, and electronic files. We expect that the highest volume of text-based output will be via display monitors, mediated by database clients such as the main collections management applications. In addition, the system must be capable of exporting data to operating system files. These files will then be distributed to end-users via the network, or portable media, such as floppy disks, optical disks, and tape.

The greatest output challenge, however, concerns printing. Under the current system, most specimen catalog cards, tags, and labels are written by hand, in India Ink, on special stock. Only box labels for skeletal material are printed from the system. An appreciable amount of manual work could be avoided if it were possible to print specimen information directly onto cards, tags and labels. The print quality of reasonably priced laser printers (e.g., 600 x 600 dpi) is now sufficient for all needs, but two problems remain: fixation of the ink to labels used in fluid collections (alcohol, formalin, and glycerin), and the ability to print on the special stock papers.

Catalog and accession cards should not pose an insurmountable problem. Most laser printers can accept relatively heavy stock paper and the 4"x 6" card size, at least via manual feed. The availability of an automatic feeder (paper tray) is the only remaining special requirement for printing cards.

Although many if not most specimen tags are written in the field, a significant number are written at the Museum. Problems associated with tags are that: they come in a variety of sizes, some very small; several types are acquired with fastening strings pre-attached; and currently, information is written on both sides of a tag. It is unlikely that any mass-produced (reasonably priced) printer will be capable of printing on pre-cut tags. The two remaining alternatives are to print data onto adhesive label stock, and then affix labels to tags, or to print data onto an uncut tag stock and then cut tags from the stock and tie strings. The label option poses a potential problem because adhesives can degrade and fail. Printing to uncut tag stock eliminates the effort to write information manually on tags, but incurs the effort to cut tags and tie strings.

Dry labels for skull vials, skeleton boxes, etc., apply only to a single specimen, contain full specimen data (e.g., collector, collector's number, date, and locality), and are currently printed with a dot matrix printer. In the new system, dry labels with full specimen data should be printed on a laser printer.

Specimens in the fluid collections are commonly stored more than one to a jar, even though jar-mates may not have been collected at the exact same time and place. All specimens are identifiable by their tags, which contain at least the catalog number, but the jar label typically contains only summary information, such as taxon name, one or more political unit names, and a "range" of catalog numbers. ("Range" does not imply that all intervening catalog numbers are contained in the jar.) Unlike specimen tags, jar labels are not a "repository" for remarks or annotations by researchers. If the specimens are re-identified or re-arranged, a new jar label is produced and the old one is simply discarded. Given that wet label information is not critical and that many museums are getting good longevity results with laser-printed labels, the MVZ may want to consider generating wet labels with laser printers. This would be, however, a relatively low-priority requirement.

At present, we expect that all dry labels, catalog cards, and accession cards, will be printed from the system. Specimen tags, and perhaps wet labels, will continue to be produced manually.

**Volume:** It is impossible to project accurately the storage volume required for text-based data without knowing both the target data structures and the specific database management software. The objective here is to provide only a rough estimate (within a factor of two). The estimate is limited to data that will be managed by the DBMS, i.e., those data contained within the various collection catalogs and the transaction management applications, but excludes text-based data created for HTML documents (for use with the Web server).

The numbers of catalog records expected after five years are shown for the eight largest MVZ collections in Table 1. Projected numbers are based on current collection sizes and annual growth rates. Growth rates for the three main collections are based on the most recent five years in which the collection operations were "normal" (i.e., excludes 93-94, when the collections were being moved). All other growth rates are estimated. Record sizes (Char/Rec) are estimated from current record structures and assume no storage efficiency is gained in converting from a flat-file to a relational design.

Table 1. Projected catalog growth and data volumes.

Catalog	# Records	Growth/Yr	5-Yr Est	Char/Rec	Ann Data Vol.	Proj. Vol (MB)
Herps	222,000	1,407	228,635	1,000	1.41	228.64
Birds	176,000	465	176,000	1,000	0.47	176.00
Mammals	183,000	1,327	190,999	1,000	1.33	191.00
Eggs & Nests	14,000	-	14,000	1,000	0.00	14.00
Tissues*	19,500	1,600	27,499	200	0.32	5.50
Photos & lantern slides	11,000	-	11,000	1,000	0.00	11.00
Curator's Slides	15,000	500	17,500	1,000	0.50	17.50
Vocalizations	8,000	300	9,500	1,000	0.30	9.50
<b>Totals for Catalogs</b>	<b>648,500</b>	<b>4,000</b>	<b>675,133</b>		<b>4.32</b>	<b>653.13</b>

\*

Storage requirements for transaction management data are projected in Table 2. The amount of data required to document a given transaction is estimated as the sum of two components: transaction-level data, e.g., transactor name and address, dates, summary description of material, etc., plus item-level data, i.e., descriptive and tracking information about each item in the transaction. MVZ does not itemize accessions because complete specimen information, including accession number, is captured in catalog records. A separate detailed specimen description under an accession would be redundant. MVZ also does not intend to itemize borrowed material because paper-based records are sufficient to satisfy institutional needs. Yearly transaction activity for accessions and loans is based on actual five-year averages. Estimates of deaccessions and borrow activity (not currently documented) were provided by MVZ staff.

\* The Tissue Collection is estimated to grow at half the rate of the three main collections.

Table 2. Projected transaction rates and data volume.

Transaction Type	Trans-level data		Average #		Total		5 Yr Est (MB)
	Bytes/Trans	Items/Trans	Bytes/Item	Bytes/Trans	Trans/Yr		
Accessions	1,500	59.25	na	1,500	54	0.4050	
Loans	1,000	30.96	1,000	31,961	78	12.4010	
Deaccessions	1,000	5	500	3,500	5	0.0875	
Borrows	1,000	?	na	1,000	15	0.0750	
Total Projected Transaction Data (5 yr)							12.9685

Total projected volume of alphanumeric data, based on collection catalogs, transaction management records, and a DBMS “overhead factor” of 100%, is shown to be approximately 1.3 GB (Table 3). Database overhead includes table indexes, but not the temporary space required by the DBMS for data manipulations and calculations.

Table 3. Projection of disk space required for text data.

Source (5 yr proj.)	Volume (MB)
Catalog subtotal	646.13
Transaction subtotal	12.97
Total Text-Data	659.10
Overhead Factor	x 2
Required Disk Space	1,318.20

**Data Usage Patterns:** The usage patterns associated with text data, i.e., the absolute or relative frequencies at which data are created, read, updated, and deleted, are hard to predict because the new system is intended to make data access and manipulation easier. Improving system capabilities should have no appreciable effect on the data deletion rate, should appreciably increase data creation and update rates, and should dramatically increase the data read rate, particularly once collections data are made accessible via the World Wide Web (WWW) and linked to analysis and visualization software, such as a GIS.

Create – Data will be added to the system primarily in three ways: cataloging new specimens (ca. 4,000 specimens or 4.32 MB per year), bringing new collection catalogs on line, and documenting new collection transactions (ca. 2.6 MB/yr). Additional text data will be created in the form of HTML documents for the World Wide Web server, but estimating the volume or accumulation rate for these documents must wait for specific projects to be defined.

Read – Estimating the rate at which data are read, i.e., simply displayed or printed, is much more difficult. The only components of the data “read” rate that have been documented in the past are the number of requests for collections information and the number of records provided in response to those requests. The last six years of information requests are summarized in Figures 2 and 3.

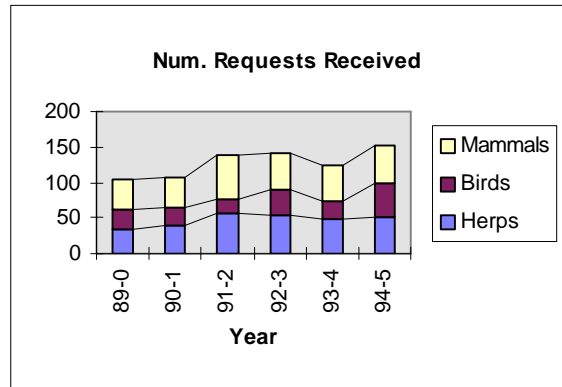


Figure 2

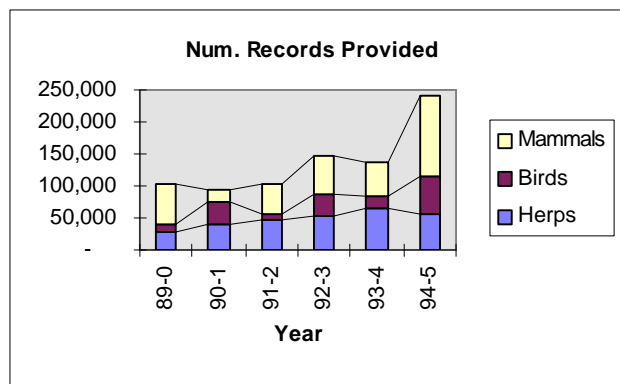


Figure 3

In these past six years, the Museum has received an average of 127 requests per year and has provided an average of 137,000 records per year. Although there is variability among years, both statistics appear to be increasing. Within the most recent year (1994-5), more than a third of all collection records were provided to external users.

Information requests received from external users, however, represent only a small fraction of data read from the system. While the MVZ research staff is much smaller than the external user community, their research is much more focused on MVZ material and they use collections information more frequently. The MVZ Curatorial Associates and Assistants, however, will in all likelihood continue to be the largest users of collections data (as measured by data volume). In the course of their daily work, these staff members regularly issue queries against the database that return hundreds and thousands of records. It is reasonable to expect that most records will be read somewhere between 10 and 100 times more often than they will be updated.

Update – The frequencies of updates are also difficult to estimate. Updates are made for a variety of reasons, such as eliminating data format inconsistencies among records, correcting data entry errors, updating taxonomic and geographic nomenclature, and documenting the preparation and research work done on specimens. The curatorial staff estimate that as many as five percent of specimen records can be updated each year.

Delete – Collection records are essentially never deleted. The only need to delete a catalog or transaction record arises from the entry of “dummy” or otherwise completely erroneous entries. Any record that represents a real object will simply be updated to reflect the current disposition or description, even if the object is consumed in analysis, lost, or donated to another institution.

## 2. Image Data

Digital images do not yet comprise a significant proportion of Museum information, but as noted above, several functional areas will require at least modest capabilities for capturing, storing and displaying images. Potential subjects of digital images include: specimens and preparations, hand-written field notes, correspondence, photographs, slides (photographic transparencies), technical illustrations, fine art (paintings), and maps.

**Input:** None of the image-based projects is defined well enough at present to identify the most appropriate image capture technology. The Museum currently has only two image capture devices, a gray-scale scanner and a video camera with a frame-grabber board. Additional color scanners and digital cameras are available on campus for use in small projects. Experimentation with existing equipment, as well as consultation with the Museum Informatics Project, will be included in the planning phase of any image-intensive project. The Museum should also consider doing initial photography on regular film and having the film digitized onto a Kodak photo CD when the film is developed. This is frequently the most cost-effective way to acquire both “archival-” and “on-line-” quality images simultaneously. Images used in GIS and mapping will most probably be acquired already in digital form, via the network or on CD-ROM, either from commercial sources or academic collaborators.

**Output:** Virtually all desktop computers are now purchased with monitors that are at least moderately capable of supporting graphics work. The important choices that must now be made concern the size of the monitor, and the speed and memory capacity of the graphics card. Twenty-inch monitors or larger, capable of displaying at least a thousand pixels in each direction (e.g., 1280x1024) in 24-bit color, are recommended for any graphics intensive work (e.g., GIS/mapping).

Day-to-day needs for hard-copy graphics output can be satisfied by full-featured desktop laser printers (600 x 600 dpi, with at least 2 MB of RAM). Special circumstances that call for higher resolution, color, or larger format can be accommodated by other units on campus (e.g., a color laser printer in the Museum Informatics Project), or taken to a commercial printer off campus.

**Volume:** The amount of storage space required by an image depends on image size (resolution), color depth, image content, and file format. An estimate of image data volume is provided for each project.

Mammal skulls will be photographed in black and white, and each image intended for on-line use should be cropped and processed (i.e., all background pixels dropped to absolute black) to minimize storage requirements while maximizing image quality on a typical computer monitor (i.e., 800 x 600 pixels). An 800 x 600, 8-bit grayscale (256 shades of gray) image of a mammal skull, in JPEG format is estimated to require between 30 and 100 KB. If each

skull is represented by five images, a maximum of 500 KB will be required for each skull, and approximately 170 MB for the entire mammal type collection.

The vast majority of the images in the photograph collection are full-tone black and white. If each image requires between 100 and 200 KB, all 8,000 photographs would require between 0.8 and 1.6 GB of storage space.

A crude survey of field notebooks indicates that the average notebook contains approximately 300 pages. An existing, 600 x 900, binary, GIF image of a notebook page requires a little less than 20 KB. By these parameters, the entire digitized field notebook collection (670 volumes) would require approximately 4.02 GB of storage space.

Within the GIS and mapping realm, hard disk space will be needed not just for data files obtained over the network, but also for data extracted from CD-ROM. It is usually the case that a particular project needs data for only a particular region, while a CD-ROM contains information about one or more layers over a much broader area. The relevant data are thus copied from CDs to hard disks to eliminate the need for CD swapping and to speed performance. A desktop computer/workstation that will be used extensively for GIS/mapping should have both a CD-ROM drive and at least a gigabyte of hard disk space available for digital maps, images, and working files (in addition to space dedicated to software).

**Data Usage Patterns:** Images representing collection items, whether images of mammal skulls or old photographs, are expected to have a single usage pattern. Digital images are most likely to be created in a single sustained effort to establish an on-line image resource. It is unlikely that any single image, once edited to balance file size and image quality, will ever need to be updated. At most, the image will be replaced entirely when new imaging or network technology is developed to support higher quality images. Images will be used most commonly by remote users, and should be stored so as to maximize retrieval and delivery performance.

Image data supporting the GIS and mapping functions will be used primarily by MVZ staff in data analysis and visualization. It appears now that these data should create, in essence, a library or resource base that can be enhanced and used by multiple individuals. It remains to be determined whether demand will be high enough to require support for multiple simultaneous users.

### 3. Audio

The need for processing sound (audio) data derives from the Natural Sounds Laboratory collection, which is composed primarily of bird vocalizations, recorded in the field, on magnetic tape (analog). Research on bird vocalizations involves comparative analysis of digitally-produced spectrographs (sonograms), which in modern laboratories entails digitizing the analog recordings. An additional motivation to digitize analog recordings concerns their long-term preservation. Magnetic tapes are notoriously unsuitable as archival media, with life-spans that range from 5 to 50 years, depending on the particular formulation of the tape. Converting analog tapes to digital form is finding greater acceptance as a strategy for the long-term conservation of audio information. Digital information (data files) can be refreshed or copied without further degradation, while copying analog media always entails some loss in fidelity.

**Input:** At present, all sounds have been recorded on magnetic tape in analog form, either on one-quarter inch reel-to-reel tapes or on cassettes. The MVZ sound lab has equipment for converting analog signals to digital form, including: tape recorders, speakers, an amplifier, a Frequency Devices Series 900 tunable low-pass filter (used as an anti-aliasing filter), and a Gateway PC (Intel 486 DX2-66) equipped with a Data Translation A/D D/A converter board (DT2821-G, capable of sampling analog signals at up to 250 KHz), the SIGNAL/RTS software, and a Bernoulli drive (150 MB capacity). The SIGNAL processing card has the capability to sample analog signals at 30, 44.1 (the digital audio tape standard), 48 (the audio CD standard), and 88 KHz, and to create digital representations (files) of these signals.

**Output:** The sound analysis workstation has the capability to play back audio signals stored in digital form and to create spectrographs from both analog and digital sources.

**Volume:** The sound tape collection contains 455 master tapes, which hold up to 550 hours of sound recordings, as estimated from tape length and recording speed. The volume of data generated from an analog source depends on the sampling rate. The vast majority of recordings are of bird vocalizations, which can be represented adequately in digital form when sampled at 48 KHz. An hour of digital sound sampled at that rate, however, produces almost 350 MB of digital data. Digitizing the entire collection thus could require up to 190 GB of digital storage space.

**Data Usage Patterns:** The rate of digital audio data creation has two components that should be considered separately. Conservation of the collection will be undertaken as a single project that will entail digitizing all analog material onto a medium with better longevity characteristics than magnetic media (i.e., optical or magneto-optical media). The sound-tape conservation project is not strictly part of the collection information system re-engineering project, and will require separate planning and funding. It is being considered here only to provide the briefest introduction to what will be entailed in the project.

The second component of digital audio data creation is that which results from ongoing field work and analysis of vocalizations still in analog form. Current MVZ staff and affiliated researchers add approximately 20 hours of new sound recordings to the collection each year. The rate at which vocalizations are digitized and analyzed is sporadic, but well within the capabilities of existing equipment.

Update and delete rates for digital audio data are essentially nil, as any analog recordings digitized within the framework of data analysis are likely to be kept in digital form.

One final aspect of working with digital audio data should be recognized. Digitizing the entire collection would result in a volume of data that would be very expensive to keep "on-line" (i.e., on a hard disk, or even a multi-disk CD-ROM player). The entire collection will probably span some 50 to 100 units of removable media. Whether the demand for digital audio data will be high enough to warrant the cost of a "near-line" solution, such as a CD jukebox, remains to be determined. Local researchers should expect that data will have to be loaded manually, i.e., copied from removable media to a hard disk, prior to analysis. The number and size of files a researcher may require in a single analysis cannot be known precisely, but it is not likely to exceed one or two hours worth of recordings, i.e., 300 to 600

MB of data. Adding a gigabyte of hard disk space to the sound analysis workstation will minimize the need for extraneous file management and is likely to be well worth the cost.

#### **4. Video**

The MVZ collections include a series of 16 mm films, which the Museum intends to convert to video tape in order to enhance their accessibility. It is unlikely however, that significant amounts of these films will be converted to digital video within the 5-year scope of this project.

#### **D. Users**

Several aspects of system requirements are determined by the activities, locations, and other attributes of the user community. The numbers of users and the kinds of work they do (i.e., input, query, or update) determine several aspects of system capacity and also influence how the physical database should be designed to optimize performance. The topology of the network needed to support the flow of information between users and the system is determined by the physical locations of users. The needs for system security, particularly the need for different levels of authority, are determined by the responsibilities or functions of different users or user groups. Finally, the times that users actually do their work determines when the system must be up and available to support that work. In this section, the user groups are defined and described according to the tasks performed by each, the levels of authority that must be supported by the system are outlined, the working hours and locations of users are described, and the typical and maximum numbers of simultaneous users are estimated.

##### **1. User Groups**

###### **a) Curators**

The primary responsibilities of Curators are to pursue the research and education missions of the Museum. Their primary collections information needs are for reports and data analyses. In the past, all information requests, including requests from Curators and other MVZ staff, have been processed by the Curatorial Associates. The Curators, therefore, have had little direct interaction with the collections databases. The technical skill level among Curators varies from basically computer literate (word-processing and e-mail) to near-power-user. With an easy-to-use interface, all Curators should be capable of retrieving data from the system independently of the Curatorial Associates. A modest variety of pre-defined query screens and reports would satisfy most of their needs, but some might also make significant use of an easy-to-use, *ad-hoc* query tool that provides greater flexibility.

The other significant activity performed by Curators in their research mission is the creation and recording of collections information in field notes, laboratory notes, and on specimen tags. Under the current system, Curators pass this information to the collections management staff, who then take responsibility for entering all data into the system according to protocols that enforce data consistency. In the future, it may become common place for field workers to capture a significant amount of their research data on laptops or PDAs in the field. If the Museum's Curators adopt this practice, it will be possible to extract relevant data from these research data sets and import them into the collections databases. Capturing data electronically at the time they are created, in some sense, moves the boundaries of the system into the field, makes field workers more directly responsible for the initial quality of collections

data, and removes at least one transcription/interpretation step from the data capture protocol. Coordinating data formats and content standards between a researcher's "electronic notebook" and the collections databases, however, will require effort and will impose constraints that reduce the researcher's "autonomy" in designing their own "notebooks". On the other hand, the Museum's researchers already appreciate the value of consistency in field notes, as evidenced by their use of the "Grinnellian" format. Moreover, consistency between electronic field notes and the collections should allow researchers to take relevant collections data with them into the field, to combine them with data about material just collected, and thus to monitor the progress of field and laboratory work more effectively as it happens.

As part of the teaching faculty in the Department of Integrative Biology, Curators are also responsible for preparing course materials. They may, therefore, either supervise or take direct responsibility for creating materials to be posted on the MVZ WWW-server.

**b) *Curatorial Associates***

The Curatorial Associates are currently, and will continue to be, the most demanding and most knowledgeable users of the system. They spend a significant portion of their time actually using the system. They may be called upon to perform all information-related tasks, e.g., data entry and proofing, data quality management, and basic system administration (creating new accounts, adjusting authorities, etc.). The only system-related tasks they are not expected to perform are system development and maintenance, which are likely to require extensive technical skills.

Curatorial Associates have the ultimate responsibility for establishing and maintaining the quality of collections information. The Curatorial Associates devote a significant portion of their time to training other users, particularly Curatorial Assistants (below), in both physical and information-related aspects of collections management.

**c) *Curatorial Assistants***

Curatorial Assistants perform the bulk of basic cataloging and collection transactions work (cataloging, accessions, loans, shipping). They are responsible for entering, proofing, printing, and updating both catalog and transaction data. The Museum demonstrates its on-going commitment to collections and collections-based research by training students, both undergraduate and graduate, in the effective practice of collections management. Because the tenure of Curatorial Assistants is limited to one or two semesters, the long-term average level of expertise in both computers and collections management is lower than it is in comparable institutions. The system should assist Curatorial Assistants in following policy and work-flow protocols, and to the greatest extent possible, should prevent them from entering invalid data.

**d) *Graduate Students, Research and Teaching Assistants, and Laboratory Technicians***

Graduate students perform a wide variety of activities related to collections information. They create and record collections information either through their own research or by being employed as research assistants on projects conducted by Curators. Like Curators and Curatorial Associates, graduate students use collections

information in research analyses. Graduate students may be employed as teaching assistants and called upon to develop or maintain course materials. In addition, graduate students are sometimes employed as Curatorial Assistants, and are thus required to perform the physical and information-related collections management activities described above.

In addition to the general uses of the specimen collections described above, Curators, Curatorial Associates, graduate students and laboratory technicians all use the tissue collection. In the course of their research activities, they create and modify information about tissue samples, which will need to be recorded in the tissue collection database. The MVZ tissue collection is currently being reorganized, and new protocols for managing the collection and its related information have yet to be established. The largest information-related challenge here will be to maintain concurrence between the tissue database and the actual contents of the tissue collection. The new protocols will require these individuals either to update the database themselves, or to fill out forms or logs, which will then be checked for accuracy and entered by a Curatorial Associate or Assistant.

***e) Other researchers and the general public (virtual visitors)***

This user group is a “catch-all” category and includes scientists from other academic institutions, conservation organizations, and governmental agencies, as well as historians, sociologists, philosophers of science, undergraduates, and the lay public. The system should provide these users with complete basic descriptions of the collections and collection items, and the ability to extract data sets made up of collection records. These capabilities should promote more effective use of the collections and should enable these users to formulate their requests for specimen loans, off-line information, and collection visits more precisely. This, in turn, should enable the Museum staff to deal with each request more efficiently, and ultimately, to serve a greater number of users.

***f) Technical support personnel (system and database administrators, software developers, etc.)***

As noted above (section V.B.1, Database management capabilities), a relatively sophisticated database server will be required to manage MVZ collections data. This type of DBMS platform, in turn, must be set up and managed by technically skilled personnel. In addition to the database server, which will constitute the core of the MVZ collections information system, the requirements described thus far call for:

- custom-developed applications that enable users to work with the collections databases;
- the establishment and maintenance of an MVZ web-server; and
- the development of at least minimal GIS/mapping capabilities within the Museum.

Each of these three areas constitutes a relatively significant and distinct set of technical skills. The functions performed by technical personnel are described below under typical job titles. The list does not imply that a separate employee must exist for each role. Several of the skill-sets overlap, and the University has technical support units with personnel that may be able accept some of these responsibilities.

Server Operating System Manager (OS) – manages system security; manages system physical resources (disk allocation, disk partitioning) and system parameters (number of users, memory configuration and allocation); and performs operating system updates.

Database Administrator (DBA) – manages database physical resources (disk allocation, disk partitioning); allocates and monitors database space; creates database logins; configures database software; optimizes database performance; and performs database backups and restoration.

Database Owner (DBO) – modifies table structures; creates and manages indexes; monitors and manages performance; creates and manages integrity constraints; and manages user authorities (grants permissions to individuals and/or groups).

Application Developer (Programmer/Analyst) – works with other technical personnel, Curatorial Associates, and other users to design databases and create applications; writes, documents, and manages both client and server code.

Web-Server Developer/Administrator (Webmaster) – establishes and maintains web-server platform; works with other technical personnel and content experts to create and maintain information resources on the web-server; develops and maintains modules that allow the web-server to act as a client to the collections databases.

Personal Computer Manager – installs and maintains personal computers (Mac and DOS/Windows) and peripherals, such as printers and scanners; installs and upgrades both hardware and software components; works with end users to backup and restore PC-based software and data; capable of configuring and providing basic instruction for using a wide variety of desktop applications.

## **2. Levels of authority**

### ***a) The needs for different levels of user authority***

Concerns that create a need for different user authorities (capabilities) are:

- to protect the system from malicious or unintended damage;
- to protect the integrity, accuracy, and consistency of data;
- to restrict access to data that are unverified or have yet to reach the Museum's standards for quality (e.g., have not been proofed);
- to restrict access to precise locality data that are deemed sensitive – i.e., where unrestricted dissemination of such information might constitute a threat to endangered species, populations, or habitats; or would represent unauthorized disclosure of information about private property.

**b) Authority Classes and Capabilities**

Each authority class below is named for the predominant user group it contains. These designations are for convenience only and not intended to imply that a user with a particular job-title or status within the Museum could not be granted a different level of authority, either permanently or temporarily. All users of non-public applications will be required to logon to the system with a user-id and password.

**(1) Curators**

If the current pattern of work flow is maintained in the new system, Curators will have the authority to read all collections data, but to modify (create, update, and delete) little or none. One of the primary reasons for restricting update capability in the past was to insure that updates of specimen information within the system were coordinated with updates of paper records and all other aspects of physical collections management. Hence, all data modification functions were delegated to the collections management staff. If data modification permissions are given to individuals outside the collections management staff, a new set of protocols will be needed to insure consistency.

If Curators or their graduate assistants take responsibility for creating HTML/Web-based materials in support of teaching activities, they will also need permissions to create, read, update, and delete relevant files on the Web-server.

**(2) Curatorial Associates**

In day-to-day activities (normal operational mode), Curatorial Associates must have the authority to create, read, update, and delete the content of any portion of the database, including all authority files. As a matter of good practice, however, a Curatorial Associate should execute a special login to acquire the authorities of a database-owner and thereby gain the capability to change the database structure, such as drop table, delete database, etc. Such changes must typically be made in concert with modifications in both server and application code, and must be made with extreme caution.

**(3) Curatorial Assistants**

Curatorial Assistants must have the authority to create, read, and update most data about collection items. As a security measure, they should not be able to delete entire records. The system should allow them to tag entries for deletion, and then allow Curatorial Associates to perform the actual deletion after reviewing the tagged records. Curatorial Assistants must have the capability to add data to some authority files (e.g., the person table), but should not be capable of adding or changing records in most (e.g., taxonomic and geographic nomenclature). Without the capability to modify authority file content, Curatorial Assistants will have to ask the Curatorial Associates to make necessary changes.

**(4) Graduate students and affiliated researchers**

Graduate students should have the capability to read all public data. A graduate student should never have the capability to process collection transactions, unless he/she is currently working as a Curatorial Assistant. If a graduate student generates data that are destined, in part, to become MVZ collections data, he/she may be given the capability to enter, edit, and use those data according to the protocols established for that project (see comments under Curators, above).

(5) Public

Users in this category should have the capability to read all public data (i.e., all data that are not deemed “sensitive”), but will not have the capability to modify any data.

(6) Technical support personnel

Technical support personnel must have the authority to create and modify all software and data appropriate to their functions (see above), but only those functions.

### 3. User Locations

The locations from which users need to conduct business determine the topology of the communications infrastructure, the partitioning of both data and software among computers, and the number of workstations and their capabilities.

The system must allow MVZ staff and visitors to perform information-related tasks from their desktop computers, as well as from computers located in special work areas, such as the sound analysis and molecular laboratories. Museum staff should be capable of logging onto the system and performing basic tasks from any workstation within the Museum, not just their normal workstation. Analytical tasks, such as sound analysis or certain GIS/mapping functions that require special hardware and software, may be available, however, only at special shared workstations. In addition to locations within the Museum proper, a number of affiliated faculty and staff have offices and laboratories in other campus locations. Some MVZ users would like to perform the complete range of information management functions from home via modem. This is not an absolute requirement, but would allow users to maintain flexibility in their work schedules and would allow some “crises” to be solved without actually being at work.

The system must allow external users to query the collections databases from remote locations. Two trends in wide-area information sharing are relevant to this requirement. First, the proliferation of the Internet and generic client-server software, particularly the World Wide Web, has clearly demonstrated the utility of this “vehicle” in delivering information to a global audience. The system must, therefore, support a “query and reporting”-type interaction with external users via the hyper-text transfer protocol (HTTP; i.e., a WWW server). Second, recent innovations by bio-informatics specialists have demonstrated that the World Wide Web standard can be used to provide end-users with single-point access to multiple distributed collections databases. The MVZ will need to participate as an information provider in such a network, and must be capable of “publishing” collections data electronically according to the semantic and syntactic standards that will be

emerging as the foundation of this incipient network within the systematics collections community.

#### 4. System Availability

The requirements for system availability are determined by users' work schedules, the degree to which their work depends on system availability. While it is trivial to state a requirement that the system must be available 24 hours a day, 7 days a week, 52 weeks a year, actually building-in the redundant network and server components that guarantee this level of availability could raise the system cost by an order of magnitude. The cost of system down-time must be weighed against the cost of building in guaranteed availability.

Normal and extended user work hours are estimated in Table 4. Normal working hours are intended to encompass at least 85% of the time that users in a particular group will need access to the system. Extended hours increase the boundaries to encompass 99% of access times. All times are expressed as local time, either Pacific Standard or Pacific Daylight-Savings time.

Table 4. User Hours.

User Group	Normal Hrs (85%)	Extended Hrs (99%)
Curators	8:00 - 18:00 M-F	8:00 - 18:00 Sun-Sat
Curatorial Associates	7:00 - 19:00 M-F	8:00 - 18:00 Sun-Sat
Curatorial Assistants	8:00 - 17:00 M-F	8:00 - 17:00 M-F
Graduate Students	7:00 - 22:00 M-F	7:00 - 22:00 Sun-Sat
Technical Support	8:00-18:00 M-F	24hr Sun-Sat
External Users	5:00 - 24:00 M-F	24hr M-F

The work hours of internal users' are based on casual observations. The work hours of external users are projected by:

- 1) Assuming that, relative to their local time zones, external users make 90% of their connections in normal work hours (between 8:00 and 18:00), 10% in the evening hours (between 18:00 and 24:00), and 0% in the early morning hours (between 0:00 hrs and 08:00).
- 2) Assuming that 85%, 10%, and 5% of connections will be from North America, Europe, and Austral-Asia, respectively. (These figures approximate the geographic distribution of connections received by the UCB Museum of Paleontology server.)
- 3) Adjusting external user-times to Pacific time.
- 4) Assuming weekend activity to be negligible.

Scheduled down-time. The system server should not be taken off-line intentionally during normal working hours, 07:00 - 19:00 M-F. Regularly occurring downtime, such as may be required for back-ups, should be scheduled between 00:00 and 06:00, if possible. This would not affect local or even North American users significantly. It would block the first four work-hours for Australian and East Asian users, and the last two work- and first four evening-hours for European users. Taking the system off-line every day at this time would block an estimated 3% of attempted user connections. Other foreseeable outages, such as modifications to the database server hardware or software, should be scheduled to occur on weekends.

Unscheduled down-time. The impact of an unscheduled system outage depends on what tasks or system uses happen to coincide with the outage. Under "normal" conditions, collections management staff can fill their time with tasks that do not require the system, and both internal and external information users can tolerate 2-3 days of system unavailability. The worst time for the system to fail would be during a project in which extra staff have been hired specifically to do information-related tasks, or when a user is trying to develop an information product under a deadline, such as a figure for a presentation or statistics for a grant proposal. While it might be difficult to determine the exact "cost" of system down-time, it would be difficult to justify an estimate larger than \$1,000 per day. The time to recover from system failure, on the other hand, is determined by availability of replacement parts and the availability of technical personnel to actually swap parts and bring the system up again. The availability of parts and technical personnel, their costs, and options for payment and service contracts all need to be determined before the appropriate annual "insurance expenditure" can be specified. As an initial target for system availability, the Museum should strive to achieve a recovery time of less than three days. In other words, system managers should be able to restore complete system functionality within 2-3 days of an unforeseen disaster, such as hardware failure. This may require a service contract with the hardware vendor.

## 5. Typical and Maximum Number of Simultaneous Users

The numbers of potential users are shown by user group in Table 5.

Table 5. Numbers of Users.

User Group	Number of Users
Curators	5
Curatorial Associates	2
Curatorial Assistants	2-4
Graduate Students and Other Staff	30-40
External Users	?

The heaviest users of the system will be the Curatorial Associates and Assistants. Their jobs do not require them to use the system constantly, but it will not be uncommon for as many as four to six to be logged on simultaneously. The Curatorial Associates typically will perform large numbers of queries, reports, and multi-record updates. Curatorial Assistants, on the other hand, typically will perform data entry and single record updates.

All other users, both internal and external, will perform predominantly query and reporting operations. This is likely to be the case even if data entry protocols are revised such that

researchers take responsibility for some aspects of data entry. It is hard to predict, however, whether the internal or external research communities will create the larger demand on the system. Internal researchers are expected to create a much larger demand per individual, because their research tends to be intimately related to the collections. The external research community, on the other hand, is likely to be many times larger. The historical number of data requests from external researchers (see data usage patterns under text data, above) indicates that at least one request will be made every other day. The combined demand of both internal and external researchers will have to be one to two orders of magnitude higher before system performance is likely to be effected.

If the system takes more than 5-10 minutes to complete large complex queries, it may become important for the system to allow a single user to perform several tasks simultaneously. One strategy for implementing this capability is to allow a single user to execute multiple logins, and thereby obtain the capability of executing several tasks at the same time, either from a single or multiple workstations.

### **E. System and Data Security**

The human resources required to build a collections database are so extensive that responsible stewardship requires protecting it from system failure, human error, theft, and natural catastrophes such as fire, flood, and earthquakes. Redundancy is the only safeguard against data loss and should be built into the MVZ system through a carefully designed backup protocol.

The purpose of creating backups is not just to safeguard years of work, but also the most recent weeks or days of work. All computer usage should be guided by the adage: "If you don't want to have to do again, you should back it up." Backup systems, in turn, should be designed according to the adage: "If it isn't easy, it won't happen."

Mitigating the risk of data loss through backups applies not just to databases, but to all data that reside on any computer, including the word-processing files, data sets, e-mail, software, and configuration files that reside on personal computers. Even the installation and configuration of software on a personal computer represents an appreciable amount of labor. Although the investment residing on a personal computer is typically far less than on a server, personal computers have proven to be at greater risk to software viruses, human error, and theft. Most individuals that manage their own computers do not invest in a backup system, such as a tape drive and software package. Instead, they intend to backup their data files to floppies. The reality is, however, that diskette-based backups require more file management and diligence than the user anticipates, particularly when a computer with a large hard disk (100 MB or more) has been in use for more than 6 months. Lapses occur and the individual tacitly begins to accept the risk of data loss. Institutional settings, however, provide an opportunity to reduce the per-individual cost of performing backups, as measured by both hardware/software costs and labor costs.

The new MVZ system must provide the capability to backup all servers containing MVZ data, including the database server, Web server, and GIS server/workstation. The Museum should also give serious consideration to acquiring either a sub-system or a service for backing up personal computers and shared workstations.

Properly designed backup protocols strike a balance between a large number of factors, as outlined below.

- **Purpose.**
  - **Recovery from disaster.** All backup protocols serve the fundamental purpose of enabling information to be recovered from a backup copy in the event of a disaster.
  - **Version recovery.** Re-writable backup media, such as tape cartridges, are often recycled when the information they contain is superseded or outdated by a more recent backup. Extracting copies from a backup cycle and retaining them in an archive for a longer period is a practice commonly used to create a partial audit trail or version history of changing information. The backup archive preserves the state of the information at specified intervals and allows any “archived” version to be recovered. The audit trail is only partial, however, because backup events are not triggered by information changes, but only by the passage of time. If “audit trailing” is developed into the MVZ database itself, a backup archive would be superfluous as audit trails triggered by actual data changes provide the ultimate in version recovery. It should also be noted that creating a backup archive creates a whole host of additional concerns, such as longevity of the backup media, long-term data format and media compatibility, and protecting the backup archive itself (see below).
- **Data modification rates.** Not all data change at the same rates. Many software program files are never modified and need to be backed up only once. Other programs can require significant effort to configure and configuration files should be backed up after each modification. Data files typically change much more frequently and should be backed up in accordance with the effort required to re-create them.
- **System availability.**
  - **Down-time for backups.** Many backup systems require information to be taken off-line (made unavailable to users) in order to back it up. Some databases, operating systems, and backup applications, however, do allow backups to be created without taking the system off-line. Incremental backups, which copy only changes made since the last full backup, can decrease both the time and storage space needed for backups.
  - **Down-time before recovery.** Information critical to business operations is typically backed up in ways that minimize the time required to get the system up and running again after a disaster. With tape-based backup systems, protocols that minimize the time required for backups usually increase the time required for restoration.
- **Delayed discovery of data loss.** While catastrophic damage to a computer system is usually more than obvious, it is not uncommon for more subtle errors to go undetected in computer files, particularly large databases, for more than a year. Mitigating the risk of undetected errors is the strongest motivation for creating a backup archive.
- **Verification of backups.** Backup systems can fail in subtle ways that make them appear to be working correctly while data are actually not being duplicated correctly. Automatic

verification at backup time lengthens the backup procedure, but guarantees at least the initial integrity of the backup copy. Backup tapes should be checked regularly for readability in another tape drive. Tape-head alignment problems can result in backups that pass verification, but are, nevertheless, unreadable by another tape drive.

- **Protecting backups.** If backups are not protected in a vault or stored off-site, catastrophes such as floods, fires, earthquakes, and theft, are likely to destroy backups as well as on-line data. Backups should also be protected against more subtle destructive forces, such as humidity and mold. Finally, it is important to recognize that the replacement and upgrading of hardware and software is a natural part of systems operations. Software and hardware changes should not be made without a full understanding of how these changes may effect the utility of existing backups, particularly a backup archive. Upgrading database software can destroy backups almost as effectively as a fire.

**Database Server.** The creation and modification rates of MVZ collections data are variable and unpredictable, but commonly large enough that the MVZ database should be backed up daily. The MVZ database system is not critical for the Museum's daily operations, at least not the way that a transaction processing system would be critical to a commercial bank. Minimizing the time required to restore the database is therefore less important than minimizing scheduled down-time and the number of backup tapes.

The backup cycle and archive protocol should allow daily versions to be recovered for at least a month, monthly versions to be recovered for at least a year, and quarterly (or perhaps semiannual) versions to be recovered in perpetuity. MVZ should consider exporting quarterly backups to a software-independent, documented, data structure (e.g., a delimited text file), duplicating the backup, and sending a copy of each quarterly extract to an off-site repository. Server software should be backed up any time it is upgraded or reconfigured.

Operation of the TAXIR platform (CMSA) by the campus Department of Information Systems & Technology includes only daily backups. Thus, the Museum's Curatorial Associates have created and maintained the backup archive of the TAXIR database by submitting batch jobs to the computer center that copy TAXIR data files to magnetic tape. It is noteworthy both that the backup archive has, in fact, been used to restore the database several times, and that even this system is not easy enough to be considered sufficient. The new MVZ system should continue to generate a backup archive, at least until audit-trailing is built into either the database server software or applications. Creation and maintenance of the backup archive should be integrated into the normal backup cycle. The MVZ must also transcribe the existing archive onto new backup media as part of the system migration.

**Desktop Computers.** The campus has recently begun to offer a backup service on a recharge basis of \$25 per workstation per month. The MVZ currently contains 22 personal computers (Mac and DOS/Windows) and one Sun Sparc II workstation. Using the campus backup service for all machines could incur annual charges up to \$6,900. Limiting backup services to computers that MVZ staff use on a daily basis could reduce the charges by half, but would leave the less frequently used workstations unprotected.

A relatively sophisticated backup solution, capable of working with MS-Windows, Macintosh, and UNIX clients, would entail the largest initial investment (ca. \$3K-6K for backup software, a 4-8 GB DAT drive, and tapes), but would pay for itself in a few years. Such a system would require a small amount of daily attention from a system administrator, but would be invisible for most users.

Another alternative that should be considered is the acquisition of three separate, less sophisticated, and less expensive backup systems – one for each of the platform-types present in the Museum (MS-Windows, Macintosh, and Sun). This less sophisticated alternative would require an initial outlay of approximately \$3K, but it would entail more effort to operate, as three separate tape drives would have to be managed.

### **F. Maintenance and System Flexibility**

Conventional wisdom in software engineering holds that application maintenance – including the fixing, adjusting, and enhancing of software to keep pace with changing user needs – can constitute up to 80% of the total cost of developing and operating custom-built applications over their entire life-cycle. The point here is not to project the exact cost of maintaining the Museum's system, but simply to acknowledge the fact that user expectations evolve, and consequently, hardware and software must evolve to keep pace.

The Museum's experience with collections management software (TAXIR and supplemental PC-based programs) has been atypical. Virtually any other system in which the core software had not been modified or upgraded for 15 years would not be functioning. Two factors have contributed to the longevity of TAXIR. First, the IBM/CMS platform has been relatively stable and upgrades of the operating system have included a high degree of backward compatibility. Second, the TAXIR software itself is fairly simple and flexible. The logical structure of a TAXIR "databank" is a simple flat file, and records can be manipulated with a reasonably straightforward command-line interface or with "scripts" of commands. More importantly, the Curatorial Associates have been able to add and modify data fields as needed for each new collection catalog.

The requirements described thus far call for a system that provides a much larger suite of capabilities than the MVZ's TAXIR-based system now provides. It must be understood that the core database software in the new system will be just one component, and that the overall increase in data management capability will be achieved at the cost of greater complexity. Assembling the system will require a significant amount of custom code. For example, the most important system modules, the cataloging and transaction management applications, will require:

- 1) database server code – the stored procedures and triggers that guarantee data integrity and boost database performance,
- 2) client applications that enable users to enter and edit their data without having to confront the full complexity of the actual data structures, and
- 3) standard queries and reports that allow users to display and print data.

Modifying the system after it has been built may require, in some cases, much more effort than might be apparent to an end-user. Something as simple as adding a new field requires not just a minor change to the database, but might also entail, for example, modifying 4 stored procedures, 3 screen forms, 4 error messages, 3 pre-defined queries, and 2 reports. Before making the actual code changes, the maintenance programmer must first determine which program elements need to be changed, and afterwards must test the new code and deploy the changes across the entire system.

Some aspects of system maintenance, therefore, can require nearly the same skills as the initial system development.

The new MVZ system should be designed and developed to maximize flexibility and to minimize the cost of system maintenance. The factors that can promote these characteristics are discussed below. Some apply only to the selection of system components, while others identify programming and design strategies that have been shown to reduce the cost of system maintenance.

### **1. Scalability**

The system should be scalable, such that demands for greater capacities can be met by replacing or adding components rather than the entire system. The system should readily accommodate, for example, the addition of hard disks or other storage media to achieve greater data capacities, or additional memory or CPUs to improve performance. From the outset, the system should be configured to accommodate optimistically projected needs, and not just current needs.

A simple example concerns the configuration of RAM in personal computers, which commonly contain four SIMM sockets for RAM. RAM SIMMs come in a various capacities, including 4, 8, and 16 MBs. If the current requirement is for 16 MB of RAM, it should be configured as either 2 x 8 MB, or 1 x 16 MB, so that at least two sockets are available for additional memory. A configuration of 4 x 4 MB would occupy all four SIMM sockets, and would leave no room for expansion.

### **2. Extensibility**

Extensibility refers to a system's ability to incorporate new functions. For purposes of discussion these extensions are divided roughly into minor, moderate, and major, according to their impact on end-users.

#### ***a) Minor***

Small extensions are those that only marginally effect the users' perception of the system. These include such things as modifying the content rules for a data field, or the development of a new report. The system should be designed to maximize both the ease with which small maintenance changes can be made and the likelihood that such changes can be handled by MVZ staff. Wherever possible, configuration files and authority files should be used to control application behavior and database content. Generic tools with point-and-click interfaces should give users the ability to customize queries and reports of existing data. Users should have the ability to save their customized query and report templates, and to share their templates with other users.

#### ***b) Moderate***

Moderately significant extensions are those that change the way a user performs an existing business function, such as automating a formerly manual sub-function and modifying the system to hold new information. Moderate extensions may not appear to the user as something completely new, but may in fact represent the development or re-development of custom applications. They are likely to require a skilled programmer and may well be the most expensive kind of adjustments to make. Minimizing the cost of this kind of maintenance can be achieved primarily by

anticipating future needs, and creating modular, well-documented applications. This requirements analysis and the MVZ Information Model have both been framed to provide a solid estimation of the Museum's information management challenges over the next five years and beyond.

**c) Major**

Major extensions to the system refer to the automation of new business functions, and may entail the addition of new kinds of software. Examples include: configuring a Web-server to be a client of the database; enabling the system to support the capture, processing (modification), and delivery of images; or configuring a GIS to work with the collections databases. Although major extensions will look completely different to users, some of them may, in fact, be less expensive to implement than moderate extensions because they can be achieved by configuring different classes of existing software to interoperate, rather than by developing additional large customized programs.

The descriptions above provide only a broad indication of the kinds of software modifications that will be necessary to keep the system's capabilities on track with user requirements. It is impossible to project the level of technical support that will be required for system maintenance without first seeing the working system and the rate at which user-expectations evolve. It can be stated unequivocally, however, that the skill sets needed to maintain the MVZ system will include relational database design and tuning, SQL programming, object-oriented software development, and a basic knowledge of UNIX, Mac, and Windows system administration. The experiences of other organizations in managing comparable information resources should lead the MVZ to anticipate that a programmer with these skills will be needed on a more than half-time basis.

### **3. The Use of Industry Standard Components**

The MVZ system should be assembled to the greatest extent possible from commercial-off-the-shelf (COTS) hardware and software. System components should be purchased from companies or manufacturers that have: 1) achieved a significant market share, 2) are likely to be long lived, and 3) have a history of providing a smooth upgrade path between product versions. Using COTS products provides two important benefits. First, vendors must upgrade products regularly to maintain or increase their market share. Upgrades offer new features and keep pace with the evolving information technology environment. The marketplace thus creates an environment in which consumers can leverage their resources in obtaining continued "maintenance work" from commercial IT vendors.

Second, the prevailing strategy in commercial software development has shifted from applications that work in isolation, to products that work well with other applications. A product with a significant market share becomes, in essence, a feature of the IT environment, and other products are more likely to interoperate it. Using mainstream COTS products provides a continuing opportunity to incorporate major new functions in the system.

### **4. Technical Documentation**

The Museum's information management needs are sufficiently complex that custom-built software applications will be the most effective way to accomplish the Museum's information management objectives. The MVZ should establish a technical documentation file for all

system components, or at least all components managed directly by the MVZ. This is particularly important for application code, because documentation substantially reduces the total long-term costs for system maintenance. Software design guidelines and naming conventions should be established, described, and followed in all application code. While extensive comments should be embedded in the code to provide the lowest level of descriptive detail, one or more separate documents should describe all software components, their interdependence, and strategies for their re-use.

## **G. System Architecture Constraints**

The Museum and the UC Berkeley campus have already invested heavily in a variety of information technologies. The architecture of the new MVZ collections information system should take the best advantage of previous technology investments and should call for new investments only to extent that greater efficiency or functionality is required.

### **1. UCB network and the Internet**

Most buildings on the UC Berkeley campus, including the Valley Life Sciences Building in which the MVZ is located, have been wired with category 5 unshielded twisted pair (UTP) cable to provide Ethernet data communication services. TCP/IP is the dominant networking protocol, but Apple-Talk and IPX (Novell) protocols are also supported. The MVZ sub-net is connected to other campus sub-nets and the Internet via the campus backbone. Dial-in access to the campus network is provided, to some degree, by a series of IS&T-operated modem pools, which support TCP/IP communications through SLIP and PPP. When classes are in session, however, the demand commonly exceeds the number of lines, and it is difficult to get a free line.

### **2. The Instructional and Collections Computing Facility (ICCF) database server**

The Instructional and Collections Computing Facility operates several UNIX-based computers that provide a variety of computing services to campus users. Most relevant to MVZ's needs is the Sun 4/690 platform and the Sybase System 10 RDBMS it supports. ICCF holds an unlimited-user site license for the Sybase software and makes it available to campus users, including collection facilities such as the MVZ, without a usage fee. The only cost-sharing required by the end-user organization is the purchase of sufficient disk space for the organization's database. The database server is also equipped with an Exabyte 8-mm DAT tape drive (5 GB) for backing up the databases. Annual support contracts are maintained with vendors for both hardware and software.

In addition to the database server, ICCF provides a full-time UNIX system administrator and a full-time database administrator. The system administrator's responsibilities include system security, configuration management, and operating system upgrades. The database administrator manages all aspects of Sybase configuration and maintenance, establishes and executes the database backup protocols, and is available to database users for consultation on database design and performance tuning.

### **3. Desktop environments and applications**

The MVZ now has 22 personal computers, of which 13 are Macintosh and 9 are Intel-based (DOS-Windows). The Macintosh computers are used by Curators, graduate students, an illustrator, and the administrative support staff. The Intel-based computers are used by

Curatorial Associates, Curatorial Assistants, and graduate students. The two specialized data acquisition and analysis workstations (sound analysis and morphosys computers) are also Intel-based.

Important applications currently used by MVZ staff and affiliates on desktop computers include:

- Word Processing: WordPerfect (Mac, DOS); MS-Word (Mac, Windows)
- Spreadsheet: MS-Excel (Mac, Windows)
- Database: Access (Windows), FileMakerPro (Mac), dBase (DOS), Paradox (DOS, Windows)
- E-mail: Eudora Pro (Mac, Windows)
- WWW browser: Netscape, Mosaic (Mac, Windows)
- Misc communications: ftp, telnet, TN3270
- Technical Illustration and Graphics: Aldus Freehand (Mac), Powerpoint (Windows), Adobe Illustrator (Mac)
- Statistical Analysis and Graphing: Cricketgraph, Deltagraph (Mac); Systat, Statview, SAS, SPSS, NTSYS (various)
- Molecular Biol. / Phylogenetics / Pop. Genetics: PAUP (Mac), Phylip (Mac, DOS), Mega (DOS), Biosys (DOS), DNASys (Mac), ESEE (DOS)
- GIS: ARC/View (Mac)

Macintosh and Intel-based desktop computers clearly have become the most important computing platforms for MVZ users. The new MVZ collections information system should continue this trend by enabling collections information to be integrated into these environments as completely as possible. In particular, direct interaction with the MVZ database should be through an interface that is consistent with the user's normal desktop environment, and collections information should be easily extractable so that familiar desktop applications can be used to perform analyses and data visualizations not provided by the data management components of the system. The relatively even split between Macintosh and Intel-based computers indicates that client applications for the MVZ database should be portable across the two platforms.

#### **4. UNIX workstation**

The Museum also has a Sun SPARC-2 workstation with 800 MB of hard disk space, a 250 MB tape drive, a CD-ROM drive, and an attached SPARCprinter. It has been used recently as a development platform for multi-media HTML documents, but it currently does not support a publicly accessible Web server. It receives daily-use only as a print server.

### **H. Summary**

The preceding descriptions and analyses of MVZ information management requirements should provide the MVZ system developers with a strong rationale for selecting the specific software and hardware components that will comprise the new MVZ collections information system. The analyses should also give the MVZ curatorial staff a better appreciation for the complexity of the coming system, and the levels of skill and effort that will be required to develop and maintain it.

In brief, the new MVZ system should be constructed around a sophisticated, multi-user, relational database server, which will provide a cost effective repository for most of the MVZ's text-based collections information. The database should be tuned to favor query and reporting speed rather than input speed, because the data are read from the system much more frequently than they are created or updated. The database server must be capable of working with a wide variety of client software, including a suite of applications that enable technically unsophisticated users to perform their work functions easily, a flexible query and reporting tool, a World Wide Web server that delivers collections information to external users, and GIS/mapping software that enables collections data to be visualized in combination with other spatially-referenced environmental data.

In the short-term, the success of an information system can be measured by the system's ability to meet the organization's requirements. In the long-term, however, success must be measured by the system's ability to meet the organization's changing requirements. We have endeavored, therefore, to make this requirements analysis as prospective as possible without engaging in fantasy. We also note, however, that even within the planning and analysis phase of this project, research protocols and user-requirements have been changing and expanding perceptibly. Some of this is due to changes in the way science is conducted, some of it is due to the users' increasing familiarity with new technology. At the moment, the system is just an abstraction, and users are describing their future needs according to their current understanding of what they do and what they think is possible. Once the system is in place, users will adjust to it and may well discover that some decisions made early in design were, in fact, too short-sighted. This is the nature of software engineering and MVZ management should fully anticipate the eventuality of software maintenance.