Version : DRAFT 5/97

Inventory & Monitoring Program

Data Management Protocols

1997

U.S. Department of the Interior
National Park Service
Natural Resources Information Division

Preface

The elements of data "management" are not new inventions. Several existing documents stipulate the ideals of data management for National Park Service units and Inventory and Monitoring (I&M) programs. These draft Data Management Protocols do not attempt to comprehensively list, review, or synthesize those criteria. Rather than a listing of mandatory policies and regulations, these guidelines are aimed toward aiding park personel with their immediate and future data management concerns. Much of the content herein comes from experiences gained at Shenandoah National Park (SHEN) which is an ecological monitoring prototype park of the I&M Program. Shenandoah had a natural resource data collection effort spanning many years before any formal "guidelines" for park-based programs were ever written. These resource studies were often initiated for very specific purposes and without the foresight that they would become "long-term." In the past, little time and no staff were available for professional data management, and data integrity suffered. In light of this and other examples, comparing past performance to "new" standards is clearly unfair and hardly recognizes the effort and commitment that preserved our resource data over the years. The immediate challenge in the National Park Service is to inventory, organize, and manage our valuable data resources (past, present, and future) in an efficient and cost-effective manner. Effective data management will preserve the quality and integrity of our data archives and provide the basis for informed resource management decisions for years to come.

Acknowledgments

This document is a result of the ongoing effort by the National Park Service I&M Program to organize and coordinate natural resource inventory and monitoring endeavors throughout America's parklands. The project has benefitted from the input and review of many individuals through the years, especially Steve Tessler when he was Data Manager at Shenandoah National Park. The focus of the text came from needs identified by data managers at prototype parks in the Long Term Ecological Monitoring Program. Thanks to everyone who contributed, whether or not your name is listed below.

Mark Adams, Leslie Armstrong, Tom Blount, Bob Carr, Linda Dye, Dave Demarest, Frank Deviney, Tim Goddard, Dave Grabor, Victoria Grant, Roger Hoffman, Ron Holmes, Michael Kunze, Sam Lammie, Kieth Langdon, Joe Meiman, Jill Minter, Jon Paynter, Linda Petit-Waldner, Dave Ryn, Shane Spitzer, Steve Tessler, Lisa Thomas, Dean Tucker, Gary Williams, Randy Winstead

Joe Gregson, Ed.

Table of Contents

Chapter 1 - Introduction

Background and Purpose

The Natural Resources Inventory and Monitoring Guideline (NPS-75) was published in 1992 and is a fundamental reference for natural resource management programs. Along with describing the objectives and rationale of I&M programs, NPS-75 "provides data administration and reporting guidelines for the program(s)." Data management is presented as a major topic in NPS-75 and is recognized as a critical component in the success of I&M at national parks. The importance of actively managing natural resource data at parks cannot be overstated, since this data will provide information affecting park management decisions for the foreseeable future.

The primary mission of the I&M program is to provide NPS unit managers with scientifically-based and statistically-valid data upon which resource management decisions can be based. Short-term goals are to increase the credibility, efficiency, and usefulness of resource research programs. In the context of this document, current goals are to assist parks in developing active data management plans (DMPs), identify and catalog existing data collections, and provide general guidelines for data handling. Longer-term goals of the I&M program are to complete biotic and abiotic inventories of the parks and include perpetual monitoring of key "trend and status indicators" to help protect and preserve each park's resources and ecological integrity for future generations.

One attractive feature of the parks' data resources--that they are "long-term"--can also be a weakness; caution regarding data integrity is always in order. The presumption that long-term data is of uniform quality needs to be evaluated on a regular basis. Sources of variation in the quality and accuracy of data include both human and systematic factors. Each resource study can minimize data errors with detailed instructions and specific standard operating procedures (SOP's) for how to collect and record the data (i.e., data collection guidelines). Unfortunately, there are no existing "off the shelf" SOP's for the validation and long-term maintenance of the data once collected. Another concern is the consistency of documentation of the data (i.e., metadata)--the what, where, when, why, and how the data generating programs were implemented. Finally, if there are few or no applications for the data, even the "usefulness" of a dataset becomes debatable.

The admission that data problems exist does not negate the usefulness and informational value of the database, but simply reflects the need to develop a strategy to insure data integrity and availability over time. To carry out the mission of providing data to our primary users--park managers--it is also essential to integrate all the various natural resource studies into a single, coherent, multidimensional view of the park; this was not anticipated when these projects began and is a laborious proposition now. Concerns and shortcomings of existing database endeavors can be best addressed by active data management.

What is data management?

Information is everywhere. Both good and bad information, biased and objective, old and new, detailed and summarized, can be found. But to *create* new information, data must be used. Therefore, information and data are not synonymous. Information is synthesized or derived from data. Consequently, the source and quality of the underlying data is important when information is used to make long-term decisions. A lot of information *may be* based on good data, but much is not. Good data, however, is *always* based on the "truth" and reality. Because of the close dependence of data on the truth, collecting and managing data is a serious business.

Data collection is the foundation of a total data management strategy. What and how to measure are determined by scientific and statistical criteria. A data management plan implemented during the collection phase helps produce unambiguous records for simplified yet more complete data entry, documentation, integration, and retrieval. Careful database design and planning promotes efficiency and minimizes errors. This may add a minor cost to a data collection effort, but advance planning for long-term data management pays dividends later on.

With the advent of the computer age, data management is often confused with the *tools* that support it--computer hardware and software. Manipulating data on a computer does not make it "managed" any more than putting books on a shelf "makes" a library. At a basic level, organization and maintenance of the items cause them to be managed; at a higher level, management includes constant vigilance against corruption, loss, and misuse. While computers allow larger data sets to be manipulated and speed the quest for useful information, computerized data is uniquely susceptible to *accumulating errors* through careless handling, systematic problems, and poor security. When the data can no longer be guaranteed to represent the "truth" they have lost their value.

A good data management strategy, therefore, includes procedures to ensure the integrity, security, and availability of up-to-date datasets. One approach is to apply a "50-year" test. When encountered 50 years from now, the data should be *certifiably* accurate, with accompanying documentation (metadata) that describes: the extent and purpose of the data; the history of where, when, and why the data was collected, how and by whom, what was done to it along the way, and references to how the data was used. The "50-year" touchstone can help guide data management in the present and suggest goals for future management efforts.

To discover useful information in data and explore it in meaningful ways requires the anticipation and building of relationships between the pieces of data and levels of information. Planning for the exploitation of known and potential relationships among data is a vital part of data management and is the primary challenge of designing the structure of individual files. Database design and integration provide feedback to and from the data entry process to enhance the quality and efficiency of data collection.

In the end, a good database design, combined with a simple access interface, can allow a casual user to easily explore data relationships, visualize their inherent characteristics, alter scale and focus, and answer questions about the data as they arise. This is an appropriate vision for well-managed data. The lists below summarize what data management is, and what it is not.

Data Management is:

- a serious and cost effective endeavor
- active rather than passive and best guided by a preconceived plan
- basic organization and maintenance at the record and file level
- a guarantee of the integrity and security of data
- good database design that promotes exploration and utilization of data
- a service function to provide reliable, understandable data to users
- often done on a computer, which must be well maintained

Data management is *not*:

- simply entering data into a computer or a file
- simply storing data in a notebook, computer, or file cabinet
- working with or analyzing data using computers and software
- using data without regard to its source, quality, and original purpose
- keeping the data to yourself, or in a form only you can understand

Perils of data mismanagement

It is unnecessary to dwell for long on the negative results of ignoring or underestimating the importance of good data management. Data can be lost through accident or disaster, corrupted through mishandling or neglect, rendered legally indefensible due to inadequate documentation and quality assurance, or found to be useless beyond a narrow purpose due to poor database design. Data can also be incorrect. Remember Murphy's Law (of Data): *without attention, whatever can go wrong with the data will go wrong,* and the problems will not be discovered until a large group of people are depending upon the data to make emergency decisions. The ultimate cost of poorly managed data can be astronomical, but most major problems can be avoided with good data management practices, procedures, and policies.

Document Overview

This document presents guidelines to help NPS units develop comprehensive data management programs. The organization of this document follows three main themes: 1) preparation of an active data management plan (DMP), 2) inventory and cataloging of existing data sets as part of the DMP process, and 3) general data handling procedures to encourage more uniform data protocols. The DMP chapter follows the format of a draft data management plan to aid in the preparation of one at the park level. The Dataset Catalog chapter illustrates a front-end database tool to inventory the park's datasets in a format that is both adaptable and compatible with a servicewide I&M Dataset Catalog. The Data Handling Procedures chapter reviews a workable approach to uniform data handling protocols. The Appendices contain examples and documentation to clarify or expand ideas presented in the other sections.

Chapter 2 - Data Management Plan

Introduction

The Data Management Plan (DMP) is a document that comprehensively organizes and stipulates data management policies and goals. The DMP is not just a report to be written; it is a critical step in the process of designing and implementing an efficient data management program. By the time the first DMP is completed, many aspects of the data management program will already be in place or actively under development. Indeed, the DMP might more correctly be viewed as a progress report for a park's active, evolving data management program.

Scope

The idealized scope of a comprehensive DMP is to include all data management endeavors at the NPS unit. However, from a natural resource, I&M, or individual park's viewpoint, the inclusion of all data management activities may be impractical or impossible. As with many specific data management activities, the overall scope of the DMP will need to be evaluated on a case-by-case, site-specific basis. Small units may be able to define an all-encompassing plan, while larger parks with complex issues and management structure may require a more restricted approach. The common thread is that each site will stipulate the scope of data management program within the DMP itself.

Goal

The goal of this chapter is to provide a common DMP framework for all NPS units to share. Following this introduction is a suggested DMP Outline (provided in WordPerfect format) and general instructions for completing the DMP. The idea behind this approach is that as datasets, information, policies, and programs are assembled, reviewed, and planned, the outline can be filled-in and expanded as needed. Once the necessary aspects of each data management element are identified and addressed, writing the DMP should go quickly. Again, the DMP is an organized reflection of both active and planned data management endeavors.

Data Management Plan Outline

NPS Unit Name

Draft Outline: April 1996

Executive Summary
Acknowledgments (optional)
Contents

Introduction
- Background & Purpose
- Scope
- Data management needs/goals

Section 1: Current Status and Data Resources
- Computer Resources
  - Hardware resources
  - Software resources
  - Minimum computer standards
  - Computer protection and maintenance
- Data Resources
  - Relationship of "non-spatial" data to GIS resources
  - Data resources (list)
  - Estimate of current data load
  - Minimum data standards
    - Existing database design and file formats
    - Assessment of existing metadata
- Human Resources

Section 2: Data Management Strategies
- Introduction
  - Goals of park data management program
  - Relationships with existing data standards and programs
- Administrative structure
- Data acquisition considerations
  - Permitting, agreements, contracts, etc.
  - Data collector relationships (as applicable)
- Quality Assurance and Quality Control (QA/QC)
  - Protocols and standards
  - Verification, validation, and editing
  - Data documentation & metadata standards
  - Data summaries and analyses

- Data maintenance (working datasets)
  - Update procedures
  - Version control
  - Added or derived data
  - Backup routine
  - Documentation and cataloging
- Legacy Data
  - Legacy data list
  - Legacy data update plan(s)and protocols
- Data security
  - Risk analysis & system security
  - Duplication of critical files
  - Equipment contingency planning
  - Access control
- Data archives and storage
  - Documentation and tracking (Dataset Catalog)
  - Metadata standards
  - Physical location of duplicate datasets
  - Scheduled rotation of storage media
  - Data purging policy
- Data applications (standard reports, query tools, decision support)
- Data dissemination
  - Freedom of Information Act (FOIA)
  - Availability and version of active and/or legacy data sets
  - "Request for Data" forms, policy, disclaimer, and fees
  - Data formats, conversions, and distribution mechanisms
  - Tracking/auditing of disseminated data and resulting products

Section 3: Accomplishments, Current Activities, and Long-term Goals
- Accomplishments
- Current Activities
- Long-term Needs and Goals
- DMP Evolution and Update Schedule

Section 4: Implementation of Data Management Plan
- Feasibility Evaluation/Budget
  - Interfacing the DMP Budget with the RMP
- Implementation schedule

Appendices (site-specific data management documentation)
- Park Dataset Catalog List

Glossary (optional for site and/or DMP specific terms)

References

Data Management Plan Instructions

This chapter is designed to aid data managers in the development and writing of an active data management plan (DMP). The text follows the Data Management Plan Outline presented in the previous section. Although the exact content of each DMP will vary from park to park, the outline and instructions provide a common basis for all DMPs. The DMP instructions are not rigorous standards, but emcompass an organized set of suggestions that may be adapted as needed for individual parks or site-specific considerations.

Executive Summary

The Executive Summary of the park Data Management Plan should be written last and presented first. A draft of the DMP should be reviewed by divisions/programs that will be using and/or supporting the plan. The plan should be complete and any support commitments confirmed before the summary is written.

The Executive Summary should probably be framed as a statement reflecting the park's desire to enhance management efficiency and effectiveness. Generally, the summary might contain statements reflecting the state of the park's data management program, resources, collaborative agreements, implementation timeframe, costs (equipment, software, FTEs), and benefits that will evolve from data management.

Introduction

The Introduction should contain a statement concerning the park's assessment of need for active data management planning and briefly cover the anticipated benefits the program will entail. This section should address the following items.

Background & Purpose. This section may be summarized from information in this document and site-specific considerations. This section might include: association with the I&M Program; concerns about long term data management, legacy data concerns, etc.

Scope. How does active data management planning relate to GIS and other park activities? Data management and GIS plans should be part of a wider effort to coordinate and consolidate all data management activities throughout the park with cooperators, clusters, field areas, and other agencies.

Data Management Needs/Goals. What data management issues need to be addressed? Review concerns for critical legacy data, database integrity/security issues, etc. that pertain to the park. What are the general goals for park data management? Some goals are general as noted herein, but others are certainly site-specific.

DMP Section 1: Current Status and Data Resources

A necessary step in data management planning is to prepare a comprehensive inventory of available resources. Resources for data management chiefly consist of computer resources, data resources, and human resources. Other resources, such as office and storage space, may also be important. This section reviews the current status of data management at the unit in the context of available resources.

Computer Resources

Computer resources consist of both hardware and software. These assets need to be documented to ascertain existing capabilities, help organize diverse uses, and guide future procurement plans. In addition, an understanding of "what's out there" may produce immediate dividends.

Computer resource standards are important aspects of data management and can promote more uniform working environments and sharing of limited resources. Maintenance chores are simplified when systems adhere to some minimum standard configuration. System maintenance is also part of a computer standard, because day-to-day work tasks depend on functioning systems. A good place to spell out these standards is in the data management plan.

Hardware Resources

A list or table of significant hardware resources should be included in this section (e.g., Table 1). Important notes for hardware include: CPU speed, memory size, storage media and size(s), available I/O devices, etc. A brief summary of major uses and users could be enlightening. Now is a good time to think about, justify, plan, and spell out future hardware needs. For example, plan to upgrade data management computers to 16Mb RAM if possible, since the demands of GUI's (graphical user interfaces) and the growing size and complexity of datasets require considerable memory.

Software Resources

Another important aspect of a data management plan is a comprehensive inventory of available software resources. An example inventory of the primary software supporting the Shenandoah's I&M program is listed in Table 2 along with Local Area Network (LAN) and support programs. Software resource decisions are affected by past (Word Perfect 5.1, Lotus 123, and dBASE III+) and present (Microsoft Office) software standards. These and other considerations are discussed below.

Table 1. Shenandoah I&M Computer Hardware (as of November 1994)

Hardware Type	Description
Desktop	1 IBM AT 80286, 8 MHz, 512Kb RAM, 40Mb HD, VGA
and Tower Computers	1 Dell 80386, 20 MHz, 2 Mb RAM, 100 Mb HD, VGA 1 Dell 80386/87, 20 MHz, 4 Mb RAM, 300 Mb HD, VGA
	2 Dell 80486SX, 20 MHz, 4 Mb RAM, 120 Mb HD, VGA 1 Swan 80486SX, 25 MHz, 4 Mb RAM, 170 Mb HD, SVGA
	1 Dell 80486DX2, 50 MHz, 8 Mb RAM, 270 Mb HD, SVGA
	2 Dell 80486DX2, 50 MHz, 16 Mb RAM, 210 Mb HD, SVGA
	1 Swan 80486DX2, 66 MHz, 16 Mb RAM, 340 Mb HD, SVGA
LAN File Server	1 Swan 80486DX2 EISA, 66 MHz, 16 Mb RAM, 2 mirrored 1.0 Gb SCSI drives, VGA
Printers	1 HP 4Si Laserjet, with 6 Mb, Ethernet, PostScript, and duplexing
	4 Okidata ML-321 wide carriage, 9-pin dot matrix
	1 Okidata ML-320 narrow carriage, 9-pin dot matrix
Mice	Assorted Logitech 3-button, Logitech 2-button, Microsoft 2-button
Tape Backup	Colorado Memory Systems Trakker 250, portable (parallel port), 250 Mb capacity per tape
	Colorado Memory Systems PowerTape 4000, external SCSI, 4 Gb tape capacity (network backup)
Modems	3 Intel 14.4 faxmodems, 2 Hayes-compatible 2400 modems
Field Data Recorders	2 Polycorder 1600 hand-held computers, 512 Kb 1 Polycorder 286XL hand-held computer, 2 Mb

In reference to past data format standards, legacy databases are typically stored as dBASE III+ format files. However, there are limitations to the dBASE file format, and some facilities are using Access or FoxPro database software. At present, the dBASE format is good for many database implementations, especially data exchange. Most spreadsheet, database, statistical, and graphics software packages can use or import this format without modification.

Most parks currently have a mix of both DOS and Windows software. Although some parks have successfully converted to standardized software suites (e.g. Microsoft Office at Channel Islands), it is not necessary to immediately convert to a pure Windows shop. Some machines are incapable of running Windows 3.1 (much less Windows 95!) efficiently. Adding more memory is only a partial solution since

Tabel 2. Shenandoah I&M Computer Software (as of November 1994)

Application or Function	Software Package and version
DOS Environment Shell	Norton Commander 4.0
Word processing	Word Perfect for DOS 6.0b
DBMS	FoxPro for DOS 2.6; dBASE IV 2.0 for DOS
Simple DBF viewers/editor	DBBrowse, a freeware viewer/editor; DBView (Norton Commander)
ASCII file editors	QEdit 2.15 and TSE 2.0 (both SemWare); MS Edit; NCEdit
Data entry in the field	Quick Collect 2.2 for Polycorders
Reporting tool	R&R Report Writer for DOS 4.0 and 5.0
Code generator	SoftCode 3.0, with templates for BASIC and FoxPro
Programming language	Visual BASIC for DOS 1.0, Professional Edition
Graphics and presentations	Harvard Graphics for DOS 3.0; Stanford Chart for Windows 2.1
Desktop GIS tool	Atlas GIS for DOS 2.1, and for Windows 1.0
Spreadsheet	Quattro Pro for Windows 5.0
Statistics	SPSS for Windows 6.1, with Professional Stats
Support Utilities	Norton Utilities 8.0
Tape Backup	Colorado Backup for DOS 4.05, for Windows 2.0

Windows performance is dependent upon several factors. User familiarity and efficiency with DOS versions of some programs preclude the urgent need to upgrade the software (and related training). Conversion to Windows and Microsoft office should be strategically planned for optimum efficiency.

Minimum Computer Standards

One basic data management endeavor that can be done is to standardize computer systems. The level of standardization is dictated by budgets, distribution of resources, personnel, etc. Minimum computer standards should be considered and addressed in the DMP. For example, Microsoft Office should optimumly have: MS DOS 5.0 or above, Windows 3.1 or above, a 486 processor or above, 8 Mb of RAM, a CD ROM drive, a 256 color SVGA monitor or above, and 85 Mb of free space on a hard disk

drive. However, each NPS unit will still have to critically evaluate their own need for formal standards. Other considerations are discussed below.

The NPS, like most organizations, has come into the computer age. Since this has occurred piecemeal rather than strategically, the organization, consistency, and compatibility of computers, software, and users often leaves much to be desired. Many facilities do not have a full time computer support person, and most have experienced the conflict and confusion of Macs vs. PCs, DOS vs. Windows, Word Perfect vs. Word, etc., and the myriad difficulties that mixed hardware and software entail. This disorganization implies an inherent danger to data and productivity. Indeed, from a computer support and maintenance point of view, tremendous savings can be realized in time, training, installation, trouble-shooting, and improved data-recovery options when all computers have a minimum configuration. The dependence upon computer resources for both basic and advanced support functions further illustrates the need to recognize and manage them as a strategic investment.

An important reason for a standard computer setup is to streamline computer support. This does *not* mean that each computer must be the same model with the same software, components, and peripherals, only that they all adhere to some *minimum configuration*. Non-standard systems require extra time to troubleshoot, and unique problems are more difficult to diagnose and correct. With standard configuration and setup, maintenance and distribution of upgrades can be more automated. Often, simple problems can be diagnosed and fixed over the phone, and solutions to common problems can be shared. One full-system backup has the potential to restore several machines and make new system installations easier.

Networking of similarly organized computers is also more efficient with standardized computers. Common directory structures, program installations, batch files and utilities make a single menu system easier to set up and maintain. Access to common data files is also enhanced, and standards facilitate the *use* of data. Likewise, careful planning and control over common environment variables allows more efficient batch operations and network resource allocation.

Computer Protection and Maintenance

Data management and analysis requires working computers and peripheral equipment. Although this may seem trivial, many times the protection and condition of hardware resources are ignored. Failure to adequately protect and maintain computer equipment can result in data loss, missed deadlines, needless and expensive equipment replacement, or worse. Preparation for the safety and uninterrupted operation of critical computer systems is an important component of data management. Considerations include: power and line protection, scheduled testing and maintenance, minimum physical maintenance and audit, hardware diagnostics, diagnostic software, hard drive maintenance, and spare equipment. Discussion of these considerations is included in the appendix on Computer Protection and Maintenance.

Data Resources

Knowledge of existing datasets and activities is important for efficient data management. How will data managers know what data to collect (and how to manage it) if they do not know what they already have? Hence, a comprehensive and up-to-date inventory of existing and current data resources is needed. A dataset inventory (next chapter) or an active plan to complete one should be included in the DMP.

Relationship of "Non-spatial" data to GIS resources

The scope of data resources will vary from site to site, but a critical aspect of data management is to define data relationships. Some users think of GIS as separate and distinct from "non-spatial" natural resource data. However, almost any type of data has a spatial component, and this distinction is probably not warranted. On the other hand, from a management or application viewpoint, the distinction may be quite clear. Data management relationships with existing or planned GIS programs at each site should be considered and documented in the DMP. Where possible, integraton of GIS into the data management process can avoid duplication of effort, enhance personnel cooperation, and provide more efficient support for resource planning. In addition, much of the information needed for the DMP may already exist in a GIS plan.

Data Resources List

Existing datasets may reside in many forms. From small, disorganized analog files to large and complex, geospatially referenced digital databases, any extant data about park resources may prove vital for resource management decisions. Ideally, the dataset inventory list will be available to resource managers at all levels. The organization and sharing of the dataset inventory should be through the mechanism of the I&M Dataset Catalog. As explained in the next chapter, the Dataset Catalog provides a predefined database structure and software implementation to assist in the organizing and sharing of park-based dataset inventories. A subset table or report from the Dataset Catalog database can be referenced or inserted as an appendix to document the data resources list.

Estimate of Current Data Load

A summary of existing data collection/management efforts will aid in future planning. Review commitments of time, personnel, dedicated hardware, software, etc. in a concise format or table. The conceptual flow of the data might be illustrated on a flow diagram. Sources of the current data load may include:

Current incoming data stream (including specimens)
Verification and Validation of current and legacy data
Analytical data products and reports
Other data products (ranger reports, maintenance schedules, etc.)

Minimum Data Standards

As previously discussed, much efficiency and productivity may be gained by adhering to a minimum standard. The past database standard for the NPS was dBASE III+, which is a functional basic database design, especially for sharing data. For purchasing and upgrade considerations, the current database software standard is Microsoft Access (which supports the dBASE III+ format).

Existing Database Design and File Formats. Brief descriptions in this section include database design(s) and file formats used at an individual site. Specific items that may be reviewed include:

- Standard file formats and structures
- Database design and integration criteria
  - Critical fields: time and location
  - Normalization and linkages
- Data dictionary (strict definitions of data in a dataset)

Assessment of Existing Metadata (documentation and data about a dataset). Included in the Appendix is a Minitemplate for Geospatial Dataset Documentation. The Minitemplate is designed to implement metadata documentation standards published by the Federal Geographic Data Committee (FGDC) for legacy (prior to Jan. 1995) geospatial (i.e. GIS) data. Newer geospatial data must comply with the more extensive general FGDC metadata standard. Whereas this level of metadata is not yet mandatory for "non-spatial" data, similar standards are pending. Hence, accumulating information about extant and future datasets is critical.

Several metadata categories from the Minitemplate discussed above are incorporated in the I&M Dataset Catalog record and field structure. This provides a "first approximation" for accumulating metadata. A summary or list of existing metadata, both analog and digital, should be included in this section.

Human Resources

Human resources are perhaps the most valuable and often the most difficult to obtain components of data management. In the context of the current status of park data management, a listing of current users and personnel that are active and/or available for data management activities, their position, status, and contact information will be useful for planning a more formal data management structure in the next section.

DMP Section 2: Data Management Strategies

Introduction

This section essentially embodies the Data Management Plan. Included are several generic principles and considerations that apply to any long-term data collection and management effort. If a program is just starting, this section will provide a foundation for planning data collection and management--even before any field work is done. Experiences with updating legacy datasets show that most of the problems and expense can be avoided with some advanced planning. On the other hand, legacy datasets do exist and must be addressed. Therefore, the primary goals of the DMP are to delineate a process for managing legacy and current datasets while planning and implementing for future efficiencies.

Goals of Park Data Management

The goals of park data management include providing accurate, efficient, and effective information and support for resource management and protection. Each unit needs to know: what data are available, in development, or stored (both on- and off-site); the quality, timeliness, and uses of the data; how to incorporate this data into resource management decisions; and how the data will be managed over time. As more baseline inventory and legacy data become available, how will the park assimilate these materials and put them to productive use? Will all data be archived on site or do alternatives exist or need to be planned? These and other considerations discussed below should be incorporated into short- and long-term data management goals.

The mandate to carry out inventory and monitoring at parks provides the basic structure to define data management goals. The data-related mission of I&M is to provide scientifically and statistically sound data for decision-making, and the principal goal of data management is to ensure quality data for this task. The data must:

- be truthful
- be precise and meet acceptable accuracy standards
- be accessible to users in a useable form
- be meaningful in the context of the decision being made
- contain clearly defined relationships to other relevant data
- be protected from unauthorized alteration, corruption, or loss
- be maintained with integrity and remain certifiably intact..., indefinitely

Truthfulness, precision, and accuracy are the realm of quality control. Dissemination and access can be problematic but must be addressed. Ensuring that the data are meaningful, both for decision-making and in relation to other data, comes with database design, data maintenance, documentation, access tools, and user input.

Archiving and providing long- and short-term data security round out the primary data management activities.

In addition to the goals discussed above, each park will have unique data management needs to be addressed. For example: how do existing and planned "non-spatial" data relate to existing GIS data, capabilities, and planning? Should GIS be integrated into the DMP? What DMP projects will be incorporated into the Resource Management Plan? What are feasible short- and long-term goals? Indeed, some data management goals may be preconceived, but several goals will evolve along with the DMP.

Relationships with existing Data Standards and Programs

As discussed in the Current Status section, existing data standards and programs should be reviewed for the DMP. Standards that are applicable and/or in use at the park should be listed. Precise standards used in specific databases can be included in an appendix. Although many standards are still under debate, those directly related to each site's data management endeavor are important and should be documented.

The issue of standards will continue to be debated until standards actually exist that all agencies are able *(forced?)* to observe. However, universal standards are difficult to achieve and implement--especially in a rapidly developing technical environment such as computer hardware and software. One issue with new standards is the pre-existing investment in hardware, software, training, and experience. The true cost of changing to a new standard (whole-scale abandonment of existing programs, data structures, etc.) is prohibitive. Yet as discussed previously, some standardization can greatly enhance maintenance and productivity. A comprehensive debate over data standards is beyond the scope of this document, but a brief discussion of the design, effect, and examples of standards is included below.

One example of a comprehensive database standard is the Environmental Protection Agency (EPA) STORET format for water chemistry data. STORET is a standard primarily because it has been the main repository of water chemistry data for many years, and computer programs are available to analyze STORET data. Failure to support the STORET standard, as is true for most park water data, risks incompatibility with other agencies and NPS divisions. Although widely used, STORET has its share of "standard problems." STORET is a mainframe FORTRAN system which is only now being updated to use modern software tools and interfaces. Until the update is complete, interacting with STORET is awkward. More importantly, data stored in STORET have little quality control and are not certified by the EPA. As such, STORET is a "user beware" system. STORET data are used in the Baseline Water Quality Inventory and Analysis reports produced by the NPS Water Resources Division for the I&M program. More information about the Baseline Water Quality Reports, the STORET system, and other EPA databases is included in an appendix.

On a different level, the ASCII format for data files is another standard (ASCII stands for American Standard Code for Information Interchange). ASCII files are often referred to as "text" files. ASCII data can be "imported" or "exported" to and from just about any computer operating system or program. Although ASCII is a "universal" format for sharing data, the format does not store many attributes. Thus, ASCII is primarily useful for raw data (like the kind being addressed in this document).

The examples above illustrate the need to identify, understand, and incorporate relevant standards whenever possible. Standards and/or standard import/export capabilities should be built into all database structures. When standards are adopted after a database is implemented, relational links and/or translator services for maintaining compatibility are needed.

Finally, standards should not inhibit practical application and experimentation. For example, the Dataset Catalog record design presented in the next chapter was developed as a standard set of fields for both a park-based and servicewide inventory, but parks may define and append additional fields to their dataset catalogs as needed. As long as the servicewide field definition is intact, individual park catalogs can be shared and aggregated into a comprehensive database. Good standards should provide utility and adaptability as well as portability.

Administrative Structure

Each park should define the administrative structure of data management personnel. The listing may be brief and totally encompassed within the existing organizational structure, but data management administration roles should be clearly designated. In the context of data management planning, failure to establish an administrative structure can lead to inadequate computer standards, difficulties in customer access to current datasets, and the loss of data integrity over time.

Clearly defined areas of responsibility and authority regarding data and computer issues establish chain-of-command and structured working relationships. Since management boundaries differ at each site--especially in a distributed organization like the NPS, where matters may need resolution at several levels (e.g., park, cluster, division)--each site must evaluate its own unique situation. Several data administration categories (computer support personnel, data manager, project manager, and end user) with brief job descriptions are reviewed in the appendix Data Personnel Hierarchy and Responsibilities.

Data acquisition considerations

When data are acquired by a park, either directly or indirectly, issues such as data ownership and copyright may need consideration. At the simplest level, any data collected by park staff or directly paid for with taxpayer monies belongs to the U.S. government. However, where outside groups or private entities are involved, the data

may not be free to the public domain. As much as practicable, data needed and used in park management should be owned, either outright or through contractual agreement, by a government agency. Several issues that may be addressed in the DMP are discussed below.

Permitting, agreements, contracts, etc.

Although casual and non-invasive data collection in parks is not controlled, anyone who collects samples or specimens of any kind (except legally harvestable items), marks trails or trees, or otherwise disturbs park inhabitants or environs while collecting or generating data is a data *collector* and must get formal permission to do this work. Data that are approved for collection require a research and/or collecting permit. As a condition of the collection permit, some sort of data ownership and proprietorship agreement could be formalized in writing before such data collection is allowed. In data collection contract agreements, the ownership of the data and resulting products should be clearly stated. Site- or project-specific documentation and procedures addressing data ownership should be reviewed in this section.

Data collector relationships

Various data collector relationships may exist that may need documentation as briefly discussed in the examples below.

1) *Volunteers, contractors, and cooperators*: data collected for park projects initiated or paid for by the park, NPS, or other government agency should belong to the park. Formalized agreements and/or contracts may be needed on a case by case basis and should be reviewed in the DMP.
2) *NPS and Interagency data endeavors*: although data ownership is probably not an issue, the full compliment of data management activities and concerns (e.g., QA/QC, database integrity, maintenance, dissemination, etc.) should be documented clearly in partnership agreements and reviewed in the DMP.
3) *Donations, loans, and copyrights*: any special conditions related to datasets that are archived or in use should be briefly documented with reference(s) to documents, contracts, etc. that affect the quality, maintenance, and availability of the data.

Quality Assurance and Quality Control (QA/QC)

Although the functional "lifetime" of hardware and software is decreasing rapidly, data are "forever." Both producers and users of data need to document and know the quality of their data. This is especially important for sharing data and is the intent of several government directives (NPS GIS Sourcebook, 1993). In short, data QA/QC cannot be over emphasized and should be formalized in the DMP. Generalized data handling procedures are provided in the last chapter of this document and may be referred to as needed. The Data Handling Procedures chapter may cover all data

management needs at some sites, but other scientific or specialized QA/QC considerations should be stipulated in the DMP. Items that should be addressed or referenced include:

Protocols and standards:
- applicable scientific measurement protocols
- applicable and documented SOPs

Verification, validation, and editing
- applicable and documented SOPs

Data documentation & metadata standards
- applicable and documented SOPs
- documentation data (e.g., Dataset Catalog, etc.)

Data summaries and analyses
- applicable and documented SOPs to evaluate precision and accuracy

Data maintenance (working datasets)

As with QA/QC, the Data Handling Procedures in the last chapter of this document may provide many of the SOPs needed for data maintenance. Again, where other or special considerations exist, they should be addressed. Items to consider include:

Update procedures
- applicable and documented SOPs

Version control
- applicable and documented SOPs

Added or derived data
- data derived or calculated from and included in the dataset
- data added from outside sources and included in the dataset

Backup routine
- applicable and documented SOPs

Documentation and cataloging
- applicable and documented SOPs
- documentation data (e.g., Dataset Catalog, etc.)

Legacy Data

A comprehensive plan for dealing with legacy datasets is a critical data management goal. Assembling and entering documentation for legacy datasets in the Dataset Catalog is an important first step. However, a plan for updating, validating, implementing, and archiving the legacy data should be documented in the DMP. If any legacy data are to remain inactive, out of date, or be considered for purging, these should also be addressed in this section.

Legacy Data List

A list of known legacy datasets should be included in this section. The list could be a report or table generated from the Dataset Catalog discussed in the next chapter.

Legacy Data Update Plan(s) and Protocols

This section should briefly document:

- Legacy datasets to be updated
- Legacy data updates completed
- Progress on current legacy data update endeavors
- Legacy datasets update schedule
- Any legacy data not slated for updating

Legacy data will often require specialized processing and filtering before it can be considered "valid" for use. In general, the same steps, albeit modified, as used to verify, validate, document, and maintain current working datasets will be needed for legacy data. Site- or dataset-specific procedures that are in use, being developed, or planned for legacy data updating should be included in this section. An example of comprehensive validation techniques used for complex legacy STORET data by the I&M baseline water quality reports is included in the Appendices.

Data security

Protection of important data from damage, loss, corruption, and/or vandalism are critical components of active data management. How secure are your datasets? Several concerns that should be reviewed in the DMP are discussed below. More information is also available in the Data Handling Procedures.

Risk analysis & system security

Resource datasets, especially those that are in day to day use, risk losing their integrity in numerous ways. An analysis of the risks to datasets and system security will head off many potential data disasters. Some security measures, such as office door locks, media safes for back up copies, and network login procedures, may already be in place, but a comprehensive plan to maintain data integrity is needed.

Duplication of critical files

As discussed in the Data Handling Procedures, critical files should be backed up and/or duplicated on a regular basis. Critical files and datasets that are in general use and/or on multi-user systems may need replacement on a moment's notice. Critical files should be identified, and local copies of master files should be available for fast replacement. This information may be reviewed or listed in this section.

Equipment contingency planning

As discussed in the appendix Computer Protection and Maintenance, failures of critical equipment must be repaired or replaced efficiently. A brief review of or reference to a written equipment-failure contingency plan should be included in the DMP. The importance and needs of this section will vary from site to site.

Access control

One of the best ways to provide for data security is to restrict access to the master dataset. Whereas dissemination of a dataset may have general restrictions (e.g., archeological or paleontological sites locations), master and archival datasets should have limited access by almost everyone. Access to master datasets may be restricted via a combination of physical location (locked doors, safes, cabinets, etc.) and/or software mechanisms (write protection or password protection) to prohibit unauthorized personnel from accessing or altering the master data. Access control documentation and procedures should be addressed here and in the Administrative Structure paragraph above.

Data archives and storage

Documentation and Tracking (Dataset Catalog)

A useful tool for dataset documentation and tracking is the I&M Dataset Catalog presented in the next chapter. Individual sites may append fields to the dataset catalog for site-specific needs for data documentation. Documentation should contain sufficient information that a future user not involved in the data collection phase can use and base decisions on the data with confidence. Documentation procedures and implementation strategies should be reviewed in this section.

Metadata standards

Existing and proposed metadata standards are discussed in Section I and the Appendices. One possibility to implement metadata documentation is with fields appended to or a relational database associated with the I&M Dataset Catalog. Pertinent metadata standards and activities should be reviewed in this section.

Physical location of duplicate datasets

The physical or geographic location of duplicate archived datasets is an important concern. For some locations, archiving duplicate data in another building in a secure data safe or vault may suffice to protect the data. On the other hand, parks in geologically unstable areas, such as the Channel Islands within the San Andreas transform zone, need to arrange for duplicate data to reside in an area that would be unaffected by a disaster such as an earthquake. In addition to regular on-site back up

and maintenance, some sites will need to plan for regular updates of duplicate data (and system backups?) stored at another location. This may be accomplished via an agreement with another NPS entity. At this writing, a centralized data archive does not exist, so data storage agreements must be made between individual units.

Scheduled rotation of storage media

Storage media, such as floppy disks, magnetic tape, etc., generally do not have an indefinite shelf life. When these media are used to store master datasets over time, their integrity should be questioned and inspected regularly (at least annually). For example, a dataset stored on floppy disks in plastic sleeves at Shenandoah was lost by degradation of the plastic sleeves. Planned media inspection, rotation, and/or replacement will ensure the long-term viability of the stored data.

Data purging policy

Occasionally datasets will need to be abandoned and purged from the system. However, before any dataset is relegated for eternal loss, several concerns should be considered. Just because a dataset has no present application does not mean that it has no value and should be destroyed; purging should be a last resort. Most often, preserving and cataloging the data for the future may be the best policy--even if the data seem irrelevant today. Considerations to be addressed before purging include:

Do the data reside elsewhere for retrieval if needed in the future?
Do the data, dataset, or represented endeavor have any:
- historical or cultural value?
- scientific value?
- administrative value?
- educational value?
- past, present, or future value?

If the data still appear to be valueless, the purging proposal should be reviewed by all interested/affected parties. If all concerns are satisfied, the dataset may be destroyed. A systematic policy for dataset review before purging should be developed if needed.

Data applications (standard reports, query tools, decision support)

Each NPS unit will have several data applications that will range from servicewide to site-specific programs. Documentation of the park's data applications will aid personnel at all levels to understand the flow and uses of the park's data and data management programs. Standard reports, query tools, and decision support implementations that are planned for or result from data management activities should be listed and briefly explained in this section. Individual program(s) and documentation may be included in the Appendices.

Data dissemination

Freedom of Information Act (FOIA)

The Freedom of Information Act (FOIA) legislates a presumption that all records in the possession of the agencies and departments of the Executive Branch of the U.S. government are accessible to all people. FOIA and the Privacy Act establish regulations and procedures for accessing government data and information and formulating cost recovery associated with dissemination (NPS GIS Sourcebook, 1993). In short, public information should be accessible to the public at a reasonable cost. Although a full discussion of the FOIA is beyond the scope of this document, some generalizations can be made. In general, most data or documents that do not contain personal information about employees or specifically legislated exceptions (such as archeological, paleontological, and cave entrance locations) must be released under FOIA. However, there are some exceptions that relate to endangered species. When presented with a FOIA request, consult with your park or cluster FOIA officer. The FOIA officer will assist in determining: is the park required to release the data and if not how to refuse the request, is there precedence, the scope of the data for release, charges for reproducing the data, and other issues. *The first time a particular data set is released sets a precedence; thus take care in your practices*. For more information, refer to the NPS GIS Sourcebook (1993) or an NPS coordinator in the area of the sensitive data. FOIA concerns/policies for sensitive datasets should be reviewed here.

Availability and version of active and/or legacy data sets

Ideally, datasets will be fully validated and documented before dissemination. However, many legacy datasets will have Dataset Catalog entries long before the data are updated and checked for accuracy and reliability. Thus, the version of the disseminated dataset is important as well as the metadata information that is supplied with it. Also, if specific agreements exist concerning the copyright, ownership, and/or dissemination of a dataset, these must be respected.

"Request for Data" forms, policy, disclaimer, and fees

Forms. To keep track of requests for data and the resources devoted to data dissemination, "Request for Data" forms are suggested for each transaction. The forms may reside in a computer or physical database (or both). An example "Request for Data" form is included in the Data Handling Procedures. Site specific forms should be included in the Appendices.

Policy. A formal policy and procedure to handle data requests should be developed and stipulated in the DMP. The policy might include: a specific contact person or office, a required written Request for Data, protocols for assembling and sending the data, a standard metadata template with liability disclaimer to include with the data,

procedures for billing/collecting handling fees, etc. These and other considerations reviewed in the Data Handling Procedures should be included here.

Example Distribution Liability and Disclaimer Notice:
(Note: example is based on the Metadata Minitemplate and has not been Solicitor reviewed.)

The National Park Service shall not be held liable for improper or incorrect use of the data described and/or contained herein. These data and related graphics or documentation are not legal documents and are not intended to be used as such.

The information contained in the data is dynamic and may change over time. The data are not better than the original source(s) from which they were derived. It is the responsibility of the data user to use the data appropriately and consistent within the limitations of geospatial and/or natural resource data in general and these data in particular. Any related graphics or documentation are intended to aid the data user in acquiring relevant data; it is not appropriate to use any related graphics as data.

The National Park Service gives no warranty, expressed or implied, as to the accuracy, reliability, or completeness of these data. It is strongly recommended that these data are directly acquired from an NPS source and not indirectly through other sources which may have changed the data in some way. Although these data have been collected, acquired, and/or processed at the National Park Service, no warranty expressed or implied is made regarding the utility of the data on any system or for general or scientific purposes, nor shall the act of distribution constitute any such warranty. This disclaimer applies both to individual use of the data and aggregate use with other data.

Fees. Formal guidelines for fee implementation have not been developed. However, reasonable costs associated with data dissemination requests may be charged to the user. Fees must be determined on a site by site and project by project basis. Any fee system in use or planned should be included here.

Data formats, conversions, and distribution mechanisms

In general, the responsibility for converting data from the NPS format to the requesting user's format ultimately falls on the outside user. However, whenever possible, requests for specific formats should be honored as a customer service function. Therefore a list of available output formats will be useful for dissemination planning. Conversions that may be time and/or resource intensive might carry an appropriate fee (which must be calculated in advance). As with data formats, several distribution mechanisms might be available and should be documented. Possible distribution mechanisms might include: floppy disks, e-mail, central archive (Web or FTP site), etc.

Tracking/auditing of disseminated data and resulting products

A system for tracking disseminated data and resulting products (e.g., an academic paper from park-provided resource data) is suggested. This system might consist of analog forms or a relational database associated with the Dataset Catalog. Periodic updates of the park's bibliographic database should also be planned.

DMP SECTION 3: Accomplishments, Current Activities, and Long-term Goals

Data is a prominent feature of any I&M Program, and one measure of the value of the program is the extent, accuracy, and availability of the long-term data it generates. Although data management often requires tedious attention to database details, the *service function* of sharing data and making it accessible to users is equally important. In addition to the implementation and periodic re-evaluation of data management strategies, data managers also need to market the value of the natural resource data to park management and make it generally available to more users. This section of the DMP communicates the progress made to date, reports the current status and prioritized task list of planned work, and documents long-term goals for the data management system.

Accomplishments

Documented completion of data management tasks illustrate the progress and evolution of an active data management scheme. Even with the first draft of the DMP, several data management tasks will be accomplished in the planning process. An annotated outline, such as in the example box at right, can be adapted as required to record program accomplishments.

Example Accomplishments List

Computer standardization
- DOS 6.22 installed on all PC systems
- 4 PC systems upgraded to 16 Mb or RAM
- 2 PC systems upgraded to Microsoft Office
- etc.

I&M Dataset Catalog 70% complete

Legacy datasets updated and validated
- Stream chemistry and hydrology database

Analysis of long-term datasets initiated through contractors
- Weather and climate data digitized

Stream chemistry and hydrology data forms updated
- New data protocols and forms were created and used in the 1995 field season for obtaining and recording stream chemistry, habitat, and discharge data.

Ethernet Network installed and operational

Internet access (at least for e-mail)

Software standards emerging
- 5 more copies of Microsoft Access purchased

I&M Staff introduced to database software and related tools

Servicewide Bibliographic Initiative completed for park

First draft of (this) Data Management Plan written

Involvement with data issues outside the park

Polycorders used: mixed results

Standardized backup procedures in place

Current Activities

Both ongoing and planned activities also need to be documented to illustrate pending data management endeavors as well as the current work load. One method to accomplish this is via a prioritized task list that details both current and future projects. In addition, a comprehensive, prioritized task list will greatly aid the development of a feasible budget in the next section. An annotated outline, such as in the example box below, can be adapted as required to document current activities. An exhaustive task list may be included in an appendix and concisely outlined in this section if needed.

Example Prioritized Task List (excerpted from Shenandoah's draft DMP)

Purchases (all priority 1)
- Media Safe(s)
- LAN Anti-virus software
- LAN group scheduling software
- CASE software for schema design and evaluation
  - Window's-based Computer Aided System Engineering (CASE) tool.
- RAM upgrades for PC's (to 16 Mb)

Review all data forms (1)
Data-entry modules: creation and/or evaluation (separate priority rankings)
Stream chemistry and hydrology data (1)
- To simplify data entry, database input screens will mimic the new forms and the module will perform all calculations and summaries that were formerly done by hand. STORET, as a national water database standard, will serve as our target of compatibility for the data structure wherever applicable.

Data Management tasks module
- (3) A relational module is needed to allow users to enter and query data related to the *status of datasets*. The module will hold information about current versions, editing history, extent (spatial, temporal, record numbers), and available products from specific datasets. It will include or link to a main metadata information database, historical documentation of the data, and references to summary reports and publications resulting from the data. Linkages to IAR and RMP would also be beneficial.
- (1) As an interim solution, an official database ledger notebook will contain the most current information on our master data--current date and time stamps, number of records, last update, physical structure and field descriptions. Edit and update work forms will be each file's "history."

Metadata format and development (1)
Put copies of all data on file server (1)
Full Internet access from desktop PC's (2)

Complete legacy data entry and error-checking of remaining datasets (1)
- Some raw field data were never computerized, only calculated for summaries. These data need to be fully entered and validated.

Observational database evaluation (2)
Update Park Species List(s) (2)
Continue training staff with tools and practices of good data management (1)
Public relations efforts (1)
Data dictionary and schema for integrated system of all existing datasets (1)

Long-term Needs and Goals

Long-term needs and goals for the data management system need to be considered, documented, and periodically re-evaluated along with the evolution of the DMP and program. Items included in this section will necessarily range from an educated wish list to desperately needed items that could not be funded in past budget plans. Again, an annotated outline format might best relay this information in the DMP. An example listing of long-term needs and goals is included in the example box below.

DMP Evolution and Update Schedule

The objective of data management planning is to enable NPS personnel to effectively manage ever-increasing data loads--both now and in the future. The focus of informational needs will vary over time as will data collection and analysis loads. For example, as critical legacy datasets are cataloged, verified, validated, and archived their management needs will decrease significantly. Hence, the DMP will necessarily evolve with time to correspond to changing demands.

Example Long-term Needs and Goals List

- Networking and communications
  - Hire a full time system support person
  - Expand Local Area Network
  - Direct cc:Mail Access for all workstations
  - Study Feasibility of a Wide Area Network
  - Complete Internet Access
- Documentation retrieval system/integration
  - I&M reports
  - RMP's and IAR's
  - Bibliographies
- Park-level data coordination
  - Train/involve all staff with data management
  - Identify, locate, and catalog Park information
  - Development of interactive database system

Evolution How will data management needs change at the park over time? Planning for the initiation, continuance, and completion of data generating projects is imperativel. As in Resource Management Planning, the DMP should document future needs and projects. Devise and briefly review a strategy for the DMP and associated projects to develop with time (e.g., a scheduled review process). Remember, that all goals will not be immediately achievable; the DMP should allow for incremental growth as needs and capabilities emerge.

Update Schedule Action-oriented management plans themselves require periodic review and revision. How often and with what focus will the DMP be reviewed? Time frames and will vary by site, but a regular DMP review and update are essential. An optimal schedule may be to append annual reviews to the DMP for comprehensive inclusion every 5 years.

DMP Section 4: Implementation of Data Management Plan

This section is where the DMP interfaces with the park Resource Management Plan (RMP) and/or other program funding sources. In that light, both this section and the Executive Summary should be as complete, concise, and accurate as possible to quickly enhance a reader's general understanding of data management needs, capabilities, and goals. For additional information about the RMP refer to the Resource Management Plan Guideline (1994) and annual RMP updates.

Feasibility Evaluation/Budget

The feasibility evaluation should contain a brief cost analysis and/or budget that reviews funding needs and sources, staffing needs and allocations (i.e., FTEs), and other pertinent resource and budgetary items. The DMP may refer to a more detailed budget document or the park RMP as needed. The information may be presented in an outline or annotated outline format. Several potential cost/budget categories are shown in the example box.

Example Feasibility Evaluation/Budget Outline

Staff and time requirements

Hardware and software procurement and maintenance
- Evaluation and update of network needs
- Evaluation and update of hardware needs
- Evaluation and update of software needs
- Media rotation for archives and backups

Recovering and updating legacy datasets

Integration of the Parkwide information system

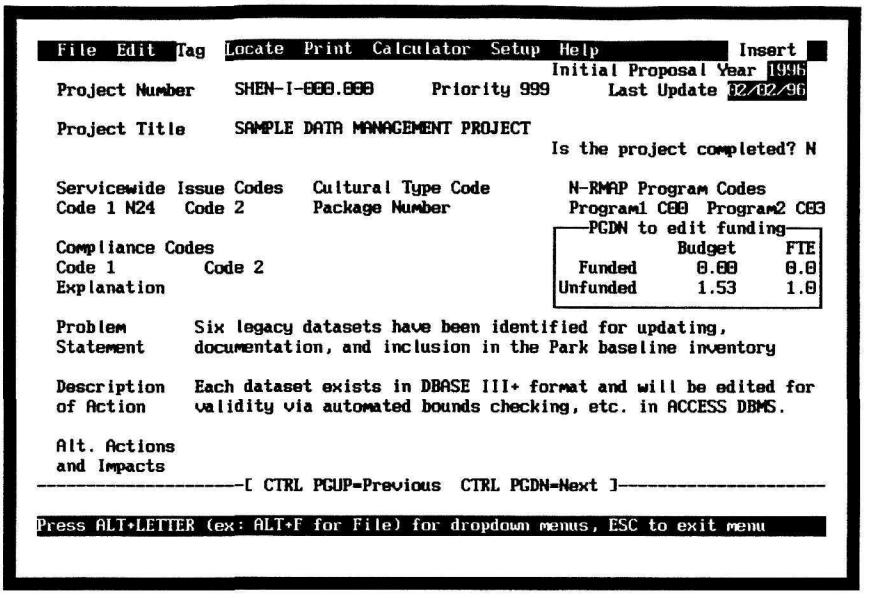
Locating and computerizing other types of data

Data dissemination and auditing

Interfacing the DMP with the RMP

At some point, the DMP will need to interface with the park and servicewide RMP process. Funding sources for extant data management efforts and personnel will already exist in some capacity, and these funding sources may need to remain in place as such and/or converted or added to for budgetary considerations. For more information about the RMP process at an individual site, consult with the Chief of Natural Resources at the park. In general, the RMP requires projects seeking funding to complete a RMP Project Statement with related project codes and priority number and enter the information into the park RMP database. Data management-related projects may need to be included in a comprehensive project or as individual projects as dictated by site-specific concerns. Several RMP project statement requirements and suggestions are illustrated in the example box below. For additional information about the RMP refer to the Resource Management Plan Guideline (1994) and annual RMP updates.

Example Resource Management Plan Software
project statement data entry form and codes



In the RMP Project Statement, Project Number and Priority codes are park-specific assignments. The illustrated Servicewide Issue and N-RMAP Program Codes are suggested but may be adapted as needed on a site-specific basis (N24 = Other Natural, C00 = Collections and Data Management, C03 = GIS/Data Management). A sample RMP Project Statement is included in the Appendix.

Implementation Schedule

When will the planned data management strategies become a reality? This section should concisely outline the implementation schedule for all aspects of the DMP. If a Prioritized Task List has been completed in the previous section, the list may be generalized and outlined here with dates attached. Projected initiation and completion dates should be as accurate as possible. An exhaustive implementation schedule can be included in the Appendices if desired.

Chapter 3 - I&M Dataset Catalog

Introduction

At the park-based data manager's meeting in Denver (1995) and Fort Collins (1996), the group identified creation of a catalog of park-based data as its highest priority. The purpose of the catalog is to provide a single source for information *about data* on each park's natural and cultural resources with the additional goal of combining each catalog at the national level into a servicewide catalog of park-based data. A format was developed for catalog entries that consisted of a single page per dataset, and the group discussed what constitutes a "single" dataset. The group also reviewed the Investigator's Annual Report (IAR) and Resource Management Plan (RMP) style of reporting. Although neither the IAR nor the RMP supplies all the information needed in the Dataset Catalog, the IAR subject list was adopted. The group also adopted generic cataloging terms used in national metadata standards (which are much more detailed than the current effort) that should enable linking of the Dataset Catalog to other 'catalogs.' Steve Tessler, the data manager for Shenandoah National Park, volunteered to lead the effort by turning the catalog format into a database that could support ad hoc queries (e.g., "What data do we have about mammals?"). The current Dataset Catalog system has evolved from the original work at the conference and Shenandoah National Park.

The current version of the I&M Dataset Catalog consists of a database with 36 fields per record that can be exchanged in .DBF, .MDB, and/or ASCII format for sharing between NPS units and inclusion in the servicewide database. Five relational database tables provide look-up references for catalog records. An on-line database system using the Internet is also being developed. The catalog is not intended to be an exhaustive metadata listing but may provide the platform for initiating comprehensive metadata standards. For servicewide use, the 36 fields are strictly defined and should be adhered to rigorously at the park level. On the other hand, parks may append additional fields to their park Dataset Catalog for more complete dataset documentation--both now and in the future. For example, the servicewide database structure does not contain any memo fields, but each park may wish to append a memo field to include a complete abstract in their dataset documentation. A generalized overview of the I&M Dataset Catalog Field Descriptions is presented in the example box below.

I&M Data Catalog Field Descriptions

CATDATE - Last update of this catalog record

PARKCODE - Park's 4 letter identity code (1st 2 letters of 1st 2 words).

PARKNAME - Park's name and contact information (relational database table)

ADMIN - SSO/Cluster for aggregate sorting (relational database table)

SUBJECT - Database subject (from IAR subject list)

KEYWORDS - Descriptive key words provided by user (Keyword Thesaurus available)

TITLE - What the dataset is called - brief descriptive title (70 characters)

VERSION - Version of dataset for version control and duplicate entry checking

PROJ_ID - Park identification of data collection project (PARKCODE + ?)

DESCRIPTION - Brief content, purpose, goals, design, methods, limitations, etc. (40-50 words)

RELDOCS - Documents related to the database such as protocols, reports, etc. (35-40 words)

DATES - Time period of data collection for this database - start and end dates (2 dates)

MULTI-DATES/NOTES - Start/end dates for multi-time data (up to 8 periods of record - 16 dates)

STATUS - (NEW, ACTIVE, INACTIVE, HISTORIC)

UPDATFREQ - Maintenance update frequency: how often data is collected (e.g., 2/year, etc.)

PLACES - Spatial relationship to park: Inside Park (IN), Outside Park (OUT), Park Area (IN&OUT)

LOCATION - Brief description of study area location - place name, site, etc. (15-20 words)

LAT - Approximate Latitude of the center of the data study area (dec. degrees, rel. database)

LON - Approximate Longitude of the center of the data study area (dec. degrees, rel. database)

DATATYPE - Georeferenced Digital Raster (GEORAS), Georef. Digital Vector (GEOVEC), Non-georeferenced Digital Raster (DIGRAS), Non-georef. Dig. Vector (DIGVEC), Georeferenced Digital Database (GEODB), Non-georef. Digital Database (DIGDB), Organized Analog Database (ANAORG), Unorganized Analog Data (ANAUNO)

TABL/LAYR - List of database tables or GIS layers in a composite database (15-20 words)

SCALE - Point data (POINT), Map Scale(s) (e.g., 24K, 100K), Raster cell sizes (e.g., 1000 feet)

QUALITY - Unknown (UNKNOWN), Unverified/Unvalidated (NONE), Formally Verified (+VER), Formally Validated (+VAL), Formal Metadata Standard Met (+META)

FORMAT - Software version, digital file, and/or analog format used for data (40-50 words)

ORIGIN - Contact info for principal investigator/study generating original data collection/design

CONTACT - NPS Park Division, address, phone, etc.

DISTRIBUTION - Distribution mechanism (e.g., WWW/FTP address, contact person, etc.)

FILE_LOC - Location/computer system where the digital data resides

ACCESS - Data access options - No restrictions (PUBLIC), Access restrictions (RESTRICTED)

Phase I: Data Resources List

The first step in cataloging datasets is to identify and list all of the park's various (and often extensive) data resources. Subsequently, Phase 2 will actually generate the catalog. The Phase 1 list will be appended to the park's Data Management Plan.

There are many "states" in which datasets may reside: hardcopy only (analog), digital, and other data stored and/or managed by outside agencies, contractors, etc. Also, any "old" and/or other park data should be included if known (e.g., thesis datasets).

As a first approximation, the data manager should compose a preliminary dataset list and circulate it among appropriate park personnel and investigators for additional input. Inquire about potentially-missed items that should be included (e.g., projects given collecting or research permits). The list should be comprehensive. In most cases, the individuals that are responsible for or supervised the collection of the data (e.g., Project Managers) should fill out the Dataset Catalog Entry Form. There may be much needed discussion about who is actually "in charge" of the support for each dataset. Indeed, this is a good time in the data management planning process to ascertain and/or assign responsibility of legacy datasets to individuals. In addition, all datasets should be evaluated for cataloging as individual, aggregate, or comprehensive datasets on a case by case basis. The main goal at this stage is to build a list of *all* the park's data resources as illustrated in the example below.

Example Dataset Identification List - Shenandoah National Park

Dataset Identity	Data Collected By	Staff Contact/Supv.
------- *ANIMALS* -------		
Invertebrates: LTEM Aquatic	SNP	Manager(s) Name(s)
Invertebrates: Aquatic, Other Agency	EPA	Manager(s) Name(s)
Invertebrates: Terrestrial	?? (Simpson)	Manager(s) Name(s)
Pests: Gypsy Moth, Pop'n Monitoring	SNP	Manager(s) Name(s)
Pests: IPM Spray Program	SNP, USFS	Manager(s) Name(s)
Reptiles: Snake Surveys, General	SNP	Manager Name
Amphibians: Aquatic, Basinwide	U Richmond (Mitchell)	Manager Name
Mammals: Bear Survey	SNP	Manager(s) Name(s)
Birds: MAPS	??	Manager Name
Birds: Peregrine	SNP	Manager Names Assistant 1&2
Fish: Basinwide Surveys (BVET)	VaTech (Dolloff)	Manager(s) Name(s)
Fish: "FISH" Project (Acidification)	UVa (Bolger)	Manager(s) Name(s)

Dataset Identity	Data Collected By	Staff Contact/Supv.
------- *PLANTS* -------		
Forest Vegetation: LTEMS	SNP	Manager(s) Name(s)
Forest Vegetation: Cover Types	SNP	Manager(s) Name(s), GIS
Forest Vegetation: Old Growth	SNP	Manager(s) Name(s)
Plant Distribution: General Diversity	SNP	Manager(s) Name(s)
Plant Distribution: Rare Plants	SNP	Manager(s) Name(s)
Plant Distribution: Communities	SNP	Manager(s) Name(s)
Lichens: Indicator Species	??	Manager(s) Name(s)
Fungi: Diversity	??	Manager(s) Name(s)
Water Quality: Stream Chem/Disch.	SNP	Manager(s) Name(s)
Water Quality: STORET	EPA	Manager(s) Name(s)
Water Quality: Basinwide Survey	UVa (Webb)	Manager(s) Name(s)
Water Quality: Rainwater, Acidity	SNP	Manager(s) Name(s)
Water Quality: Wastewater	SNP (Maintenance)	Manager Name
Air Quality: Meteorology	SNP	Manager(s) Name(s)
Air Quality: Visibility, Camera	SNP	Manager(s) Name(s)
Air Quality: Visibility, Transmissom.	SNP	Manager(s) Name(s)
Air Quality: Pollutants, Other Agency	EPA	Manager(s) Name(s)
Streams: Habitat Description	SNP	Manager(s) Name(s)
Streams: Fish Habitat (BVET)	VaTech (Dolloff)	Manager Name
Geology: General	SNP	IMS (GIS)
Soils: General	SNP	IMS (GIS)
------- *Reports & Bibliographies* -------		
NPFauna/NPFlora for SNP	UCDavis (Quinn)	Manager Name
Va Heritage Survey for SNP	?? Va Heritage	Manager Name
ANCS (specimens)	SNP	Manager(s) Name(s)
Resource Mgmnt Proposals (RMP)	SNP	Manager Name Asisstant
Investigators Annual Reports (IAR)	SNP	Manager Name Asisstant
I&M Annual Reports	SNP	Manager Name
Park Planning Documents: Extensive	SNP	IMS
Park Bibliographic Database	SNP (planned)	CfR, IMS
------- *Miscellaneous* -------		
Tree Tag Database & Identifier	SNP (planned)	CfR
BVC Observation Database	SNP (Education)	I&E
GIS Spatial Databases	SNP	IMS (GIS)

Dataset Identification

The following three statements and examples will help define what constitutes a separate line (an eventual dataset catalog entry) in the data resources list above. The criteria also suggest when to combine entities that share the important traits of objective, methodology, and/or investigator.

1. A "single dataset" in the catalog contains data collected for a single objective.

   Example: The park's aquatic macroinvertebrate monitoring protocol calls for taking insect samples, measuring discharge, recording water chemistry parameters, and performing a suite of habitat measurements. Whereas the "meat" of the project is getting the bugs, the other data are important companions and may be critical to the interpretation of patterns of diversity and community in the insect data. However, the companion data are not unique to macroinvertebrate monitoring, so the bug data stands alone.

   The chemistry, discharge, and habitat data also stand alone. Depending upon the objectives of the field foray, one, two, or all could be sought, with or without any biological data. Stream chemistry measurements are essentially the same "thing" (and viewed the same) whether the task is bug collecting, fish shocking, or demonstrating techniques during I&E training sessions. It is a separate data/information entity.

2. Where the same information objective is pursued using different methodologies, each should be cataloged separately but share Subject and Keyword elements.

   Example: Air quality sampling at Shenandoah provides an example where visibility in the park is measured using camera, transmissometer, and via the IMPROVE (multi-agency) devices and protocols. Each dataset tells us something about visibility, but their methods, time intervals, and resulting datasets are quite distinct; thus, each gets its own catalog entry. Using the same keyword(s) (e.g., visibility) will identify *all three* datasets, and allow the catalog explorer to make their own distinctions about usability for their particular needs.

   A second example is Shenandoah's three "bird census" methodologies: the MAPS program (net capture), Breeding Bird Survey (point counts), and transect surveys (visual and voice). Whereas each bird census results in complementary, but different, kinds and qualities of data; they are cataloged separately.

3. If the same kind of data is gathered by separate investigations and/or different SOPs, each dataset gets their own catalog entry.

   Example: Stream chemistry parameters are measured by park staff during resource monitoring activities, wastewater effluent evaluation, and potability studies. Even if all three sought only temperature, pH and DO, the difference in

who collects the data is significant and can reflect real differences in equipment, methodology, training, accuracy, and resolution. Each gets its own catalog entry. This redundancy is instructive and may indicate the need to consolidate these activities or at least standardize the equipment and procedures used.

In another example from Shenandoah, water chemistry characterization is the target of several studies done by the University of Virginia. *Whenever* an outside investigator generates the data it is a separate catalog entry. Again, all of these studies would share the bulk of their keyword terms and be identified as a group during any query of "water chemistry" studies in the park.

Phase 2: Cataloging Datasets

Preparing a catalog entry for a dataset takes about 5-10 minutes each if you are intimately familiar with the work. Filling in the form may also be quicker if the corresponding RMP or IAR for the project that collected the data is reviewed. An example completed form is provided at the end of this chapter, and a blank Dataset Catalog Entry Form is provided in the Appendix.

Not all field values in the actual catalog database are on the form; some will be assigned by the data manager or obtained from I&M/GIS, and others are redundant (e.g., park code, name, and address). Therefore, all fields on the Dataset Catalog entry form must be completed. The project manager's work on a catalog entry consists of:

1. Selecting a Subject and several Keywords
2. Writing a short dataset Title (i.e., an official title) and determining Version
3. Writing a concise (250 character or less) description of the data
4. Indicating the times, places, physical categories, etc. covered by the data
5. Listing important documentation relating to the data
6. Determining Time frame(s) for the data collection effort(s)
7. Assigning Status and Update Frequency information for the dataset
8. Determining Places and Location(s) that the data were collected.
9. Determining the Latitude and Longitude of the center of the study area
10. Determining the Data Type and names of Tables and/or Layers
11. Assigning Quality values and determining Scale(s) of the dataset and/or layers
12. Identifying the current data Format(s)
13. Identifying and documenting the Origin of the dataset
14. Identifying and recording the Park Contact person(s) for the dataset.
15. Considering the sensitivity of the data for possible Access restrictions

Instructions for the Dataset Catalog Entry Form

This section contains generalized instructions for completing the Dataset Catalog Entry Form. Numbers in parentheses refer to the respective field widths for the topical catalog records.

Subject and Keyword Terms. (30 & 70) Subject (up to 30 characters) and Keyword (up to 70 characters) terms are to be selected from the IAR subject list (mandatory) and the Servicewide Bibliographic Database Keyword Thesaurus (optional), respectively. The Subject term is general, and there are more than 50 terms to choose from. The IAR Subject List is included in the Appendix or may be available online in the Dataset Catalog software. The project manager will also choose appropriate Keywords from his knowledge of the data and/or from the Keyword Thesaurus. The Keyword Thesaurus is implemented as a Windows Help file so that it is searchable, clip-able, and paste-able into any Windows software. At least the first keyword should be selected as a general term with the others being more specific. Keywords should include common names, taxa, environment, site, etc. If this sounds complicated, rest assured that it is not. Scan the list(s) and the "picking" will be apparent.

Dataset Title and Version. (70 & 10) You are given 70 characters for a concise dataset title. Avoid the temptation to describe the project, and instead have the title *describe the dataset*. Consider that a subject or keyword search will result in a list of these titles, so the title should be distinct enough to direct the data explorer to the proper dataset(s) for their needs.

The dataset definition requires a Version assignment. If the dataset has no assigned version, now is the time to assign one of 10 characters or less. The Version identifier can be a simple alpha-numeric assignment that may be incremented with each successive quality control step, revision, and/or update. In the servicewide Dataset Catalog, the assigned Version will provide duplicate entry control.

Project ID. (20) The Project Identification field (up to 20 characters) should be assigned a unique, park-based value (ideally, the project's RMP Project Number if available). The Project ID may best be assigned by the Data Manager in coordination with the Chief of Natural Resources. The primary emphasis here is that the Project ID should be unique to the dataset/data project and not duplicate a different project's RMP Project Number recorded in the IAR/RMP databases.

Dataset Description. (250) The Dataset Catalog allows 250 characters (40-50 words) for a brief description of the dataset. The description must be concise and direct. Items included in the description may include:

the "project" from which the data are derived,
general methodology (#sites, sampling frequency, protocol, equipment),
reference(s) to concurrent/related data.

Be creative with the description, but *focus on the data*. You will probably go above the size limit with the first draft but will quickly be able to pare it down to the basic facts and write a concise yet complete description of the dataset.

Related Documents. (210) List the documents that relate to the dataset. Although some projects may have an extensive document list, the list must be edited down to 35-40 words (210 character limit) for the Dataset Catalog. The list may include brief titles, project proposals, published protocols, work plans, significant reports, analyses,

IAR/RMP IDs, etc. A later version of the catalog may link all documents in the service-wide Bibliography Database (by code) to datasets in the Database Catalog (also coded) to allow a more comprehensive examination of data and document relationships, but be selective this first time around.

Related Datasets: (140) List any other datasets that are closely related with the current catalog entry. Related datasets may include physical or biotic data collected at the same time or in close proximity to the current dataset.

Dates. (Time/Date & 160) Begin Date and End Date indicate starting and ending points (e.g., 12/13/1986-08/19/1996) of data collection. The overall time span of the dataset should be recorded as well as any individual periods of record if time gaps with no data collection occurred. For one-time-only studies, list the single year and give more details in Multi-Dates/Notes or the dataset description. For multi-time period datasets, include a list of the start and end dates for each period. 8 individual periods of record can be included in the Dataset Catalog.

Status and Update Frequency. (10 & 10)
The "Status" field (10) can have only one of three specified values:

NEW	a dataset in the planning/implementation/collection stage
ACTIVE	data are still being added to the dataset
INACTIVE	data is no longer being collected or planned for collection
HISTORIC	older, historic data collected in past projects

The "Update Frequency" field (10) states the interval at which new data are appended to the dataset, e.g., 2/year, daily, biannually. Use your own best descriptor for the update interval for this field.

Places. (6) Places refers to the spatial association of the dataset relative to the park. Were all of the data collected *inside* (IN) or *outside* (OUT) of the park boundary? Did the study area include places that are both inside the park as well as outside park boundaries (IN&OUT)?

Location. (100) Location refers to the geographic position of the dataset. Use descriptive term(s), place name(s), site(s), etc. (e.g., parkwide, North District, East Side Streams, Brooks Creek, Mt. Marshall, etc.). If the dataset covers relatively few sites distributed throughout the park, choose parkwide but also give other place names and/or some details in the Dataset Description.

Longitude and Latitude. (Double) The Longitude and Latitude of the approximate center of the study area must be determined for geospatial referencing. If the centroid of the park is the center of the study area, this location may be obtained from the table included in the Appendix; otherwise, the centroid location must be determined from a map. Report the Lat/Lon in *decimal degrees* (DD)--not degrees, minutes, and seconds (DMS). To convert DMS to DD, first divide minutes by 60 and divide seconds by 3600--then add the degrees, converted minutes, and converted seconds together.

UTM Zone. (Integer) The UTM zone of the dataset study area.

UTM Northing and Easting. (optional) If UTM Northing and Easting values are available, needed, or desired for the approximate center of the study area, enter them here. If the centroid of the park is the center of the study area, this location may be obtained from the table included in the Appendix; otherwise, the centroid location must be determined from a map.

Spatial Bounding Rectangle. (not on form - optional) Relational table in Dataset Catalog application with geographic/UTM coordinates of the bounding rectangle of the study area. For more details, see the software application or the Appendix.

Data Type. (6) The Data Type is the physical (i.e., analog) or digital medium in which the data exist. For identification of the Data Type in the Dataset Catalog, choose one of the acronym codes below:

Spatially Georeferenced Datasets

Digital Raster Data (e.g., GRID, IDRISI, etc.):	GEORAS
Digital Vector Data (e.g., ARC, Atlas, etc.):	GEOVEC
Digital Database (e.g., DBASEIII+, ASCII):	GEODB

Non-georeferenced Datasets

Digital Raster Data (e.g., SURFER, etc.):	DIGRAS
Digital Vector Data (e.g., AutoCad, etc.):	DIGVEC
Digital Database (e.g., DBASEIII+, ASCII):	DIGDB

Analog (i.e. Hardcopy) Datasets

Organized Database (e.g., field forms, tables, etc.)	ANAORG
Unorganized Database (e.g., unsorted files, etc.)	ANAUNO

Table or Layer Names. (200) For a composite dataset of several GIS layers or database tables, list the names of the multiple data record types (up to 200 characters). This field allows composite datasets to be listed as a coherent unit in the catalog without separate entries, but careful documentation of all tables/layers is essential for this scheme to be practical and useful for future reference.

Scale(s). (130) The Scale or spatial resolution of a dataset is an indication of its spatial accuracy. In addition, composite databases with several tables or GIS layers may vary in scale. The Scale or resolution should be listed respective to each table or layer (up to 130 characters). If the data consist of individual points check "Point". Otherwise, list the map scale as a representative fraction (e.g., 1:24K, etc.). If the data do not contain spatial coordinates, record N/A for "not applicable."

Dataset Quality. (15)
Data quality in the Dataset Catalog has three components:

1) state of formal verification,
2) state of formal validation, and

3) state of metadata relative to formal standards.

Some datasets have been "unchaperoned" by the current generation of resource management personnel and contain many unknowns about verification, validation, or use of the data. The Quality field is essential, so choose the most accurate answer regarding data quality. The Verification and Validation quality components are described in more detail in the Data Management Plan and Data Handling Guidelines. It is very important to identify any datasets that have not been critically reviewed. Selections for the Dataset Catalog are listed below.

Condition:	Value	Attributes
Unknown:	UNKNOWN	quality of the data including verification and validation attempts is unknown
Not Ver./Val.:	?	data are known to be unverified and unvalidated
Verification:	+VER	data have been *formally* verified as an accurate transcription of the original source (usually field forms)
Validation:	+VAL	data have been *formally* checked for out-of-range errors, spelling (sitecodes, names), correct dates, and logic errors (e.g., a 2 foot high tree with a 6 foot dbh).
Metadata:	+META	data have been fully documented to meet the applicable metadata standard

Dataset Format. (80) Specify the format(s) in which the dataset resides (e.g., software version, digital file, and/or analog format(s) used to manage the data). An example entry might include: DBASEIII+, ASCII, analog field forms. Analog formats, such as photographs, may also be listed here. In general, only the major data management and dissemination format(s) should be included.

Dataset Origin. (200) If the data were generated by park staff, write the name, *position*, and contact information of the project manager. If the data are obtained from outside contractors/cooperators or other sources, write the source, principal investigator's name, *affiliation, position*, and contact information in the spaces provided.

Dataset Contact Person. (30 & Table) The "Dataset Contact Person" field is the person and *postion* where inquiries about the dataset can be forwarded. The person filling in this catalog entry should be the Contact person for this dataset. Fill in the name, *position*, and contact information in the space provided.

Distribution. (50) How will the dataset be disseminated when needed? What mechanism will be used? If the dataset will be available over the Internet, give the IP address (e.g., www.something, ftp.something, etc.) or stipulate other distribution

mechanisms such as by diskette via park contact. If a contact reference is used, be sure that the contact information is included in the Contact field of the catalog record.

File Location. (50) Physical location or computer where the original data or database resides. This could be a file cabinet, local computer, server, etc.

Access Options. (10) The deliberate withholding of publicly-owned information is valid only when those data are specifically protected under existing law. The locations of caves/caverns and archeological sites on federal lands currently enjoy this level of protection, and similar defenses should be available for threatened and endangered species and critical habitats. Typically, though, there are no legal restrictions regarding access or use of resource datasets.

Check "Public" for Restrictions if these data can clearly be shared without reservation, or check "Restricted" and describe in the space provided the danger or threat of releasing this dataset to the public in its current state. Dataset restrictions must be evaluated and planned at an official level to protect sensitive data in response to FOIA and other requests. One procedure that may allow data dissemination without endangering resources is to prepare a restricted database without geospatial coordinate fields. Access concerns should be addressed in the park's Data Management Plan.

Example Dataset Catalog Entry Form

Copy this form before use! Fill out a separate form for each dataset. Numbers after the field name indicate the character limit for that field. All fields must be completed.

Subject (30): *Limnology*

Keywords (70): *Stream, Habitat, Water Quality, LTEM data*

Dataset Title (70): *LTEM STREAM HABITAT DATA FOR INVERTEBRATE STUDIES*

Version (10): *3.0* **Project ID** (20): *PARK*******

Dataset Description (250): *Data describing stream habitat and substrate composition for macroinvertebrate monitoring sites. Substrate type, pool/riffle ratios; cover, debris, canopy, others.*

Related Documentation (210): *LTEM Protocol Manual, I&M annual reports, 1005 Aquatic Macroinvertebrate Work Plan*

Related Datasets (140): *Aquatic macroinvertebrates, water chemistry/discharge*

Begin Date: *1989* **End Date:** *1992*

Multi-Dates/Notes: (160): *N/A*

Status (10): New __ Active X Inactive __ Historic __ **Update Frequency** (10): *2/YEAR*

Places (6): In X Out ___ In&Out ___

Location (100): *Piney Creek, Rocky Creek, Oak Creek, Stony Creek, Trout Creek*

Latitude (Double): *38.492593* **Longitude** (Double): *-78.46879*

Data Type (6): GEORAS __ GEOVEC __ GEODB X DIGRAS __ DIGVEC __ DIGDB __ ANAORG __ ANAUNO __

Table/Layer Name(s) (200): *INVHAB30.DBF*

Scale(s) (130): *POINT*

Quality (15): Unknown __ *Not* Ver./Val.(?) __ Verified X Validated X Metadata __

Data Format (80): Paper X dBASE X Lotus ___ WordPerfect ___ ASCII X

Other (describe): *Field Notebooks*

Data Origin (200): Name/Pos. *John Q. Biologist, Organization Position Title*

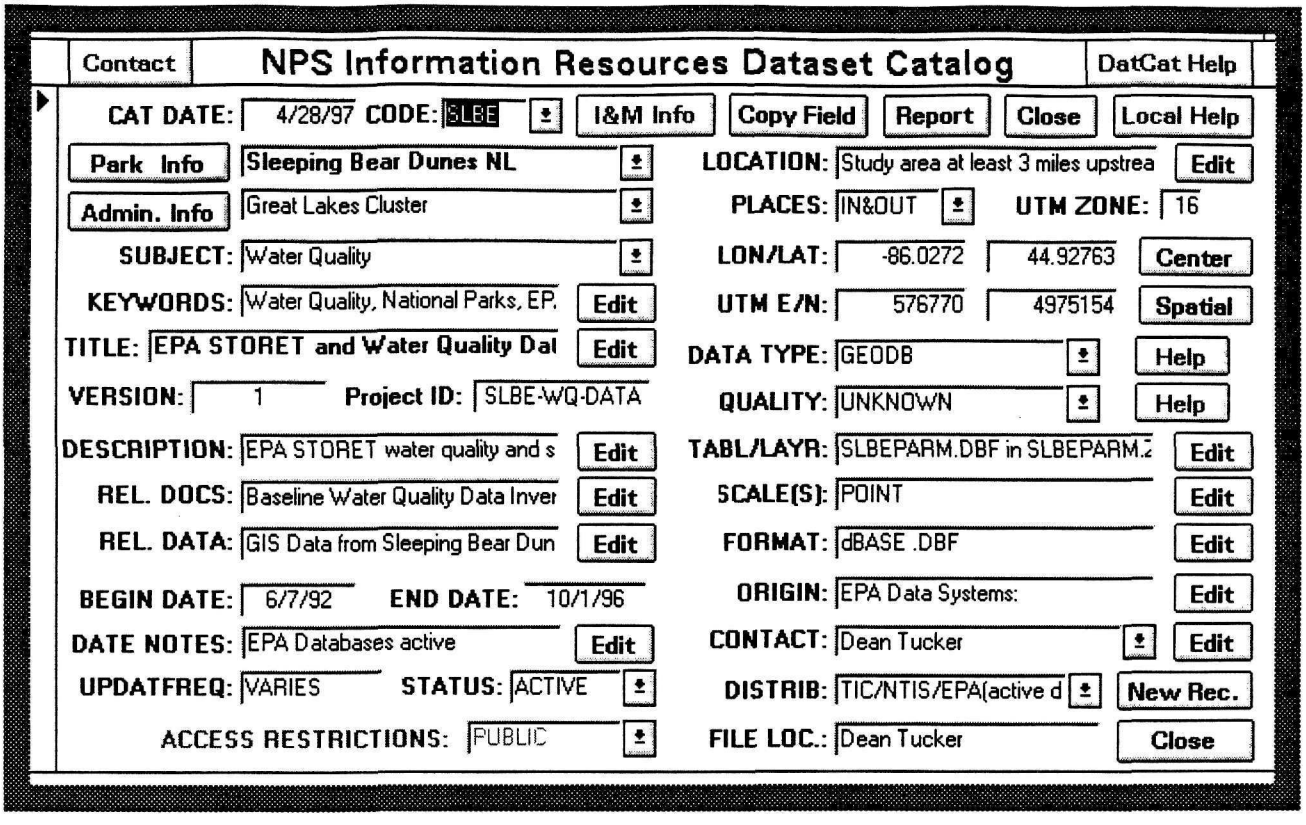
Affiliation, Contact Info *Organization, Mail Address, Phone, Fax, Email Address*

Dataset Contact: *Susan P. Datasetmanager, PARK Position Title*

PARK, Mail Address, Phone, Fax, Email Address

Distribution (50): *Dataset Contact* **File Location** (50): *I&M Server*

Access Options (10): Public X Restricted ___ (briefly describe below)



Dataset Catalog Input Form (Programmed in Microsoft Access)

The Dataset Catalog has been implemented with MicroSoft Access as a relational database incorporating the NPS address, IAR subject, the I&M bibliographic keyword thesaurus (Marilyn Ostergren), and latitude/longitude of park centroids/bounding rectangle databases. The Access database helps automate input, query, and report functions for the database and contains on-line help documentation. The Access database system is available from the NPS I&M Program (diskette or ftp).

Chapter 4 - Data Handling Procedures

Introduction

The guidelines in this chapter are intended to walk data handlers through the processes of entering data into the computer from field forms, verifying and validating the accuracy of the data, editing existing files, updating and validating legacy datasets, archiving and documenting master datasets, preparing data for dissemination, and developing backup strategies for their data and other work. The individual sections contain general instructions for the basic operations used in careful data handling.

The main purpose of this chapter is to provide basic standard operating procedures (SOPs) for data handling to ensure the quality and integrity of the data over time. The secondary purpose is to guide and instruct individuals with little knowledge or experience with data handling in a process that is clearly explained and easy to follow. Essential documentation procedures for tracking what is done to (or with) the data are included. Generally in the past, no usable notes were kept about the types of activities that were preformed on active datasets or if the data were carefully checked for errors, and this translates directly into serious doubt about the long-term credibility of the data. Indeed, as older data are reviewed, evidence for poor maintenance and systematic errors will be discovered frequently within the files. Consequently, considerable effort will be needed to "clean up" legacy data files, and this effort should not be wasted by careless activities of subsequent data handlers.

Several data management terms of general importance are discussed below. The remainder of this chapter presents more specific data handling guidelines.

Version Control is the process of documenting the temporal integrity of files as they are being changed or updated. Change includes any alteration in the structure or content of the files, and such changes should not be made without the ability to "undo" mistakes caused by incorrect manipulation of the data. The instructions in this document guide the user through incremental steps in the various procedures, and whenever a set of changes is complete the user should save the file with a unique name. Typically, numbered digits in the file name indicate which version of a file the user is working with (in addition to careful notes about the status and content of the version), so that if errors in data handling occur, the user can return to the previous numbered version and start again. In reality, version control is simple insurance for maintaining data integrity, and using good version control should be routine for all data handlers. An example of version control in practice is given in the Data Editing section.

Archiving is the process of making and maintaining copies of "*Master Data*" for the purposes of secure storage and easy retrieval. Master Data are up-to-date reference

copies of data files that are well documented and fully error-checked. Editing tasks and other activities with data are never made to the actual master data but with copies of the master. After editing, the fully documented and revised dataset, complete with new version number, is archived as master data. Archiving, therefore, is done with master data.

The simplest archiving procedure is to have at least two locations where master data and their associated documentation reside. Additional copies of master data that represent "milestones" (such as individual year subsets) can be similarly archived with the current complete master datasets. The media for archival data storage can be either diskette or tape, or both, and the content of those media should always be clearly labelled. Additionally, a tracking log should be maintained by each project manager that identifies the current master dataset contents, extent, and the locations of all archival master copies.

Documentation of datasets is second in importance only to verification and validation of the accuracy of the data in the files. Without informative and complete documentation, the content, quality, extent, known causes of variability, and utility of the data remain unknown. Again, remember that some data has already gone through years of neglect--despite the fact that "copies" were being stored. Storage is *not* the same as managing the data, and much legacy data will be found in great disrepair. The process of cleaning up undocumented data must begin with the very first record, whereas even partially documented data give a data handler or project manager the ability to focus "cleaning" efforts on the sections of the data that actually need it. Documentation involves several types of descriptions, and a formal metadata documentation protocol has not yet been adopted for non-spatial data. However, metadata standards are already mandated for spatial data and similar standards will evolve for "non-spatial" data. A good start toward understanding what constitutes adequate documentation can be found in the sections on Dataset Documentation and Archiving, Disseminating Data, the previous chapters, and in the Appendix.

Data Entry, Verification, and Validation

Data Entry is the initial set of operations where data from paper field forms or field notebooks are transcribed or typed into a computerized form (i.e., a database or spreadsheet). Where data were gathered and/or stored digitally in the field (e.g., on a datalogger), data entry is the stage where those data are transferred (downloaded) to a file in an office computer where they can be further manipulated. Specific procedures for electronic data transfer will not be discussed here, but the general procedures apply for those data, too.

Getting data from field projects into the computer seems like a fairly simple process--just type it in! However, without proper preparation and some simple guidelines, the quality and integrity of the data will be debatable. Three steps are needed in the data entry process to insure that the resulting database is certifiably accurate.

1. Data Verification. Data verification immediately follows data entry and involves checking the accuracy of the computerized records against the original source, usually paper field records. While the goal of data entry is to achieve 100% correct entries, this is rarely accomplished; the *verification* phase checks the accuracy of all entries compared to the original source to identify and correct any errors. Once the computerized data are verified as accurately reflecting the original field data, the paper forms can be archived and most activities with the data can be done via the computer.

2. Data Validation. Although data may be correctly transcribed from the original field forms (data entry and verification), they may not be accurate or logical. For example, finding a stream pH of 25.0 or a temperature of 95°C in our data files is illogical and almost certainly incorrect--whether or not it was properly transcribed from field forms. This process of reviewing computerized data for range and logic errors is the *validation* stage. This can be done during data verification *only* if the operator is intimately knowledgeable about the data, but more often this will be a separate operation carried out by a project specialist *after* verification with the goal of identifying both generic and specific kinds of errors in particular data types. Any corrections or deletions made to a dataset reflecting logic or range errors will also require returning to the original paper field records and making notations about how and why those data were changed. Modifications to the field data should be clear and concise while preserving the original data entries or notes (i.e., no erasing!).

3. Version Control. It is critically important to any scientific project that all data are validated as "truthful" and not misrepresentative given the circumstances and limitations of their collection. The procedures outlined below show the user how to practice careful *version control* while working on files so that changes are incremental, and that roll-back to a previous data handling session is possible until such time as the file being changed is certified as correct and up-to-date (i.e., ready to archive). Failure to follow standardized procedures for data entry, verification, and validation will render a dataset suspect, and such data should not be considered as, or appended to, a master dataset. Again, only the data entry and verification stages can be done by someone who is not familiar with the kinds of errors sought during validation; validation requires in-depth knowledge about the data.

Data Entry Procedures

Data entry is a simple monotone process to perform. However, data entry is *not* a trivial operation, because the value of the data depend on their accuracy. Remember that the single goal of data entry is to *transcribe* the data from paper records into the computer with 100% accuracy. Observation of the data entry guidelines below will minimize verification work. Although transcription errors are nearly unavoidable when entering lots of data, these errors will be corrected during the verification phase.

Have two people available for entering data. Although not required, when one person reads the data off the paper and another enters it into the computer, the work often goes faster with a lower error rate. If another person is not available, the 'enterer' should try to work a bit slower (=not rush) to compensate. Like many monotonous

tasks there is a rhythm to be discovered in data entry, and setting a reasonable tempo is important for avoiding mistakes or getting distracted.

Prepare your work space. *Do not begin data entry in a messy setting.* Clean a space on your desktop near the computer and do not allow other "items" to intrude into that space. You want to proceed cleanly through the process and without distractions that will make you lose your place. You will likely be working with two piles of paper documents, one with data to enter and another for ones you've completed. A pad or notebook and a few fine colored markers (green, blue, and red) are handy for notes.

Examine the documents needing transcription. Become familiar with the data forms, the differences in people's handwriting, and whether or not some "fields" are intentionally left empty on some of the forms. Some errors and/or omissions are detectable or may be suspected at this stage and may necessitate setting aside some of the forms for clarification or correction by the field staff *before* attempting to enter them into the computer. Also identify in the documents what constitutes a good stopping point, since interruptions (or the end of the work day) are likely. The best stopping point is probably when a single, complete field form has been entered; avoid quitting in the middle of a logically single operation.

Prepare the computer. Each dataset may have its own, unique data entry procedure on the computer. The specific application program and/or file to use should be provided by the project manager and be consistent with the data content and data structure. Almost always, data are entered into an empty, "fresh" database table to avoid contaminating existing data [new data are appended to the master data only after formal verification, validation, and documentation].

If the program or application used for entering the records is not familiar, spend a little time practicing first. The best entry applications will also do some validation checking of data on-the-fly, such as checking for valid ranges, dates or spelling, and warn the typist as errors are made and provide the opportunity for correction before the data are committed to a file. You need to know how to commit both a "field" entry and a complete record (e.g., often TAB is used to move between fields, and ENTER is reserved for committing the entire record [whether you finished or not!]). Also, know how to correct mistakes that will be made while typing (i.e., the effects of using backspace, back-tab, del). Once where and how to enter the data are known, you are ready to move on.

Enter the data, one logical "set" at a time. Here you simply enter the data one logical data "sheet" at a time (usually one complete field form). Record in your notes any errors made or any questions that arise about the data content; these will be useful during data verification. Initial each paper form as it is completed to avoid confusion about what has been entered and what has not (green is a positive color to use). *Interrupt data entry only at logical stopping points.* Periodically, make a working backup copy of the data for safety's sake if your software does not do so automatically. When data entry of all records for this dataset has been completed, *immediately complete the next three steps--this is not a good stopping point!*

Print a copy of the computerized data. Print a copy of all the data you entered for the verification stage (see the project manager for formatting details if you are unsure). Do *not* apply any sorting to the file, since it must be checked in the same "order" as it was entered to speed verification from matching field forms. Check to make sure *all* the data were printed (i.e., none are "off" the right margin) and are readable (font size and attributes), since this printout will be used for data verification. Do *not* prepare any reports from these data at this stage since they have not been checked and probably contain some errors.

Initial and date the original records and the printed copy. Indicate on a cover sheet or other suitable location that these data were entered; provide name or initials, and a date. Record identical information at the top of the printout of the entered data. Keep the field data and the printout together for use in data verification.

Make and store a backup copy of the data. To avoid data loss, immediately make a backup copy of the data just entered and store it in a second location (or provide a copy for safekeeping to the project or data manager). Congratulations! The data entry phase is now complete.

Data Verification Procedures

The verification phase is carried out to ensure that all the data were entered and accurately transcribed.

Two people are best here, too. If possible, use two people for the verification phase. This process involves two sets of paper simultaneously and goes much more rapidly than data entry, and thus generates a greater opportunity for confusion or losing one's place when working alone. Verification is best accomplished with one person reading the original data sheets (the "reader") and the second comparing with the same data on the printout (the "checker"). The remaining discussion assumes that a pair of people are working together.

Prepare the work space. As with data entry, a clean work space will promote better control of the verification process. Both the reader and checker should have space for their checked and unchecked pages on the desktop. The checker will need a red and green fine-tipped marker for identifying errors and indicating corrections have been made; the reader will need a marker, too (green is good). Straight edges make reading aligned tabular data much easier. Keep a notepad handy for recording any notes that might be useful during validation.

Compare data and note differences on printout. The reader reads the original data (field forms) and the checker compares that against the printout made after data entry. The three common types of error that will be found are duplicated records (entered twice), missing records (inadvertently skipped during entry) and misspellings (wrong number or code). The checker controls the speed of the reader and halts the reader when any discrepancy is found. When an error in the printout (=computerized records) is found, the correction to be made is noted in red on the printout and *not on*

the original data sheets. After verifying the data from each field sheet the reader should date and initial the original field form at the top (or where provided) stating that verification was done. Continue the reading and checking until all the data sheets for that dataset have been compared. Now you have an original set of data sheets with completion marks (both entry and verification), and a set of printouts with corrections needed marked in red.

Correct identified errors in the computer files. Return to the application used for data entry (or one provided specifically for editing) and correct the errors as indicated on the printout. Make each correction separately (i.e., avoid doing a "search and replace" that might have unexpected consequences). As each correction is made, the red mark on the printout should be "OK'd" with green. Continue until all identified errors are corrected in the computer file, and double-check the printout again for any that were missed (red without green check). Finally, date and initial the printout at the top that all errors were corrected. Save the printout with the original field form, as it serves as direct evidence of the completion of entry and verification.

Perform simple summary analyses. Use the computer to generate some simple summary statistics for the entered data. This is important because even when care is taken up to this point, it is possible to have missed a duplicate or omitted record. For example, do a count of elements that are known to be constant, such as the number of sites sampled, plots per site, or dates per sample. Be creative by asking the same question in different ways; differences in the answer provide clues to where errors reside. The more checks you can devise to test the completeness of the data the more confidence you will have that the data are completely verified.

Make and store a backup of the data. Make a copy of the verified data file(s) and store where instructed. Be sure filenames have space to accommodate version numbers. Pass a second copy of the file(s) to the project and data managers with appropriate documentation. Make a copy of the original field forms. Attach the printout to the original field form and store in the specified area; put the copy of the original form in a second location (i.e. in another building). Check with the project manager for the exact file cabinets to use if unsure about storage locations.

Data Validation Strategies

Unfortunately, there are no step-by-step instructions possible for data validation, because each dataset has unique measurement ranges, as well as sampling precision and accuracy. Nonetheless, validation is a critically important step in the certification of the data. Invalid data commonly consist of slightly misspelled species names or site codes, the wrong date, or out-of-range errors in parameters having well-defined limits (e.g., elevation). But more interesting and often puzzling errors are detected as unreasonable metrics (e.g., stream temperature of 70°C) or impossible associations (e.g., a tree 2 feet in diameter and only 3 feet high). These types of erroneous data are called "logic errors" since using them produces illogical (and incorrect) results. The active discovery of logic errors has direct, positive consequences for data quality and

provides important feedback to the methods and data forms used in the field. Validation, therefore, cannot be ignored.

Wherever possible the data entry software should be programmed to do the initial validation. The simplest validation to perform during data entry is range checking, such as ensuring that a user attempting to enter a pH of 20.0 gets a warning and the opportunity to enter a correct value between 1.0 and 14.0 (or better yet, within a narrow range appropriate to the study area, such as pH of 3.5 to 11.5 at Shenandoah). Not all fields, however, will have appropriate ranges known in advance, so knowledge of what is "reasonable" data and a separate, interactive validation stage is still important. The data entry application should also use pop-up pick lists for any standardized text items where spelling errors can occur. For example, rather than typing in a species name (where a misspelling can generate a "new" species in the database), the name should be selected from a list of valid species, and "picked" for automatic entry into the species field. Again, not all written fields can use a list, but where they can be used they should be. Unfortunately, many current data entry programs provide minimal data validation, if any. Therefore, most quality control work must generally be done *after* data entry and verification.

One of the most important activities of rigorous validation is to return to the original data media (*and* the printout, 2nd copy, etc.) to make corrections and notations about the errors that were found and fixed in the digital files. Without annotating the original field forms, the digital and paper records are out of synch. If this is discovered without adequate documentation, *all* of the data are rendered suspect. This is so important that it really needs to be repeated more strongly:

> *When validation errors are found in the original data, both the computer files* ***and*** *the original field records should be corrected.* Only when the original forms are annotated with the same corrections will the correspondence between computerized files and field forms be kept exact. Failure to correct the original field data forms will create havoc and doubt about the integrity of the data if it is later discovered that the field data and the computerized data do not match. Clearly mark any changes on the field forms so that the original data (i.e., the mistake) and the correction are legible (i.e., do not erase the original data; e.g., circle or mark through the incorrect data with a single line). And, don't forget to make the same correction notations on the other copies of the field forms. Sometimes, the wisest course of action may be to discard the suspect data item from the dataset.

The following generic suggestions can help develop a validation strategy for most datasets, and examples of validation strategies (and strange errors) are also provided.

Catalog the error types found in each dataset. Once particular validation errors are found, it is important to catalog them for that dataset. Notes on the error(s) should include a description, how detected, and how corrected. Both simple, generic errors and more esoteric and cryptic errors need to be documented. This list of errors will be a valuable reference for the next validation session and ultimately for building formal

validation procedures into the data entry process and other automated, post-entry error-checking routines.

Perform exploratory data analysis to look for outliers. Database, graphic, and statistical tools can be used for ad-hoc queries and displays of the data. Look at histograms, line plots, and basic statistics for possible logic and range errors. Such exploratory techniques will identify obvious outliers. Some of these may appear unusual but prove to be quite valid after confirmation. Noting correct but unusual values in documentation of the dataset will save other users from repeating the same confirmation themselves.

Modify field data forms to avoid common mistakes. With a catalog of validation errors and exploratory data results in hand, reevaluate the field data forms as the source of the logic errors. Often minor changes, small annotations, adding check boxes, etc. to a field form will remove ambiguity about what to enter onto the form. In fact, any time the same type of validation errors occur repeatedly in different datasets, the field form is usually at fault--not the field crew. Repeated validation errors can also mean that protocol(s) or field training is faulty, both of which are also important to recognize and correct.

Example validation problems. Below are four examples of logic errors discovered in Shenandoah datasets. These examples are informative. They demonstrate how errors can hide as well as some generic and specific approaches to finding them. For data validation, the most useful adage is: "Seek and ye' shall find." Looking for errors in a data file, however, is probably not a career option and other work awaits. At some point, the validator must stop searching for problems and accept the data as certifiably verified and validated. Subsequent data analysis will ultimately (hopefully?) reveal any remaining errors. Keep in mind that the most effective mechanism for avoiding tedious validation is to get the right data into the computer in the first place, i.e., having a comprehensive set of SOP's and data-collecting protocols for quality control: clear field methodologies, a well-trained field staff, well-organized field forms, and data entry applications with simple validation built-in. Last but not least, note that exploring the data looking for logic errors is also a good way to "get to know" the data intimately; actually finding real errors builds understanding of what is truly represented.

> Wrong date. A simple typo during data entry creates a logical "set" of data for a day, month, or year in which samples were never taken. This can become puzzling if the data are sorted by date--thus moving the entry away from its true neighbors. If sorting creates the appearance of "missing data" where a record should have been, the apparently appropriate corrective action might actually *create* duplicate records in the file rather than fix the ones that were wrong--leaving the original problem unresolved. Even when left in the original order, however, date errors might go undetected because checkers can sometimes "see" what the readers say--especially when the month and day are the items of focus and an incorrect year digit is not examined. A summary analysis counting the total records for the dataset will also be correct. A check of the number of dates or samples per year will often detect an erroneous year by revealing too

many samples or a year that does not belong, whereas the rest of the data records reveal where the correction is needed. Identifying site code errors, etc. is a similar process for incorrect values not identified during verification.

Cryptic duplicates. Shenandoah contracts out invertebrate identifications for the stream monitoring program. The contractor supplies printed tables containing species codes, names, and counts in each sample, which are entered into the computer. Cryptic duplicates occur in the data files when a single sample contains two entries for the same species (the contractor didn't realize they already had a line for that species when doing the counts and added another line later in the table). The data verification process correctly confirms the separate entries but does not recognize that they should be pooled for that sample. Summary counts of the number of "species" for that sample also show the same number as lines of original data--apparently correct. However, a count of the number of unique species (i.e., a "count distinct" query) for the sample shows one less than the line count. Returning to the original data form and comparing each line with the others for that sample eventually reveals the error of duplication, and the data file is corrected by pooling the abundance values into the first record of that species and then deleting the second. The original printed data table (our original "form") is then also corrected. Here, two different methods of making counts of the same item (species per sample) were used and compared to find the discrepancy.

Wild temperatures. Stream temperatures can show wild variation and yet be completely verifiable and valid. For example, some older data, or the occasional spurious recent record, may have been taken in Fahrenheit rather than Celsius. There is obviously a big difference in the recorded number(s). This is really a protocol problem and not a data question, but if quality control procedures during data collection were lax, these types of errors are often found only during data validation or (more annoyingly) analysis. Routinely producing a box-plot or histogram of numerical data will reveal dramatic outliers, and when the original data forms are consulted, true outliers vs. errors in measurement scale or units become apparent, as does the correction for the files (i.e., convert the measurement to the appropriate units).

Trees that shrink. Shenandoah's vegetation monitoring program includes remeasuring trees at permanent plots every five years. In one survey, the project manager discovered that some of these remeasured trees were getting smaller--recent DBHs (diameters at breast height) were less than the original measurements five years before. Tree trunks of live trees don't get smaller. Some serious detective work revealed that the data were entered accurately (verifiable), but that there appeared to be slight-to-moderate differences in the accuracy and exact methodology used by current vs. previous crews. A "Search and Compare" program was written to parse the data and identify and scale the differences between trees, revealing the extent of the "damage" in the data. Unfortunately, this problem could not be fixed by editing the data files. Rather,

it revealed a previous protocol problem that resulted in data of poor quality which are useless for their original purpose.

Guidelines for Data File Editing

Datasets are rarely static; they often change via additions, corrections, and improvements from summary and analysis. These guidelines outline basic strategies for editing data files to update records, add new records, or change the record structure. There are three main caveats to this process: 1) only make changes that "improve" or update the data while maintaining data integrity, 2) document everything done to the dataset, and 3) be prepared to recover from mistakes made during editing.

All data must be validated as "truthful" and representative given the SOPs of their collection. Proper preparation for and documentation of all changes made during editing is important. Practice careful version control during editing to ensure that changes are incremental, and that roll-back to a previous editing session is possible until such time as the file being changed is certified as correct, up to date, and ready for archiving.

Before Editing

Make notes of editing actions. The editor should carefully note any and all changes done to a data file. These working notes may or may not become part of the permanent record for the data, but are necessary for reconstructing the strategy used to change a file during an editing session. Whether or not these working notes are saved, a formal written summary and explanation should be created from each editing session, including a listing of all changes made, when they were made and by whom; this editing report will become part of the documentation permanently associated with the data file. A scratch copy of the Data Edit Report included in the Appendix can make note-taking much easier. Also, detailed notes may later prove invaluable as a guide for accomplishing specific editing tasks in subsequent sessions.

Prepare a clean work space on your computer. Work on files in a "safe" place on your computer--away from other files that might accidentally get altered or deleted. A separate directory such as /EDIT, /WORK, etc. should be created and used for this purpose. A /VERSION subdirectory is useful to store numbered working versions of the data file if the current directory becomes crowded. In that case, the main working directory might only contain the current copies of the files undergoing change, while the roll-back versions are safely tucked away.

Work only on copies of files, one-at-a-time. Never work on the only copy (i.e., the original) of a file. Make a working copy of the file, or better yet, give it a slightly different name. Choosing a shortened name for your working copy can facilitate loading, saving, and version control procedures. For example, if working with the file STG3GC.DBF you may want to shorten the name during the editing session to S*n*.DBF, where the '*n*' is the version number. If two or more files are edited simultaneously

(i.e., if they are relational), use similarly coded version numbers to remind yourself that they are both at the same stage of editing.

Work on a subset of the data whenever possible. Here you want to avoid corrupting any data that do not need editing. For example, if a field named SITECODE needs to be adjusted for only one year in a multi-year file, it is best to isolate records for that one year before any editing begins. This can be done by splitting the original file into two parts, editing the one that needs it, then recombining them later. (Remember to track the whole file and separate pieces in your notes.) Another approach is to use a "filter" tool in a database program to allow access only to the records you want. Some programs will "write protect" data fields to prevent accidental alteration. In any case, take steps to prevent accidental changes to data in non-target parts of a file.

Define Editing Strategy

Working file information. Write down the names of the file(s) to be worked on, the initial date(s), size(s) and the number(s) of records. If you will be renaming a copy of the file to work on, record the working name for the file that includes a temporary version number (e.g., "S0.DBF is the original SITE.DBF file."). If files are to be edited in more than one format (e.g., as both ASCII and .DBF) to take advantage of different editing tools, state so at the outset and include information about the file extensions that will identify them (i.e., "ASCII versions of the same-name .DBF files will have a .TXT extension.").

List and sequence the changes to be made. Before beginning a data editing session, write down precisely, in as much detail as practical, what is about to be accomplished. If several steps are needed to "fix" a file, they should be written down separately and examined carefully *before any editing begins* to evaluate whether any one change might adversely affect later steps. This examination may also reveal how to arrange a cascade of changes to be most efficient. If order of actions is important, explain why in your notes before beginning.

A common example of poor planning is using global search & replace functions indiscriminately, such as wanting to increment a few numbers by one. You might start by changing all 1s to 2s, and quickly discover you can't distinguish the original 2s anymore and all the 1s in compound numbers and other codes were also changed to 2s--not what was intended. For this example, you would record beforehand the proper strategy of replacing the highest number first, then working downward, and restricting your edit to only the field(s) of interest.

Define the tools you will use for editing. List the computer program(s) and/or file editor(s) that will be used to make changes to a file. This needs to be stated only once in your notes for the editing session, but if you perform some unusual feat of magic to accomplish a difficult task, or simply try some advanced feature out for the first time, you may want to be recording the steps you take in detail. ("Wow! Perfect! Now, how did I do that?!")

During Editing

Keep making notes. If you have properly planned and documented your strategy for the editing session, there is little left but to carry it out. Remember to record each step in the edit process, including names and versions of files--especially noting any changes in the number of records as a result of an edit. It is best to number the editing steps in your notes, and edit version tracking can be facilitated by using the step number as the temporary version number as each step is completed (e.g., S2.DBF is the *result* of step 2 in your notes of the editing process).

Save your work often. Save your files often; use the timed-autosave features of the software if available. The interval chosen for saving ongoing work is flexible and not predicated by completing and renaming a new version. For example, if there are a small series of relatively simple updates to carry out, the editor may wish to complete them all before saving, since it would be an easy task to do again; at that point, the editing may be completed and the file renamed as a new version. However, if a step in the editing process requires numerous changes to individual records, you may want to save at five minute intervals or even after each record is completed. In the latter case, keep running notes of last record edited before each last save.

Track your versions of the edited file. Be sure to keep all intermediate, numbered versions until the full edit is complete. Use compression software for file storage if disk space is limited.

After Editing

Backup your new file and the version series. Immediately make a copy of the edited file and its intermediate versions to diskette, even if only for temporary storage. This prevents a power outage or hard disk crash from nullifying editing efforts.

Review your pre-edit notes. Check off the changes you wanted to make against the notes made during editing to double-check that all changes were really made.

Formally document the file edit. Document a formal statement of what was done to the file. Include your name, the date, the file(s) that were changed and a concise list of the changes that were made and why, and the version series used during the edit. Each record that was changed does not have to be listed if the change applied was more or less global in nature, but if individual records were separately adjusted for a particular reason (i.e., to correct an error) they should be identified individually. The documents that accompany the file--including edit summaries--should detail the entire history of the file, so don't leave anything out, no matter how minor you think the change was (for example, changing a single date in a 20,000 record file still needs an explanation). Any change to a long-term database is not trivial. A sample data edit report is shown in the example box below, and a generic Data Edit Report Form is included in the Appendix.

Example Data Edit Report

Name: Steve Tessler
Date: October 7, 1993 (completion date)

File(s) Edited: AQINS.DBF from Aquatic Macroinvertebrate LTEMS project

Reason for Edit: Taxonomic code changes required to the TAXA field as per correspondence with Steve Hiner/Reese Voshell @ VPI. Changes needed and taxa-change categories are fully outlined in the correspondence and "Change Sheets" created for this edit, and include changing certain generic ID's to species, eliminating "terrestrials" and "impossible" taxa (for Virginia) in the data, and regrouping *Baetis* and *Pseudocloeon* as *Baetis* complex.

Program(s) Used: FoxPro 2.5 and DBBrowse for .DBF files; TSE pre-release 1.0 and QEdit 2.15 (both Semware) for ASCII files.

Original File Information: AQINS.DBF, Version *n*, 13,779 records, contains data from 1986 through 1992. Full error-checking pending.

Final File Information: AQINS.DBF is now temporary version 7 of *this* edit, dated 10/07/93, time 10:15:23.03a, with 13,731 records.

Editing Details:
1. Created AQINS.1 as a tab and "" delimited ASCII version of the original .DBF. Files with a number extension are ASCII.
2. Did global delete of terrestrials and "impossibles." 33 records removed; new number of records = 13,746. Saved as AQINS.2.
3. Found error while previewing for *Peltoperla* changes. 15 lines were duplicates in 3L301 2nd Qtr 1988; confirmed by checking paper records; they were deleted. Saved as AQINS.3, #Rec now = 13,731.
4. Made unusual, one-time-only changes as per Change Sheets. 19 records were changed; # rec still 13,731. Saved as AQINS.4.
5. Changed all UNID to straight taxon code (i.e., removed X's from the code). 138 X's removed, no rec# change. Saved as AQINS.5.
6. Changed genus to species for monotypic genera and *Perlesta* to *P. placida* group as per Change Sheets. 395 changes made to 15 taxa. Saved as AQINS.6, still with 13,731 records.
7. Changed names up one taxon level globally as per Change Sheets. 358 changes made, still 13,731 records. Saved as AQINS.7; loaded into AQINS7.DBF for checking and sorting. Re-saved as AQINS.DBF, Version *n+1*.

Archive the edited file and associated documentation. When the editing and documentation are complete, follow the guidelines for formal archiving of the file and its associated documentation. Dataset Documentation and Archiving steps are described below.

File the edit session report. A paper copy of the formal edit session documentation should go into a folder associated with that file or project.

Print the file, if required. Optionally, the final version of the edited file should be printed if that has been the standard procedure for the project.

Archive the file(s) and documentation. A copy of the final version of the file, copies of the edit information, and copies of all other documentation associated with the file are also archived to safe, off-site storage.

Update all computers that need the new file version. Copies of the new file(s) need to be distributed to where they are used. This process is best handled by the individual responsible for the project or the data. The project manager should always have a current listing of the latest version, date, size, and number of records in the dataset.

Update any Master record(s) listing current data files. Any Master records that note the current version of datasets should be updated. This might include a notebook to which all users have access to check on the "current status" of their data, and/or a computerized database (i.e., the Dataset Catalog) used for the same purpose. Any place where the file version/date is recorded must be updated with the new information immediately.

Dataset Documentation and Archiving

Numerous references to dataset documentation and archiving occur throughout these guidelines. This section concisely reviews and discusses the overall strategies and concerns for documenting and archiving datasets. While some of the documentation procedures and metadata standards are currently evolving, good preparation and comprehensive documentation of datasets will ease the transition to a more rigorous system. Draft FGDC "Non-spatial" Metadata Standards are included in the Appendix.

Example Dataset Documentation

- Resource Study/Data Collection Plan
  - Project Title
  - Problem Statement
  - Description of Study/Action
  - Description of Data (types, ranges, etc.)
  - Inception/Duration Dates
  - Stipulated SOPs (accuracy, precision, etc.)
  - Project Originator/Manager
  - RMP record
  - Long-Term Data Manager
  - Project Data Management Plan
- Implementation
  - Modifications to SOPs
  - Data Entry, Verification, and Validation Considerations/Modifications
  - Data Edit Report(s)
  - Legacy Data Validation Report
  - Dataset Version(s)
  - IAR Record
- Archiving
  - Compilation of all Analog and Digital Documentation
  - Location of Master Data and backups
  - Dataset Catalog Entry
  - Data Dissemination Contact(s)
  - Data Dissemination Plan and Records
  - Data Distribution Arrangements
  - Access Restrictions
  - Metadata Document

Documentation

Documentation for a dataset should begin at the conceptual stage of the study that will collect the data. Notes concerning the purpose of the study, need for the data, monitoring goals, etc. are important metadata considerations. Good sampling design and SOPs should be documented along with comprehensive plans for error checking, validation, archiving, application, and dissemination. NPS specific documentation should include RMP, IAR, and dataset catalog records. Any

modifications to the data content or procedures should be duly recorded. Whereas most of the specific documents and procedures are discussed in other sections, a brief outline of dataset documentation needs is presented in the example box at right.

Archiving

As discussed in the introduction of this chapter, formal archiving of master datasets and associated documentation is done to protect and maintain the physical and informational integrity of the data through time. Because archiving procedures include the physical storage of master datasets at separate locations or buildings, each site will need to evaluate and implement their own archiving procedures and include them in the Park Data Management Plan.

Validating Legacy Datasets

Observe the Data

The first step in validating legacy data is to determine the physical and logical nature of the extant dataset. Assemble and review both the data and all documentation relating to the dataset. Determine the original purpose for the dataset and if validation will enhance the data enough to be useful for present needs. Become familiar with the type(s), collection protocol(s), accuracy, precision, as well as the reasonable and valid range(s) for the data. As discussed previously, credible data validation requires intimate knowledge of data characteristics, and this is especially true for validating legacy datasets.

Prepare Validation Strategy

Once the dataset has been thoroughly reviewed, prepare a validation strategy. Initial objectives should include a list of valid data ranges and establish a procedure for handling discrepancies in the data. Since the data may be in either analog or digital format, determine the steps necessary to enter or convert the data into a standard database format (e.g., DBASE III+). If the data are analog, many if not all of the procedures and considerations discussed in this chapter for ongoing data collection, verification, validation, and editing will apply. If the data are already in digital format, systematic validation techniques as discussed in the Data Validation Strategies section above may apply. Regardless of the technique(s) used, a formal, written data validation strategy listing all aspects of bounds checking, editing criteria, etc. should be prepared and included in the permanent documentation for the upgraded dataset. Example legacy dataset validation strategies used in preparing the Baseline Water Quality Inventory and Analysis Reports from the EPA's legacy water quality data are included in the Appendix. The validation report for the dataset will enhance the preparation of a standard metadata document in the future.

Digitize and Validate the Data

If the dataset is analog, follow the guidelines in the section on Data Entry, Verification, and Validation. With analog data, previous preparation of DBMS software data entry forms with built-in error and bounds checking is highly recommended. Comprehensive error/bounds checking at the data entry stage will greatly enhance the efficiency of the overall data validation effort.

After the dataset is in digital format, convert the data to the format used by the software platform(s) to be used in the validation stage. If possible, print out a hard copy of the dataset for notation of suspect data entries that may need to be edited.

Validation. Using the *previously prepared* validation strategies, systematically process the data to check for logical data ranges and data collection locations. An important step for legacy data validation may include displaying the data locations accurately on a relevant map using GIS software to examine the locational position of the data collection stations (e.g., check that water gaging stations are physically located on hydrologic features). During the systematic examination of the dataset, note on the print out any suspect data that may need to be modified or omitted from the validated dataset.

Editing. Editing of the digital legacy data should be a similar process to that discussed in the Guidelines for Data Editing above. Carefully outline, follow, and document an editing strategy consistent with the *previously prepared* procedures for handling data discrepancies discussed in the Prepare Validation Strategy section above.

Documentation

The legacy dataset validation report should include: 1) an abstract or executive summary of the dataset history and validation effort, 2) the written validation strategy in enough detail that it could be repeated if needed, 3) formal data edit report(s), and 4) the archived version(s) of the dataset. Other documentation should include relevant notes concerning the dataset, entry into the dataset catalog, and a metadata report that encompasses any existing standards. Any restrictions, problems, or other considerations relating to the use and accuracy of the data should be made readily apparent in the dataset documentation.

Archiving

Once comprehensively updated and validated the legacy dataset should be archived, installed, used, and disseminated in accordance with procedures established in the site data management plan (discussed in the Dataset Documentation and Archiving section). Optimal archiving of legacy data includes a copy of the original "raw" dataset that has not been validated, edited, and/or filtered to correct apparent data errors.

Guidelines for Disseminating Data

These guidelines are currently evolving. When the need arises to distribute data, some of the specific procedures described below may need to be modified (such as providing datasets in a standard delimited ASCII format rather than customizing datasets for individuals). As data management planning and implementation become standard procedure, some of the operations described here will already be complete for each dataset. When datasets are fully validated, archived, and cataloged, they should already exist in a form easily transferred without additional preparation. The following steps describe the transfer of data with supporting documentation that fully describes the data extent and limitations. An example data file description is given at the end of this section. Although these instructions are devoted to disseminating a customized dataset in response to an individual request, the process may be adapted for assembling a standard dataset format for general or automated distribution.

Items That Precede Data Preparation

Identify the data recipient. Identify, contact, and work out a transfer protocol with the data manager who will actually receive, handle, and process the data. The next four items should be discussed with the recipient.

Identify the data to transfer. Determine whether the entire dataset or only a subset of the data is desired. If needed, provide the recipient with the present dataset structure and description before discussing the transfer. A report from the Dataset Catalog will provide much of the needed information.

Determine the format for the data. Determine the data format to send the data recipient (ASCII, .DBF, spreadsheet, etc.). If ASCII files are desirable, then determine if the data should be in a fixed-column format or delimited by a specific character. Since ASCII conversions of .DBF files from DBMSs often use quotes on character fields but not on other types, determine if this is acceptable or if all fields should either be quoted or unquoted. Special formatting efforts must be determined and approved on a case-by-case basis.

Determine the format for accompanying documentation. Determine the format of accompanying documentation (e.g., ASCII text w/ 65 column maximum width, WordPerfect 5.1, etc.).

Determine the transfer medium and method. Determine how the transfer will take place (e.g., 3.5" disks via U.S. mail, FTP to a specific Internet address, etc.), and whether file compression is acceptable. Since most data files compress very well, transfer of compressed archives should be encouraged. NPS maintains an active Internet FTP site at *ftp.nps.gov* where files may be placed (i.e., *put*) in the */incoming* directory for retrieval via FTP.

Data Preparation

List the data items to be prepared.

Prepare the data. If necessary, convert the data to be transferred to the transfer format(s). Fully document on paper any modifications, additions, or changes in structure that are carried out on the files during this process. For example, the recipient may only want selected fields from a database. During preparation of this data, record the actions taken to achieve this goal.

Data Documentation

List of files. List all file names, dates, sizes, and any directories and subdirectories that are to be transferred. Write brief, informative descriptions of each item.

File relationships. Describe any relationships between individual data files. For example, if the data are relational and fully normalized, identify the primary and foreign keys in individual files used for linkages. Include a text diagram showing these relationships whenever possible.

For each data file, prepare a table of the data file structure that includes:

the total number of records,
the size of each record,
the number of fields per record,
the names of fields, in record order,
field type, size, etc.,
a description of the field, and
codes for missing values.

Describe the full dataset, including limitations. Document the dataset with a descriptive paragraph and a disclaimer if necessary; separate comments may need to be made regarding each individual file. Be sure to explain any known problems with the dataset--such as different protocols followed over different years, changes in equipment, detection limits or resolution, etc. An example disclaimer notice is included in the Data Management Plan Guidelines (Chapter 2). An example Data Format Document is shown below.

Assemble The Transfer Materials

Review. Review all of the items above and check that any descriptive documents are also catalogued (as noted above).

Double-check that complete documentation is provided. Both printouts and/or digital copies of all descriptive documentation material should accompany the dataset.

Example Dataset Format Documentation
(modified example from Shenandoah National Park)

Park Contact Information:
Name, Position, Address, Phone, etc.

Dataset Title: LTEMS Aquatic Macroinvertebrates Database
(see accompanying Dataset Catalog Report)

AQINS.HDR -- description of AQINS.TXT (ASCII form of AQINS.DBF)
AQINS.DBF and AQINS.TXT contain 13,731 records from 8/7/86 to 9/28/92

This is a brief header definition and description of the AQINS.TXT file containing the fixed-column ASCII version of AQINS.DBF. There are 8 fields (variables) in the file we are sending for each record (row or line). There are a total of 13,731 records in the file. Each record is based on the actual presence of a single taxon in a specific sample. The following notes detail specifics of the AQINS.TXT file and describe how it differs from the original dBASE file.

Description of Fields in the original AQINS.DBF dBASE file (see LTEMS manual).

NAME	TYPE	LEN	DESCRIPTION
1. SITE	Character	5	LTEMS Site code
2. QUARTER	Character	1	Annual quarter (1-4)
3. SMPDATE	Date	8	Sample Date
4. ASMPL	Character	1	Sample # for a method
5. SMPMETH	Character	3	Sample Method
6. TAXA	Character	7	Taxon code
7. INSCNT	Character	4	Taxon count in the sample
8. METERCNT	Character	5	Calculated #/sq.meter
9. WEIGHT	Character	6	Individual weight
10. STAGE	Character	1	Individual stage
11. FUNCGRP	Character	2	Taxon functional group
12. REARED	Logical (T/F)	1	Individual reared or not

First, the fields METERCNT, WEIGHT, STAGE, and REARED were dropped from the ASCII version of the file. These were either virtually always empty or redundant with other data. They are still available in the .DBF file, or from me by request. Second, all empty or null fields have been filled with an "X" to represent missing data. Any analytical processing of the data should properly identify the "X" as missing data.

For quantitative methods (SMPMETH = PIB or SUR) the actual count of individuals of that taxon in the sample is given in both the TXT and .DBF files; so in these records a "1" represents only one individual in the sample. Most of the "counts" for the other sampling method types are null (SMPMETH = QUD, AER or RED); the mere existence of the record in the file indicates the presence of a specific taxon in that sample.

Other fields in the dataset were empty and are now filled with "X" -- reflecting the absence of corresponding data for that record. FUNCGRP is not available for some taxa, so its analytical value at all is highly questionable. A single "X" has been placed in each such field that is truly null.

The AQINS.TXT file is sorted (ascending) by the following fields, in this order: SMPDATE, SITE, ASMPL (which follows SMPMETH), and TAXA. This sort gives a "taxa-by-sample#-by -site-by-date" which was useful for checking items in the file. The data occupy the first 45 characters in the text file, including double spaces between actual data columns. A short description of data field identity, column location, maximum length, and some notes on the data are given below. The range of "Columns" occupied for each field should be consulted for importing the data using a fixed column format into other programs.

Description of fields format in the AQINS.TXT file ..... S.Tessler 7/93 SNP

# Name	Columns	MaxLngth	Notes
1. SITE	1-5	5	alpha-numeric LTEM site code
2. QUARTER	8	1	range is 1 to 4
3. SMPDATE	11-18	8	date format is numeric YYYYMMDD
4. ASMPL	21	1	range is 1 to 6 per SMPMETH
5. SMPMETH	24-26	3	PIB, SUR, QUD, AER, RED
6. TAXA	29-35	7	codes refer to TAXADICT.DBF
7. INSCNT	38-41	4	for PIB & SUR "1" is real count
8. FUNCGRP	44-45	2	see LTEMS manual; some null (X)

Compress the dataset, if required. If compression was stipulated for data transfer, compress the file collection with any subdirectories intact.

Copy the files to the transfer medium. Prepare the distribution medium and copy the files to be transferred. Recheck the list above to be sure all items are included.

Prepare a cover sheet. Write a brief, dated cover sheet addressed to the data recipient, with a return address, phone and fax numbers. If the data are compressed on the transfer medium, write a brief description of how to decompress the files and how much space they will take up when decompressed.

Assemble the printed documentation. Assemble the printed copies of all documentation that will accompany the data with the cover sheet on top.

Archive And Document What Is Being Sent

For your records, make a compressed version of the collection of data file(s) and accompanying material. A copy of this compressed collection should be backed up to a storage disk as a record of what was transferred.

Send the data!

Routine Backup Strategy for Work Computers

These guidelines are intended to help develop a backup strategy for recovering work files after simple computer accidents or full-blown disasters--events such as "losing" a critical file or experiencing unexplained file corruption, a hard drive crash, or total system loss. Few things in life are as terrifying as watching a month's worth of creative effort disappear before your eyes, especially, knowing that a little extra work the day before could have brought it back in minutes. Hopefully your system will never fail you, but it cannot be overstated that a little time and effort spent in anticipation is good insurance. Consider the alternatives: disruption of work, missed deadlines, blown commitments, data loss, or worse. Keep in mind that these pages are for planning and testing a backup strategy for *disaster recovery*--not for cataloging or archiving work files for storage and retrieval; another scheme is needed for that.

To understand what is involved in planning a backup strategy, imagine that a work computer is destroyed in a fire. A good backup strategy will allow the complete reconstruction of the system, directory structure, application files, and work files on another computer in a short time if another computer is immediately available. It would also be possible to restore just the files needed right away to another computer, so work could continue until a complete restoration was possible. The key to remember is that the restoration will only be as current as the last backup; so if backups occur only once a week, a weeks' work may be lost! And since the hypothetical fire above destroyed the computer, it is also likely that it destroyed any disks or data tapes stored along with it. Indeed, part of the backup strategy is determining a safe place to store the backup sets.

So, what exactly *is* needed for a viable backup strategy for a work computer? That depends upon the way people work. For example, a person who edits and creates lots of files and/or very large files every day would need a more rigorous backup plan than someone who only needs to backup an occasional letter. Another factor is the nature of the files to be backed up. Short memos or email may not need rigorous backups--especially if printed copies are on file. However, a long document or even a single page of a complex form may represent significant effort, so backups are essential. Data files in particular are much less useful in printed form than on a computer where they can be accessed, so they should *always* be backed up after working on them. Fortunately, many tools are available to assist in developing and carrying out a personalized backup plan.

Backup Tools

One purpose of standardizing computer resources is to provide each system with similar basic tools and utilities to perform a variety of tasks. Backup plans need to be "flexible" to match differences in work schedules and utilize different tools for different jobs. The backup tool set allows the time and effort spent backing up to be varied by individual, but the best protection is to do backups every day.

Below is a brief introduction to some of the basic backup software installed on most systems. Six "tools" are described below: three of them come as part of MS-DOS 6.2, two others may be purchased for use, and the last is an example of a custom program created at Shenandoah and adapted to provide automated backup options.

UNDELETE. UNDELETE is a MS-DOS utility that recovers a file that was inadvertently erased. Sometimes it can find and recover files that disappeared during an unplanned system interruption (such as after a power bump or system lockup that required rebooting). While not designed for backups, a discussion of disaster recovery tools would be incomplete without this one. More information on this file utility is available via the DOS online Help system. Just type "HELP UNDELETE" at a DOS prompt. However, the real secret of successfully using this utility to recover a file is to "*stop what you are doing and immediately use the UNDELETE command to retrieve the [lost] file*" (from the DOS Help system). If you keep working and write over the file's location, it is gone forever.

COPY & XCOPY. These are also MS-DOS programs. Everyone has probably used the COPY command to copy a file from our hard drive to a diskette, and that is the simplest type of backup that can be made. If only a few files are worked with each day, that may also be all that is needed for backups on a day-to-day basis. The XCOPY utility is a more powerful copying tool that can also work with entire directories, not just individual files. If you have a complex directory structure for your working files you may want to explore XCOPY in more detail. More information about both copying tools is available using the DOS online Help system.

MSBackup. MSBackup is a backup program that comes with MS-DOS. This tool may be useful as part of an overall backup strategy. MSBackup's has built-in functionality

for both backups and restores, and the ability to do full, differential, integral, or selective backups. For example, a setup called "BACKUP95.SET" could automatically backup only 1995 files. MSBackup has several optional settings and a comprehensive online help system that includes a glossary of backup terms. The program can be started by typing "MSBACKUP" at the DOS prompt and also has a Windows version.

Although MSBackup is a DOS tool and available on to most systems, the primary backup tool recommended and discussed in this document is the file compression utility PKZip. PKZip has more file handling options than MSBackup and supports command line parameters that are useful in batch programming.

PKZIP & PKUNZIP. The "PK" compression utility programs are site-licensed for use on NPS computers and are very useful backup utilities. Shareware versions exist that may be distributed with compressed datasets. PKZIP is used to compress large files or entire directory structures to a single, much smaller file with a .ZIP extension. PKUNZIP restores the original files from a .ZIP file to any specified location. Entering PKZIP or PKUNZIP alone at the DOS prompt will start help screens for the programs.

PKZip can create an editable list of files for a specific date or date range before compression. The PK utilities also have the ability to span several disks with large compressed files. The .ZIP utilities could backup an entire hard drive, but tape backups are much better for that. Directory viewing options for compressed .ZIP files are explained in the PKZIP help screens. A Windows version is also available for working with .ZIP files without leaving Microsoft Windows.

The .ZIP utilities do several things of importance here: 1) quick, efficient backups (and restores) over multiple diskettes of file(s) too large to copy to a single diskette, 2) selective searches for files to compress from a drive, 3) preserve the full path name along with the backup files, and 4) be included in DOS batch file processing to help automate backup routines. Examples of PKZip commands and the "BYE" backup applications are included in the Appendix.

Portable Tape Backup System. Although a variety of high-capacity backup devices and options are available for installation on individual systems (e.g., magnetic/optical drives, recordable CD-ROMs, removable hard disks, and tape drives), a less expensive alternative includes portable high-volume storage devices that may be shared between systems as needed. Portable storage devices that attach to a computer's parallel port include the Iomega Zip drive (100 Mb magnetic disks) and magnetic tape drives. The Zip drive provides random file access and rapid file transfer capabilities, but the discussion here will focus on portable tape drives because of their higher capacity. Two popular portable tape backup systems are the Backpack (Micro Solutions) and Trakker (Colorado Memory Systems) tape drives. The Trakker drive, which is used at Shenandoah, is described below.

The Trakker portable tape backup system is an easy device to use and can be shared on a check-out basis. The Trakker unit attaches to a system's parallel port and can make total, differential, integral or selective backups of the *entire* system onto a single,

small tape cartridge. Shenandoah's computers bear yellow stickers that list the date(s) of the last tape backup of the hard drive(s), and that date may be set in MSBackup or used with PKZip to allow between-tape backups to diskettes. Tape backups, whether using a portable, system, or network drive, should be a part of a long-term backup strategy--perhaps on a monthly, quarterly, or semi-annual schedule.

BYE, the end-of-day routine. The BYE routine discussed here is a DOS batch file adaptation of a backup system at Shenandoah National Park. The original BYE routine was written for and by the Shenandoah I&M Unit to solve two problems: a) not always knowing what new or modified files needed to be backed up each day; and b) having to remember all of the available backup options and commands. Semi-automatic daily and/or weekly backups can be made via typing "BYE" at a DOS prompt. Example BYE backup applications are included in the Appendix.

In general, the BYE applications use PKZip to assemble an editable list of new and modified files which are displayed using the MS-DOS EDIT program. The file list shows the full path of either just the daily files (files created or changed on the current date according to the system clock) or files that are new or modified since a user specified date. Files can be removed from the backup list by scrolling to the file to be deleted and typing Ctrl-Y or by using the mouse. Help for using the EDIT program is available by typing F1. Once the backup list has been edited, select Exit from the File menu. If changes to the list were made, the user is prompted to Save them. Next, the BYE application searches for a .ZIP file with the same file name and presents backup options if one is encountered. Daily and/or weekly backup .ZIP files may then be copied to a diskette, tape, or other media (or re-ZIPped to multiple diskettes if necessary) for safe storage.

All of the BYE batch application functions discussed above are available through current versions of MS DOS and PKZip. The functions may be applied manually if desired, and the included batch files may be edited and adapted to customize individual backup strategies.

A Simple, Generic Backup Strategy

Consider the ease of file *recovery* while deciding on a backup strategy: generally, the more complex the backup, the more difficult the restoration. This section outlines a simple, generic backup strategy that will suit most users. An efficient backup strategy will minimize the amount of time spent on backups each day, but remember, this plan is only a prototype example for planning individual backup strategies.

This generic backup strategy involves three levels of backup (daily, weekly, and long-term) using two tools (tape and PKZip). The procedures are listed in the general order of application.

Total system backup to tape. The first step is to perform a full backup of the system to tape (or other high capacity media). Using the portable tape drive's built-in compression option, this should occupy only a single tape for most systems. Label and

store the tape in a safe, off-site location (i.e., in another building). When another tape backup is needed (determined below), use a fresh tape and store it with the first. On the *third* tape backup, *write over the data on the first tape*. By rotating the overwriting of two tapes, copies of the last two complete system backups are available at all times. After each tape backup, write the backup date on a sticker on the computer as a reminder, then change to the "\BACKUP" directory and delete any daily or weekly .ZIP backup files that may reside there (these are now on the tape).

Daily backup using .ZIP files. After a full tape backup, the primary concern is backing up new and modified work files on a regular basis. The user can run BYE and/or manually locate and backup new and modified files. Using BYE or PKZip will save a backup .ZIP file with compressed copies of all the new and modified files for the chosen period. The name of the backup .ZIP file should be either the single day's date (e.g., BK060196.ZIP) or the start date of the backup range, so that it is easy to locate a specific backup's date(s) from the .ZIP file attributes. With this naming convention, the date range for the backup begins with the file name and ends at the file date. After creating the backup on the system's hard drive, copy the .ZIP file(s) to diskette(s) for off site storage. If needed, recompress large backup files to span multiple diskettes. Note that one concern when backing up work files from a file date search is that any software or data files added to the system with dates *earlier* than the current or specified search date will not be included for backup--these files will need to be added to the backup list or the .ZIP file manually.

Users on a network may be able to save backup .ZIP files directly to their home directory on the file server, rather than to diskettes. Be cautioned, however, that any files saved on a network server are at risk if stored in public space or if the server is located in the same building as the user's system.

Weekly backup using PKZip. Periodically, depending upon the computer use, use PKZip or MSBackup to make a differential backup to diskettes. This will include all the files that are new or have changed since the last full tape backup (or the last weekly backup if saved separately). The BYE applications in the Appendix provide several automated functions to aid in daily and weekly backups.

A differential backup is also a good idea each time you install new software or add numerous files to your system. If the changes to the system are particularly large or weekly backup disk sets become too cumbersome, consider a tape backup instead. A good rule of thumb is that when periodic differential backups exceed 6-10 high density diskettes or exceed 30 minutes to complete, it's probably time for another full tape backup.

Restoration And Purging

Before discussing the subject of purging old, unneeded backup files, let's consider how, after using the generic backup strategy above, you would go about completely restoring the system following a major disaster. The strategy of restoration follows

two simple rules: restore from the oldest backup set forward to the newest, and from the most complete backup set to the least. Here's how it would work:

Starting with a bootable but otherwise empty hard drive, you would install the tape backup software and restore the last full tape backup to the system. Next, restore the last differential backup set (the software was reinstalled from tape). Finally, restore, from oldest to newest, the files from the daily .ZIP backups. If the backups were made correctly, the new system would be current with the end of the last work day. For a 120Mb hard drive this should take less than two hours to complete. Compare that with how long it would take to even assess what had been lost if no backups existed to restore!

A good idea is to periodically test a backup by attempting to restore some files from it. All three methods used above allow some sort of selective retrieval of files from the backup set. Both the tape backup and PKZip let you optionally compare the backed-up data with the original data during (or immediately following) the backup. These safeguards should be used whenever possible.

Using the generic strategy also periodically frees you to delete, or "purge," the older, less comprehensive backup sets. For example, after a weekly differential backup you could delete the individual daily .ZIP files for the same week (since those files now also reside in the differential backup). And, after a tape backup both the previous week's daily files and the last differential backup can be deleted. Carefully consider the need for retaining multiple, serial backups of the system, and only maintain the purging schedule when the backup schedule is followed correctly.

Re-Evaluating Needs: A Testing Period

Whatever backup strategy individuals use will depend upon how they work. The generic backup strategy discussed above may be all that is needed; try it for a time and see how it works. After this test period, re-evaluate the backup strategy and adjust it as needed. And don't forget to share any unique and efficient backup solutions with the rest of us!

A Final View

The final summary includes the most important points for developing a backup strategy for the work computer. Assume that the machine will fail and critical files will need to be restored from backups. Also assume that this will happen at the worst possible time, affecting a great number of people who need (demand?) results. A variety of tools are available for backing up either work files or the complete system. Used correctly these tools minimize the time spent with backups each day while offering flexible options for complete backup security.

Make a backup schedule and follow it. Files considered too "trivial" to back up usually take the least amount of time and space; so back them up anyway. Only purge old backup sets when the newer replacement has already been created. Keep copies of

backup sets--especially tapes and weekly differentials--in another building away from the computer. Periodically test the backups by attempting a selective restore. And finally, adjust the backup strategy for maximum efficiency and protection. Remember, you have nothing to lose--except hours, weeks, and months of work...

Appendices

Data Personnel Hierarchy and Responsibilities (Example Job Descriptions)

Computer Support. Computer support personnel have a primary responsibility to service, install and maintain computer hardware and software, including network systems. They have full authority in issues of system compatibility, configuration standards, hardware, software, and security. Computer support personnel generally have no authority or responsibility for data design criteria, maintaining user data, or for carrying out procurement procedures. Their principal responsibilities are listed in the example box at right.

Example Computer Support Responsibilities

- maintain computer tools, supplies and diagnostic software to keep equipment operational
- ensure continuous network operation and maintain viable network backups
- collect and maintain data on individual setups and configurations for trouble-shooting
- respond to computer support requests in a timely manner
- log all support work to store common solutions and avoid repeat visits
- report all instances of abuse or misuse of government computer resources
- perform, supervise and/or approve all software installations to maintain existing standards (dirs, ini's, versions, etc.); record settings
- perform, supervise and/or approve all hardware installations to ensure proper setup
- provide hardware purchase consultation for adherence to existing standards
- maintain a media safe containing all original software application diskettes labelled with property numbers, and cataloged with license info, user/server assignment, etc.
- make sure the property manager keeps records of warranties and is updated on serial numbers of replacement items
- keep records of technical support contracts and special PINs and expirations
- maintain current phone, BBS, FAX and address list of vendors
- stay informed about the latest drivers, patches, , upgrades, etc. for existing hardware and software

Data Manager. The data manager has primary authority over the master and archival datasets, data security, data access and dissemination, maintenance of data documentation, and overall database design and standards issues. He generally has no direct responsibility for generating or validating data, but has authority over whether data are considered complete enough for inclusion into a master data series. As the final authority on all data design and structure issues he is also involved in the evaluation of field data forms and data entry modules. His principal responsibilities are listed in the example box at right.

Example Data Manager Responsibilities

- serve as the data "master" or gate-keeper, allowing only fully documented, validated data to become part of the master data series
- provide clean copies of master data and associated documentation to users
- protect archival copies of data and documentation
- develop a data integration schema and maintain the primary data dictionary
- insure all projects meet minimum standards for field structures and integration criteria
- provide project managers with written guidelines for data preparation, editing, post-processing and documentation
- provide basic training in the use of installed data management and manipulation tools
- explore avenues for electronic data-sharing and dissemination
- be an active member of park's ADP committee
- be a member and participator at regional and national levels regarding data standards and integration/analysis issues
- maintain contact with other data managers
- stay current on general and federal data issues

Project Manager. Individual project managers have responsibility for field work and subject matter pertaining to their specialties or assignments. Where these individuals *are not park staff* (such as contractors and cooperators) they are still similarly responsible for recognizing and complying with relevant Park data collection and data management standards (also see the section on ownership agreements). Project managers are the authorities within their projects for designing statistically sound data collection schemes, adhering to accepted scientific standards, maintaining quality control at all phases of the work, and developing and reporting from their datasets. Their primary data-related responsibilities are listed in the example box at right.

Example Project Manager Responsibilities

- supervise and certify all field operations, including training, data collection, and testing
- provide fully documented master data and version updates to the data manager
- maintain hard-copy files of data; ensure copies are stored in second location
- guarantee proper version control and associated documentation for data entry and editing
- maintain concise explanatory documentation of all deviations from SOP's
- clear all changes to data structures and field data forms with the data manager
- create general access tools (queries, reports) for users to access their data
- supervise or carry out all in-house analyses and reports on their data, store the results, and make them available to users
- be the main contact about the data content

End Users and Customers. Individual end users are typically park staff who access data and have specific responsibilities toward the computer resources at their workstation; end users may also be management and outside data customers. As are the primary customers for the data collected, end users are important to active data management. End users are the target audience in the creation and maintenance of access tools, reports, and queries. Users must cooperate with and facilitate computer support operations, produce and store viable backup and archival copies of their work, and maintain the basic skills to be effective with the technological tools required to perform their jobs. End users generally have no authority over data or standards issues unless they are also a project or data manager. However, consultation with and feed-back from users are important. End users responsibilities are listed at right.

Example End User and Customer Responsibilities

- maintain their equipment (e.g.,monitor, keyboard, and mouse) in clean, working condition and free of risk by food, drink, etc.
- keep all hardware/software manuals accessible
- keep all property stickers accessible to an audit
- produce and maintain viable disaster recovery backups of their full system and incremental work
- maintain own filing system for working files and archival copies of important work, both paper and disk-based
- maintain familiarity with the basic operation and maintenance of their system
- attempt to diagnose and correct computer problems by using manuals, vendor technical support, and other options before calling computer support; maintain notes of corrective attempts
- seek remedial or advanced training when needed
- periodically clean-up their work directories
- report unusual computer incidents or malfunctions to computer support immediately
- establish and enforce computer rules for visitors
- obtain authorization from computer support for all major system configuration changes
- obtain authorization and/or assistance from computer support for all new software/hardware installations
- virus check all disks used in their machines--even from old storage sets and co-workers
- respect LAN access and storage privileges
- not disseminate data or results of informal analyses without proper authorization

Dataset Catalog Entry Form

Copy this form before use! Fill out a separate form for each dataset. Numbers after the field name indicate the character limit for that field. All fields must be completed.

Subject (30): _____

Keywords (70): _____

Dataset Title (70): _____

Version (10): _____ **Project ID** (20): _____

Dataset Description (250): _____

_____

_____

Related Documentation (210): _____

_____

Related Datasets (140): _____

_____

Date(s) (Date/Time) **Begin Date:** _____ **End Date:** _____

Multi-dates/Notes (160): _____

Status (10): New __ Active __ Inactive __ Historic __ **Update Frequency** (10): _____

Places (6): In ___ Out ___ In&Out ___

Location (100): _____

Longitude (Double): _____ **Latitude** (Double): _____

UTM Zone (Int): _____ **Easting** (Long): _____ **Northing** (Long): _____

Data Type (6): GEORAS __ GEOVEC __ GEODB __ DIGRAS __ DIGVEC __ DIGDB __ ANAORG __ ANAUNO __

Table/Layer Name(s) (200): _____

_____

Scale(s) (130): _____

Quality (15): Unknown __ *Not* Ver./Val.(?) __ Verified __ Validated __ Metadata __

Data Format (80): Paper ___ dBASE ___ Lotus ___ WordPerfect ___ ASCII ___

Other (list): _____

Data Origin (200): Name/Source, Position _____

Affiliation, Contact Info _____

_____

Dataset Contact _____

_____

Distribution (50): _____

File Location (50): _____

Access Options (10): Public ___ Restricted ___ (briefly describe below)

Investigator's Annual Report (IAR)
Research Project Subject Categories

Air Quality
Archeology
Botany
Cave (Flora/Fauna)
Cave/Karst
Climatology
Coastal/Marine Systems
Contaminants/ Haz. Mat.
Ecology
Entomology
Environmental Monitoring
Erosion/Sedimentation
Exotic Species - Animals
Exotic Species - Plants
Fire
Fisheries Management
Flood Management/History
Forestry
Fungi
Geo-Hazard (Chemical)
Geo-Hazard (Physical)
Geographic Information System
Geochemistry
Geohydrology
Geology - Coastal
Geology - Fluvial
Geology - General
Geology - Structural
Geomorphology
Geophysics
Glaciology
Herpetology
History
Hydrology (Ground)
Hydrology (Surface)
Ichthyology
Integrated Pest Mgmt.
Invertebrates
Limnology
Mammalogy
Management/Administration
Microbiology
Minerals Management
Oceanography
Ornithology
Paleontology
Petrology/Mineralogy
Range Management
Recreation/Aesthetics
Restoration - Cultural
Restoration - Natural
Sedimentology/Stratigraphy
Sociology
Soil Science
Tectonics
Threatened/Endangrd Animals
Threatened/Endangrd Plants
Volcanology/Geothermal Water Quality
Water Quantity
Water Rights
Watershed Management
Wetlands
Wildlife Management
Zoology
Other

Geographic and UTM Coordinates of National Park Centroids*

Park Code	Latitude	Longitude	UTM	UTM	Zone
ABLI	37.53226	-85.73354	4154469	611899	16
ACAD	44.29047	-68.32905	4904137	553531	19
ADAM	42.2564	-71.01193	4679990	334039	19
AGFO	42.42169	-103.7533	4697136	602571	13
ALFL	35.58192	-101.6712	3940663	257952	14
ALPO	40.43256	-78.56721	4478400	706361	17
AMIS	29.53489	-101.075	3268866	298915	14
AMME	15.21713	145.72102	1682640	362628	55
ANDE	32.19775	-84.13035	3565778	770525	16
ANIA	56.87261	-157.8091	6303624	572592	4
ANJO	36.15589	-82.83733	4002601	334719	17
ANTI	39.47005	-77.73838	4372311	264432	18
APCO	37.38006	-78.79886	4139104	694887	17
APIS	46.96537	-90.6478	5203782	678944	15
ARCH	38.71957	-109.5883	4286396	622732	12
ARHO	38.88208	-77.07443	4305528	320060	18
ARPO	34.02061	-91.34686	3765478	652638	15
ASIS	38.06143	-75.23435	4212450	479441	18
AZRU	36.83682	-107.9992	4080766	232532	13
BADL	43.68555	-102.4818	4839814	702966	13
BAND	35.77725	-106.3277	3959855	379995	13
BEFR	39.95811	-75.1736	4422912	485172	18
BELA	65.9551	-164.401	7314845	527237	3
BEOL	38.03961	-103.4262	4211173	638112	13
BIBE	29.29738	-103.229	3242057	672019	13
BICA	45.02487	-108.2132	4989274	719555	12
BICY	25.97007	-81.08136	2872209	491855	17
BIHO	45.647	-113.6493	5058024	293556	12
BISC	25.49011	-80.20879	2819294	579524	17
BISO	36.5346	-84.66518	4045580	709024	16
BITH	30.43879	-94.3616	3368013	369249	15
BLCA	38.5751	-107.7127	4272909	263682	13
BLRI	36.55158	-80.99908	4044927	500082	17
BLUE	37.56517	-80.98436	4157367	501381	17
BOAF	42.35846	-71.06852	4691435	329646	19
BOST	42.37241	-71.05243	4692952	331009	19
BOWA	37.11492	-79.73252	4108169	612610	17
BRCA	37.5837	-112.182	4160079	395634	12
BRCR	34.50616	-88.72888	3819441	341282	16
BUFF	36.04083	-92.90694	3988280	508383	15
BUIS	17.79123	-64.62517	1967716	327732	20
CABR	32.67251	-117.2403	3614817	477473	11
CACA	32.14049	-104.5528	3555907	542176	13
CACH	36.14399	-109.3354	4001001	649765	12
CACO	41.9347	-70.04613	4642841	413270	19
CAGR	32.99705	-111.5319	3650894	450309	12
CAHA	35.47265	-75.61788	3925436	443941	18
CAKR	67.40655	-163.4886	7477306	564797	3
CALO	34.82621	-76.34389	3854396	377102	18
CANA	28.78652	-80.75503	3184183	523909	17
CANY	38.24508	-109.8793	4233395	598067	12
CARE	38.17895	-111.1765	4225478	484545	12
CARL	35.26828	-82.45115	3903561	368006	17
CASA	29.89759	-81.31343	3307299	469738	17
CATO	39.65014	-77.46383	4391619	288601	18

Park Code	Latitude	Longitude	UTM	UTM	Zone
CAVO	36.78175	-103.9705	4070953	591868	13
CEBR	37.63548	-112.8445	4166767	337252	12
CHAM	31.76719	-106.4539	3515363	362320	13
CHAT	33.99678	-84.28972	3764916	750338	16
CHCH	34.94306	-85.28749	3867870	656389	16
CHCU	36.03838	-107.9516	3992036	234067	13
CHIC	34.46045	-97.0068	3814818	683086	14
CHIR	32.01205	-109.3409	3542786	656702	12
CHIS	33.98632	-119.9124	3764269	230952	11
CHOH	39.40776	-77.90804	4365854	249611	18
CHPI	32.8429	-79.82484	3634292	609978	17
CHRI	17.74877	-64.70277	1963090	319462	20
CHRO	41.70233	-103.3468	4617834	637557	13
CIRO	42.06991	-113.7115	4660882	275670	12
CODA	48.169	-118.3502	5335746	399606	11
COLM	39.05044	-108.6913	4324702	699793	12
COLO	37.22804	-76.62055	4121196	356232	18
CORO	31.349	-110.2553	3468335	570840	12
COSW	33.79676	-80.773	3739450	521013	17
COWP	35.13147	-81.8094	3887724	426256	17
CRLA	42.9408	-122.1316	4754391	570853	10
CRMO	43.40298	-113.5196	4808434	295966	12
CUGA	36.63693	-83.58161	4057499	269182	17
CUIS	30.85828	-81.45273	3413800	456714	17
CURE	38.46764	-107.3219	4260050	297425	13
CUVA	41.25931	-81.57102	4567488	452163	17
DENA	63.28841	-151.0553	7019004	597524	5
DEPO	37.61524	-119.0866	4164969	315833	11
DESO	27.52302	-82.64362	3045273	337673	17
DETO	44.59082	-104.7151	4937319	522611	13
DEVA	36.44547	-117.0177	4033157	498412	11
DEWA	41.12647	-74.94889	4552585	504290	18
DINO	40.50704	-108.9324	4485878	675191	12
DRTO	24.63991	-82.87492	2726215	310229	17
EBLA	48.21401	-122.6863	5339915	523302	10
EDAL	39.96187	-75.15093	4423326	487109	18
EDIS	40.78562	-74.23892	4515026	564217	18
EFMO	43.07167	-91.18568	4770155	647716	15
EISE	39.79573	-77.26532	4407330	306043	18
ELMA	34.88404	-107.9939	3864077	226368	13
ELMO	35.03979	-108.3433	3880485	742348	12
ELRO	41.76198	-73.89941	4623721	591490	18
EUON	37.82613	-122.0262	4186765	585706	10
EVER	25.37178	-80.88203	2805961	511868	17
FIIS	40.69653	-73.0011	4506779	668889	18
FLFO	38.91233	-105.28	4306877	475728	13
FOBO	32.14917	-109.4505	3557833	646131	12
FOBU	41.85684	-110.7612	4633695	519822	12
FOCA	30.38435	-81.49804	3361299	452150	17
FOCL	46.13404	-123.8784	5109097	432142	10
FODA	30.59988	-103.8944	3385598	605989	13
FODC	36.48752	-87.8477	4038155	424069	16
FODO	36.48695	-87.85724	4038100	423214	16
FOFR	31.22109	-81.39459	3453987	462416	17
FOLA	42.20331	-104.546	4672235	537479	13
FOLS	38.18037	-99.218	4225644	480906	14
FOMA	29.71162	-81.23609	3286675	477163	17
FOMC	39.26382	-76.58071	4347035	363626	18

Park Code	Latitude	Longitude	UTM	UTM	Zone
FONE	39.81344	-79.59207	4407789	620512	17
FOPO	37.8085	-122.4727	4184493	546416	10
FOPU	32.0307	-80.93411	3543652	506221	17
FORA	35.93671	-75.71245	3976962	435735	18
FOSC	37.8436	-94.70396	4189625	350070	15
FOSM	35.38756	-94.43508	3916770	369659	15
FOST	43.21024	-75.45595	4784047	462962	18
FOSU	32.75637	-79.87387	3624649	605491	17
FOUN	35.90688	-105.0139	3973420	498742	13
FOUS	48.0007	-104.0365	5316608	571871	13
FOVA	45.62332	-122.6606	5052035	526459	10
FRED	38.29273	-77.46948	4240973	284029	18
FRHI	39.77529	-79.92548	4403159	592024	17
FRSP	38.2824	-77.64496	4240251	268647	18
GAAR	67.74892	-153.2972	7514717	487439	5
GARI	38.2059	-81.00486	4228454	499574	17
GATE	40.57174	-73.91054	4491576	592219	18
GERO	38.67908	-87.53555	4281093	453415	16
GETT	39.81543	-77.23252	4409446	308906	18
GEWA	38.19314	-76.92182	4228784	331701	18
GICL	33.2265	-108.2687	3679531	754534	12
GLAC	48.68326	-113.7993	5395808	293959	12
GLBA	58.84118	-136.8827	6523686	391339	8
GLCA	37.5151	-110.7726	4151836	520100	12
GLDE	39.93413	-75.14441	4420246	487661	18
GOGA	37.82843	-122.5283	4186678	541508	10
GOSP	41.62013	-112.5223	4608507	373174	12
GRBA	38.94586	-114.2571	4314138	737715	11
GRCA	36.1752	-112.6673	4004467	350055	12
GREE	38.98272	-76.89824	4316364	335577	18
GRKO	46.41335	-112.7456	5141237	365838	12
GRPO	47.99265	-89.75519	5318938	294447	16
GRSA	37.76581	-105.5614	4179774	450557	13
GRSM	35.60049	-83.50866	3942336	272736	17
GRSP	38.02731	-78.16567	4212431	748785	17
GRTE	43.81812	-110.7049	4851498	523736	12
GUCO	36.13284	-79.84182	3999102	604213	17
GUIS	30.29737	-87.76121	3351801	426800	16
GUMO	31.92353	-104.8844	3531777	510932	13
GWCA	36.98698	-94.35496	4094082	379414	15
GWMP	38.91938	-77.13838	4309796	314610	18
HAFE	39.32357	-77.72971	4356029	264686	18
HAFO	42.7905	-114.9438	4739386	668172	11
HALE	20.71963	-156.1652	2293568	795262	4
HAMP	39.41744	-76.58799	4364096	363297	18
HAVO	19.38094	-155.3267	2144497	255621	5
HEHO	41.66792	-91.35237	4614005	637166	15
HOBE	32.97623	-85.73417	3649172	618286	16
HOCU	39.33253	-83.04324	4355460	323891	17
HOFR	41.76688	-73.93855	4624224	588229	18
HOFU	40.20603	-75.76858	4450697	434589	18
HOME	40.28816	-96.83461	4461780	684069	14
HOSP	34.52349	-93.06348	3820007	494173	15
HOVE	37.39127	-109.0042	4139943	676681	12
HSTR	39.09374	-94.42244	4327933	376985	15
HUTR	35.70722	-109.5592	3952232	630340	12
INDE	39.94932	-75.14832	4421932	487329	18
INDU	41.6458	-87.07843	4610241	493468	16

Park Code	Latitude	Longitude	UTM	UTM	Zone
INTE	48.98918	-100.0627	5426577	422253	14
ISRO	47.98705	-88.88728	5316365	359180	16
JAGA	41.66491	-81.3511	4612419	470769	17
JECA	43.73107	-103.8295	4842453	594268	13
JEFF	38.62278	-90.17449	4278497	745988	15
JELA	29.81788	-90.14769	3301839	775663	15
JICA	32.03026	-84.41736	3546518	743905	16
JODA	44.62522	-119.8777	4945130	271710	11
JODR	44.08951	-110.7273	4881634	521828	12
JOFI	42.34646	-71.12395	4690215	325048	19
JOFK	38.89584	-77.05589	4307019	321703	18
JOFL	40.34578	-78.77492	4468302	688983	17
JOMU	37.9829	-122.1328	4204066	576157	10
JOTR	33.93172	-115.9273	3754909	599146	11
KAHO	19.68394	-156.0334	2179090	811046	4
KALA	21.17648	-156.9538	2342912	712449	4
KATM	58.61967	-155.0119	6499245	383147	5
KEFJ	59.81576	-150.1065	6634228	662262	5
KEMO	33.95328	-84.59211	3759393	722514	16
KEWE	47.17959	-88.52238	5226027	384646	16
KICA	36.89169	-118.5971	4083845	357687	11
KIMO	35.1388	-81.39005	3888307	464466	17
KLGO	59.36042	-135.2596	6580006	485239	8
KNRI	47.35407	-101.3854	5247052	319850	14
KOVA	67.35369	-159.1955	7470637	491599	4
LABE	41.75929	-121.5158	4623902	623387	10
LACH	48.35401	-120.6821	5358025	671722	10
LACL	60.56744	-153.5555	6714531	469548	5
LAME	35.96497	-114.3452	3983121	739406	11
LAMR	35.61908	-101.6812	3944811	257160	14
LAVO	40.49366	-121.4059	4483560	635090	10
LIBI	45.55621	-107.4168	5047364	311363	13
LIBO	38.118	-86.99682	4218701	500278	16
LIHO	39.79704	-89.64506	4408368	273530	16
LIRI	34.42291	-85.61287	3809725	627467	16
LOWE	42.64481	-71.31917	4723765	309873	19
LYJO	30.25261	-98.60512	3346662	537989	14
MACA	37.19821	-86.13118	4117010	577105	16
MALU	33.75491	-84.372	3737890	743423	16
MALW	37.54777	-77.43885	4158231	284545	18
MANA	38.81878	-77.53368	4299509	280024	18
MAVA	42.36937	-73.70413	4691387	606701	18
MCHO	45.35731	-122.6058	5022502	530877	10
MEVE	37.23843	-108.4621	4124137	725128	12
MIMA	42.45495	-71.30285	4702645	310638	19
MISS	44.8742	-93.04626	4968759	496345	15
MOCA	34.62323	-111.8114	3831364	425616	12
MOCR	34.45831	-78.11056	3816567	765438	17
MONO	39.36208	-77.39579	4359487	293588	18
MORA	46.86074	-121.7033	5190286	598838	10
MORR	40.76495	-74.53532	4512557	539220	18
MORU	43.88042	-103.4525	4859539	624320	13
MUWO	37.89662	-122.5795	4194222	536970	10
NABR	37.60466	-110.0032	4162215	587991	12
NACC	38.88432	-77.03339	4305696	323626	18
NACE	38.75671	-77.01635	4291501	324792	18
NASA	-14.25	-169.882	-1575572	620610	2
NATC	31.54671	-91.39046	3491130	652785	15

Park Code	Latitude	Longitude	UTM	UTM	Zone
NATR	33.95863	-88.89622	3758994	324787	16
NAVA	36.70198	-110.545	4061706	540640	12
NEPE	45.85051	-116.3073	5077453	553786	11
NERI	37.86505	-80.99981	4190636	500016	17
NIMI	42.79663	-98.2487	4738290	561440	14
NISI	34.14266	-82.01658	3778245	406274	17
NOAT	68.00441	-159.8383	7543410	464960	4
NOCA	48.71065	-121.2034	5396629	632171	10
OBRI	36.10792	-84.77483	3998010	700296	16
OCMU	32.83802	-83.60235	3636140	256415	17
OLYM	47.80251	-123.6624	5294343	450398	10
ORCA	42.09678	-123.407	4660388	466340	10
ORPI	32.03519	-112.8566	3545655	324685	12
OZAR	37.13954	-91.25755	4111569	654762	15
PAAL	26.0244	-97.46213	2879129	653888	14
PAIS	27.05454	-97.35727	2993370	662921	14
PECO	35.54209	-105.6811	3933175	438263	13
PEFO	35.01758	-109.8071	3875445	608836	12
PERI	36.45421	-94.03467	4034624	407281	15
PETE	37.19234	-77.47319	4118870	280476	18
PETR	35.13894	-106.7493	3889653	340628	13
PEVI	41.65453	-82.81168	4612792	349145	17
PIMA	33.18706	-111.9184	3672208	414385	12
PINN	36.48494	-121.1873	4039063	662382	10
PIPE	44.01353	-96.32443	4876640	714471	14
PIRO	46.56339	-86.31654	5156655	552383	16
PISP	36.8627	-112.7391	4080850	344970	12
POPO	32.64231	-91.40818	3612563	649309	15
PORE	38.05344	-122.8806	4211545	510480	10
PRES	37.79719	-122.465	4183242	547104	10
PRWI	38.58837	-77.38726	4273594	292071	18
PUHE	20.03046	-155.8236	2217214	204599	5
PUHO	19.42018	-155.9104	2149774	194354	5
RABR	37.07789	-110.9639	4103310	503208	12
REDW	41.37165	-124.0293	4580313	413919	10
RICH	37.47496	-77.33563	4149920	293464	18
ROCR	38.95165	-77.04938	4313200	322407	18
ROLA	48.78179	-121.1024	5404716	639399	10
ROMO	40.3554	-105.6968	4467226	440830	13
ROWI	41.83042	-71.41137	4633546	299754	19
RUCA	34.97436	-85.81591	3870642	608089	16
SAAN	29.30729	-98.42319	3241992	556015	14
SACN	45.69167	-92.36569	5059768	549389	15
SACR	45.12787	-67.13415	4998631	646734	19
SAFR	37.80834	-122.4224	4184502	550846	10
SAGA	43.49858	-72.37543	4819314	712196	18
SAGU	32.2056	-110.7417	3563066	524347	12
SAHI	40.88567	-73.49771	4526939	626568	18
SAIR	42.46832	-71.00828	4703515	334896	19
SAJH	48.49933	-123.0451	5371583	496670	10
SAJU	18.47055	-66.11884	2044549	804295	19
SAMA	42.51998	-70.88705	4709022	344990	19
SAMO	34.09474	-118.7663	3773874	337053	11
SAPA	40.89273	-73.82607	4527300	598892	18
SAPU	34.3565	-106.2064	3802148	389052	13
SARA	42.99511	-73.6347	4760961	611296	18
SCBL	41.83492	-103.707	4632042	607365	13
SEQU	36.50849	-118.5736	4041299	359082	11

Park Code	Latitude	Longitude	UTM	UTM	Zone
SHEN	38.49259	-78.46879	4263300	720760	17
SHIL	35.13843	-88.34207	3889020	377733	16
SITK	57.04713	-135.3128	6322463	481022	8
SLBE	44.92763	-86.0272	4975154	576770	16
SPAR	42.10851	-72.58058	4664443	700041	18
STEA	41.40727	-75.67006	4583973	443993	18
STLI	40.69543	-74.04276	4505176	580877	18
STRI	35.87603	-86.43079	3970147	551383	16
SUCR	35.37112	-111.5095	3914120	453714	12
THKO	39.94324	-75.14783	4421257	487370	18
THRO	47.17377	-103.4298	5225452	618993	13
THST	38.52922	-77.03817	4266297	322334	18
TICA	40.44026	-111.7088	4476653	439886	12
TIMU	30.47374	-81.49929	3371205	452073	17
TONT	33.64855	-111.1127	3723001	489546	12
TUIN	32.42944	-85.70481	3588586	621771	16
TUMA	31.56854	-111.0496	3492428	495295	12
TUPE	34.25536	-88.73717	3791640	340045	16
TUZI	34.77287	-112.027	3848139	406018	12
UPDE	41.66231	-75.04658	4612071	496121	18
VAFO	40.10047	-75.44612	4438793	461974	18
VAMA	41.79694	-73.94258	4627558	587853	18
VICC	32.37154	-90.87091	3583423	700312	15
VICK	32.3617	-90.84957	3582372	702343	15
VIIS	18.34489	-64.74203	2029102	315920	20
VOYA	48.48337	-92.83707	5369820	512040	15
WACA	35.16908	-111.5026	3891711	454232	12
WAPA	13.43553	144.68867	1486381	249736	55
WEFA	41.25807	-73.45482	4568344	629449	18
WHIS	40.61358	-122.6011	4495727	533748	10
WHMI	46.04135	-118.4617	5099462	386896	11
WHSA	32.77894	-106.3327	3627376	375186	13
WICA	43.58759	-103.4394	4827036	625985	13
WICR	37.10195	-93.4084	4106056	463709	15
WIHO	39.11945	-84.5074	4332782	715494	16
WORI	42.91168	-76.79123	4752348	353785	18
WOTR	38.93843	-77.26505	4312175	303679	18
WRBR	36.01606	-75.6702	3985736	439606	18
WRST	61.39071	-142.583	6807129	415436	7
WUPA	35.5573	-111.3946	3934720	464233	12
YELL	44.5966	-110.5462	4938022	536015	12
YOSE	37.84818	-119.5563	4191844	275081	11
YUCH	65.08983	-142.7958	7219472	415605	7
YUHO	37.24801	-108.6855	4124692	705284	12
ZION	37.29895	-113.0258	4129756	320448	12
ZUCI	35.02283	-108.8674	3877455	694565	12

* **Note:** The Park centroid locations were calculated with Atlas GIS from the digital Park boundaries and may not fall within the boundary of Parks with an asymmetrical shape.

Dataset Catalog Installation Instructions for MS Access

These instructions assume that the user is familiar with file handling, MS DOS, and Microsoft Access. Otherwise, some additional assistance may be needed.

1. Create or Make a subdirectory in the root directory called DATACAT (e.g., C:\DATACAT).

2. Copy the Dataset Catalog files into the new DATACAT directory.

3. Expand the self-extracting compressed database files.

4. Rename the DATCATxx.MDB and DATCATxx.LDB files to:

   <PARKCODE>CATx.MDB and <PARKCODE>CATx.LDB
   (e.g., CHISCAT1.MDB for Channel Islands National Park)

Note: Lower case x denotes the software or database version number.

5. Start Microsoft Access and open the database.

6. If Access cannot find some or all of the needed files, they must be re-attached with the correct path. To re-attach the files, click on the following:

   File -> Add-Ins -> Attachment Manager -> (mark all files) -> OK ->
   -> (Follow any additional instructions/steps) -> OK -> Close

7. The catalog database should now be ready to use...

For additional assistance contact:

Joe Gregson
Inventory and Monitoring Program
National Park Service
1201 Oak Ridge Drive, Suite 350
Fort Collins, CO 80525

Phone: 970-225-3559
Fax: 970-225-3585
Email: NPS ccMail (Joe_Gregson@nps.gov)

Dataset Catalog Installation Instructions for Windows Executable Database Software

These instructions assume that the user is familiar with file handling, MS DOS, and Microsoft Windows. Otherwise, some additional assistance may be needed.

1. If downloaded from the Internet:

   A. Create or Make a temporary subdirectory in the root directory called TEMP (e.g., C:\TEMP).

   B. Copy the Dataset Catalog files into the new TEMP directory.

   C. Expand the self-extracting compressed database files in the TEMP directory.

2. From Windows click on File->Run or Start->Run to run the SETUP.EXE file:

   A:\SETUP.EXE for Disk 1 of floppy disk set, or
   C:\TEMP\SETUP.EXE for the Internet downloaded set.

3. Follow the Windows setup screens and accept the default paths. **

** Note that if the default installation paths are changed, the Dataset Catalog will be unable to find the attached database tables and help files.

4. The setup program may report several errors of files that cannot be found depending on the individual Windows components and configuration. Click OK for each error message as the missing files are generally not needed for the executable database.

5. When the setup program is complete, start the Dataset Catalog by double clicking on its icon.

6. The Dataset Catalog database should now be ready to use...

For additional assistance contact:

Joe Gregson
Inventory and Monitoring Program
National Park Service
1201 Oak Ridge Drive, Suite 350
Fort Collins, CO 80525

Phone: 970-225-3559 Fax: 970-225-3585
Email: NPS ccMail (Joe_Gregson@nps.gov)

Dataset Catalog Database Dictionaries

C:\DATACAT\DATACAT4.MDB

DATACAT Table Field Definitions

Name	Type	Size
ID	Number (Long)	4
CATDATE	Date/Time	8
PARKCODE	Text	4
PARKNAME	Text	40
ADMIN	Text	30
SUBJECT	Text	30
KEYWORDS	Text	70
TITLE	Text	70
VERSION	Text	10
PROJ_ID	Text	20
DESCRIPTION	Text	250
RELDOCS	Text	210
RELDATA	Text	140
BEG_DATE	Date/Time	8
END_DATE	Date/Time	8
TIMES2	Text	160
STATUS	Text	10
UDDATFREQ	Text	10
PLACES	Text	6
LOCATION	Text	100
LAT	Number (Double)	8
LON	Number (Double)	8
UTM_EAST	Number (Double)	8
UTM_NORTH	Number (Double)	8
UTM_ZONE	Number (Integer)	2
DATATYPE	Text	6
TABL/LAYR	Text	200
QUALITY	Text	15
SCALE	Text	130
FORMAT	Text	80
ORIGIN	Text	200
CONTACT_ID	Number(Long)	4
CONTACT	Text	30
ACCESS	Text	10
DISTRIBUTION	Text	50
FILE_LOC	Text	50

CONTACT Table Field Definitions

Name	Type	Size
CONTACT_ID	Number (Long)	4
CNTCT_PERSN	Text	30
CNTCT_POS	Text	50
CNTCT_ORG	Text	50
CNTCT_ADDR	Text	50
CNTCT_CITY	Text	50
CNTCT_ST	Text	2
CNTCT_ZIP	Text	10
CNTCT_PHONE	Text	15
CNTCT_FAX	Text	15
CNTCT_EMAIL	Text	50

SPATIAL Table Field Definitions

Name	Type	Size
REC_ID	Number (Long)	4
PARKCODE	Text	4
WEST_LON	Number (Double)	8
EAST_LON	Number (Double)	8
NORTH_LAT	Number (Double)	8
SOUTH_LAT	Number (Double)	8
WEST_UTM	Number (Double)	8
EAST_UTM	Number (Double)	8
NORTH_UTM	Number (Double)	8
SOUTH_UTM	Number (Double)	8
UTM	Number (Double)	8

C:\DATACAT\PARKRECT.DBF

PARKRECT.DBF Field Definitions

Name	Type	Size
PARKCODE	Text	14
WEST_LON	Number (Double)	8
EAST_LON	Number (Double)	8
NORTH_LAT	Number (Double)	8
SOUTH_LAT	Number (Double)	8
WEST_UTM	Number (Double)	8
NORTH_UTM	Number (Double)	8
EAST_UTM	Number (Double)	8
SOUTH_UTM	Number (Double)	8

C:\DATACAT\PARKCENT.DBF

PARKCENT.DBF Field Definitions

Name	Type	Size
PARKCODE	Text	4
LAT	Number (Double)	8
LON	Number (Double)	8
NORTH_UTM	Number (Double)	8
EAST_UTM	Number (Double)	8
ZONE	Number (Double)	8

C:\SUBJECT.DBF

SUBJECT.DBF Field Definitions

Name	Type	Size
SUBJECT	Text	30

C:\DCATOUT.DBF

DCATOUT.DBF Field Definitions

Name	Type	Size
ID	Number (Double)	8
CATDATE	Date/Time	8
PARKCODE	Text	4
PARKNAME	Text	40
ADMIN	Text	30
SUBJECT	Text	30
KEYWORDS	Text	70
TITLE	Text	70
VERSION	Text	10
PROJ_ID	Text	20
DESCRIPTION	Text	250
RELDOCS	Text	210
RELDATA	Text	140
BEG_DATE	Date/Time	8
END_DATE	Date/Time	8
TIMES2	Text	160
STATUS	Text	10
UDDATFREQ	Text	10
PLACES	Text	6
LOCATION	Text	100
CEN_LAT	Number (Double)	8
CEN_LON	Number (Double)	8
UTM_EAST	Number (Double)	8
UTM_NORTH	Number (Double)	8
UTM_ZONE	Number (Double)	8
WEST_LON	Number (Double)	8
EAST_LON	Number (Double)	8
NORTH_LAT	Number (Double)	8
SOUTH_LAT	Number (Double)	8
WEST_UTM	Number (Double)	8
EAST_UTM	Number (Double)	8
NORTH_UTM	Number (Double)	8
SOUTH_UTM	Number (Double)	8
DATATYPE	Text	6
TABL/LAYR	Text	200
QUALITY	Text	15
SCALE	Text	130
FORMAT	Text	80
ORIGIN	Text	200
CONTACT_ID	Number(Long)	4
CONTACT_PERS	Text	30
CNTCT_POS	Text	50
CNTCT_ORG	Text	50
CNTCT_ADDR	Text	50
CNTCT_CITY	Text	50
CNTCT_ST	Text	2
CNTCT_ZIP	Text	10
CNTCT_PHONE	Text	15
CNTCT_FAX	Text	15
CNTCT_EMAIL	Text	50
ACCESS	Text	10
DISTRIBUTION	Text	50
FILE_LOC	Text	50

Data Edit Report Form

Name: _____

Date: _____ (completion date)

File(s) Edited: _____

Reason for Edit: _____

_____

_____

_____

Program(s) Used: _____

_____

Original File Information: _____

_____

Final File Information: _____

_____

Editing Details: _____

_____

_____

_____

_____

_____

_____

_____

_____

Resource Management Plan
Example Project Statement

SHEN-I-000.000
Priority: 999

Last Update: 02/02/96
Initial Proposal: 1996

Title: **SAMPLE DATA MANAGEMENT PROJECT**

Funding Status: Funded: 0.00 Unfunded: 1.53

Servicewide Issues: N24 (OTHER (NATURAL))
Cultural Resource Type:
N-RMAP Program codes: C00 (Collections and Data Management)
C03 (GIS/Data Management)

10-238 Package Number:

Problem Statement

Six legacy datasets have been identified for updating, documentation, and inclusion in the Park baseline inventory database. The legacy datasets include...

Description of Recommended Project or Activity

Each dataset exists in DBASE III+ format and will be edited for validity via automated bounds checking, etc. in ACCESS DBMS. A comprehensive data dictionary will be defined for each dataset to delineate valid ranges for the data field(s)...

Budget and FTEs:

----------FUNDED----------

Source	Activity	Fund Type	Budget ($1000s)	FTEs
	Total:		0.00	0.00

----------UNFUNDED----------

	Activity	Fund Type	Budget ($1000s)	FTEs
Year 1:	ADM	One-time	1.03	0.50
Year 2:	ADM	One-time	0.50	0.50
	Total:		1.53	1.00

DRAFT

LEGACY DATA INPUT MINITEMPLATE FOR: FGDC CONTENT STANDARDS FOR DIGITAL GEOSPATIAL METADATA BASED ON JUNE 8, 1994 VERSION COMPLETED AUGUST 30, 1994

Instructions for Use

This minitemplate is to be used for "Legacy" data or data that was produced and completed **before** January 1995. Any data produced after this date must fully comply with the June 8, 1994 "FGDC Content Standards for Digital Geospatial Metadata". This document will be referred to hereafter as "the Standard." (There is a corresponding template available for the entire Standard.) This minitemplate requires only a subset of metadata or data elements from the Standard. The minitemplate should always be used in conjunction with the Standard as it provides useful definitions, a glossary, and other related background information.

This minitemplate is designed to provide National Park Service metadata producers with an easy to use and abbreviated "fill-in-the-blank" style input form with which to produce metadata in Word Perfect format. The minitemplate is not intended to be an automated software application, but a simple tool to use until more sophisticated tools are developed. Other metadata formats (i.e., dBASE files) are acceptable if they include the same or similar data elements that are contained in this minitemplate.

Use of this template assumes that a copy of the data and the data documentation or metatdata will be sent to the Technology Information Center, DSC, Denver for archiving and distribution. The Word Perfect files created with this template will be converted to ASCII text for electronic distribution purposes.

Data elements ending in a colon (:) require input immediately after the colon. Data elements not ending in a colon require input in the fields nested below them. If the input required is unknown or not applicable, please fill it in as UNKNOWN or N/A but do not leave it blank or omit it.

An asterisk preceding a data element indicates that the data element is optional and can be either deleted or filled in.

Some data elements are followed by a short description or explanation in parentheses. It is recommended that these be deleted before converting the file to ASCII and distributing it.

Section 2.5 is about Source data and process used to create the subject data being documented. This section or parts of it may be repeated as many times as necessary.

Section 4.1.2.1.2 of the Standard on map projections can be particularly confusing and long. For this reason, only the most commonly used projections have been included here for the user to choose from and no numbering has been included in the levels nested below 4.1.2.1.2.

The minitemplate numbering and syntax corresponds to the numbering system and syntax of the Standard. The minitemplate numbering appears inconsistent because many whole portions and data elements of the Standard and full template have been eliminated and are not required for internal NPS legacy data documentation purposes. The original Standard numbers and syntax are retained for reference. Please refer to the Standard or full template for additional clarification if necessary.

Sections or data elements may be repeated as often as necessary (e.g. if there is more than one data source repeat the Source Information Section for each source.)

For your convenience three paragraphs have been included as a liability statement. This statement has been reviewed and approved by the Department of Interior Office of the Solicitor, Division of Conservation and Wildlife, Parks and Recreation Branch, August 19, 1994.

LEGACY DATA INPUT MINITEMPLATE OUTLINE

1. IDENTIFICATION INFORMATION

1.1 CITATION (uses Citation Info, section 8)

1. **ORIGINATOR** (name of NPS unit or program that produced the data set):
2. **PUBLICATION DATE:**
4. **DATA SET TITLE:**
5. **EDITION**(VERSION):
8. **PUBLICATION INFORMATION**
    1. **PUBLICATION PLACE:**
    2. **PUBLISHER:**
9. **OTHER CITATION DETAILS:**
10. **ONLINE CITATION** (This field should include files names, and Internet ftp site or server address):

* 11. **LARGER WORK CITATION** (If the data set is part of a larger data collection effort, e.g. American Battlefield Program.)

1. **NPS ORIGINATOR:**
2. **PUBLICATION DATE:**
4. **PROGRAM TITLE:**
5. **EDITION** (Version):
7. **SERIES INFORMATION**
    1. **SERIES NAME:**
    2. **ISSUE IDENTIFICATION:**

1.2 DESCRIPTION

1. **ABSTRACT:** (This general information includes county and state and any other important locational information about the data set. State whether the data type is raster, vector or point.)
2. **PURPOSE:**

1.3 TIME PERIOD OF CONTENT (time period for which the data corresponds to the ground)

1. **TIME PERIOD INFORMATION - DATE(s)/TIME(s)**

1.4 STATUS OF DATA SET

2. **MAINTENANCE AND UPDATE FREQUENCY:**

1.5 SPATIAL DOMAIN

1. **BOUNDING COORDINATES**
    1. **WEST BOUNDING COORDINATE:**
    2. **EAST BOUNDING COORDINATE:**
    3. **NORTH BOUNDING COORDINATE:**
    4. **SOUTH BOUNDING COORDINATE:**

1.6 KEYWORDS

1. **KEYWORD THESAURUS:**
2. **KEYWORD:**

1.9 **POINT OF CONTACT** (uses Contact Info, section 10)

1. **PRIMARY CONTACT**
    1. **CONTACT PERSON OR ORGANIZATION:**
4. **CONTACT ADDRESS:**
5. **CONTACT VOICE TELEPHONE:**
* 6. **CONTACT TDD/TTY TELEPHONE:**
* 7. **CONTACT FACSIMILE TELEPHONE:**
* 8. **CONTACT ELECTRONIC MAIL ADDRESS:**
* 9. **HOURS OF SERVICE:**
* 10. **CONTACT INSTRUCTIONS:**

2. DATA QUALITY INFORMATION

2.1. **ATTRIBUTE ACCURACY**

1. **ATTRIBUTE ACCURACY REPORT:**

2.3 **COMPLETENESS REPORT:**

2.4 **POSITIONAL ACCURACY**

1. **HORIZONTAL POSITIONAL ACCURACY REPORT:**
2. **VERTICAL POSITIONAL ACCURACY REPORT:**

2.5 **SOURCE INFORMATION**

1. **SOURCE CITATION**
    1. **ORIGINATOR:**
    2. **PUBLICATION DATE:**
* 3. **PUBLICATION TIME:**
    4. **TITLE:**
    5. **EDITION** (VERSION):

2. **PROCESS STEP**
    1. **PROCESS DESCRIPTION:**
    3. **PROCESS DATE:**
* 6. **PROCESS CONTACT**
        1. **PRIMARY CONTACT PERSON OR ORGANIZATION:**
        4. **CONTACT ADDRESS:**
        5. **CONTACT VOICE TELEPHONE:**
        6. **CONTACT TDD/TTY TELEPHONE:**
        7. **CONTACT FACSIMILE TELEPHONE:**
        8. **CONTACT ELECTRONIC MAIL ADDRESS:**
        9. **HOURS OF SERVICE:**
        10. **CONTACT INSTRUCTIONS:**

4. SPATIAL REFERENCE INFORMATION

4.1 HORIZONTAL COORDINATE SYSTEM DEFINITION

1. GEOGRAPHIC (LATITUDE, LONGITUDE)

3. GEOGRAPHIC COORDINATE UNITS: (DEGREES, MINUTES, AND DECIMAL SECONDS)

OR

2. PLANAR

1. MAP PROJECTION

1. MAP PROJECTION NAME: (PICK ONE DELETE THE OTHERS)

ALBERS CONICAL EQUAL AREA
LAMBERT AZIMUTHAL EQUAL AREA
MERCATOR
TRANSVERSE MERCATOR
OTHER PROJECTION'S DEFINITION

OR

2. GRID COORDINATE SYSTEM

1. GRID COORDINATE SYSTEM NAME:

2. UNIVERSAL TRANSVERSE MERCATOR

1. UTM ZONE NUMBER:

OR

4. STATE PLANE COORDINATE SYSTEM (SPCS)

1. SPCS ZONE IDENTIFIER:

OR

5. ARC COORDINATE SYSTEM

1. ARC SYSTEM ZONE IDENTIFIER:

OR

6. OTHER GRID SYSTEM'S DEFINITION:

3. LOCAL

1. LOCAL DESCRIPTION:

2. LOCAL GEOREFERENCE INFORMATION:

4. GEODETIC MODEL

1. HORIZONTAL DATUM NAME:

2. ELLIPSOID:

4.2 VERTICAL COORDINATE SYSTEM DEFINITION (ELEVATION)

1. ALTITUDE SYSTEM DEFINITION

1. ALTITUDE DATUM NAME:

3. ALTITUDE DISTANCE UNITS:

2. DEPTH SYSTEM DEFINITION

1. DEPTH DATUM NAME:

3. DEPTH DISTANCE UNITS:

5. ENTITY AND ATTRIBUTE INFORMATION

5.1 DESCRIPTION

1. CLASSIFICATION SCHEME, TYPE, OR NAME (list any known classes, categories, and values)

6. DISTRIBUTION INFORMATION

6.1 DISTRIBUTOR

1. **CONTACT PERSON PRIMARY** (uses Contact Info, section 10)
    1. **PRIMARY CONTACT PERSON OR ORGANIZATION:**
    4. **CONTACT ADDRESS:**
    5. **CONTACT VOICE TELEPHONE:**
    6. **CONTACT TDD/TTY TELEPHONE:**
    7. **CONTACT FACSIMILE TELEPHONE:**
    8. **CONTACT ELECTRONIC MAIL ADDRESS:**
    9. **HOURS OF SERVICE:**
    10. **CONTACT INSTRUCTIONS:**

6.3 DISTRIBUTION LIABILITY:

The National Park Service shall not be held liable for improper or incorrect use of the data described and/or contained herein. These data and related graphics ("GIF" format files) are not legal documents and are not intended to be used as such.

The information contained in these data is dynamic and may change over time. The data are not better than the original sources from which they were derived. It is the responsibility of the data user to use the data appropriately and consistent within the limitations of geospatial data in general and these data in particular. The related graphics are intended to aid the data user in acquiring relevant data; it is not appropriate to use the related graphics as data.

The National Park Service gives no warranty, expressed or implied, as to the accuracy, reliability, or completeness of these data. It is strongly recommended that these data are directly acquired from an NPS server and not indirectly through other sources which may have changed the data in some way. Although these data have been processed successfully on a computer system at the National Park Service, no warranty expressed or implied is made regarding the utility of the data on another system or for general or scientific purposes, nor shall the act of distribution constitute any such warranty. This disclaimer applies both to individual use of the data and aggregate use with other data.

6.4 STANDARD ORDER PROCESS

2. **DIGITAL FORM**
    1. **DIGITAL TRANSFER INFORMATION**
        1. **FORMAT NAME:** (GRASS 4.0, ARC/INFO 7, BIL, SDTS)
        6. **FILE DECOMPRESSION TECHNIQUE:**
        7. **TRANSFER SIZE:**
    2. **DIGITAL TRANSFER OPTION**
        1. **ONLINE OPTIONS**
            1. **COMPUTER CONTACT INFORMATION**
                1. **NETWORK ADDRESS**
                    1. **NETWORK RESOURCE NAME:**

7. METADATA REFERENCE INFORMATION

7.1 METADATA DATE:

7.4 METADATA CONTACT (uses Contact Info, section 10)

1. CONTACT PERSON PRIMARY

1. PRIMARY CONTACT PERSON OR ORGANIZATION:

4. CONTACT ADDRESS:

5. CONTACT VOICE TELEPHONE:

* **6. CONTACT TDD/TTY TELEPHONE:**

* **7. CONTACT FACSIMILE TELEPHONE:**

* **8. CONTACT ELECTRONIC MAIL ADDRESS:**

Working With Legacy Data

Baseline Water Quality Data Inventory and Analysis Project

NPS Water Resources Division and Inventory & Monitoring Program

Introduction

One component of the National Park Service (NPS) Servicewide Inventory and Monitoring (I&M) Program is the Baseline Water Quality Data Inventory and Analysis (BWQ) Project. This cooperative effort, overseen by the NPS Water Resources Division, seeks to characterize baseline surface-water quality at national park units containing significant water resources. Tapping into several existing Environmental Protection Agency (EPA) databases, particularly STORET, the national water quality database, BWQ Reports are being prepared for parks. These reports will provide each park with a complete inventory of all surface water quality data; descriptive statistics and graphics characterizing annual, seasonal, and period-of-record central tendencies and trends; and comparisons of park water quality data with relevant EPA national water quality screening criteria and NPS-75 "Level I" water quality parameters. The entire report (text, tables, and graphics) and all databases (water quality parameter data; water quality station, water gage, industrial facility discharge, drinking intake, and water impoundment locations; and other data) are provided in both analog and digital format to encourage additional analysis and incorporation into park geographic information systems.

The goal of the BWQ Project is to provide descriptive water quality information in a format usable for park planning purposes (e.g., Water Resources Management Plans, Resource Management Plans, and General Management Plans). The BWQ Reports are designed to characterize baseline water quality rather than assess specific water quality problems at a park. This is consistent with the Servicewide I & M Program's goal of obtaining basic, "Level I", water quality parameters for key waterbodies at each park. The reports are intended to be used as reference documents to help design new goal-driven water quality monitoring programs rather than as conclusive evidence of previous or existing water quality problems.

In developing this project, a number of data management challenges were confronted due to the quantity of data and the use of legacy data collected and maintained by non-NPS personnel. The purpose of this exposition is to provide some brief background on the databases used in producing a BWQ Report and then share the Water Resources Division's experiences in harnessing these databases, producing the reports, and managing the data.

Data Sources

The EPA maintains many mainframe data systems related to national water resources. Six of these data systems were used in the production of BWQ Reports: (1) STORage and RETrieval (STORET) national water quality database; (2) Industrial Facilities Discharge; (3) Drinking Water Supplies; (4) Water Gages; (5) Water Impoundments; and (6) River Reach File, Ver. 3.

STORET is the national water quality data repository. Water quality data is entered in STORET by public agencies (federal, state, or local) that collect water samples and/or perform laboratory analysis. Currently, there are over 800,000 active and inactive sampling stations and more than 150 million observations covering in excess of 13,000 water quality parameters entered in STORET. The earliest data dates back to the turn of the century. All U.S. Geological Survey (USGS) water quality data is in STORET. The STORET water quality database is not exhaustive; it only contains water quality data that public agencies have taken the time to enter.

The data within the Industrial Facilities Discharge (IFD) database are extracted from the EPA's Permit Compliance System (PCS). IFD contains the facility locations of all industrial and municipal dischargers which require a National Pollutant Discharge Elimination System permit to operate. Over 7,100 municipal, federal, and industrial facilities discharging into the waters of the United States are tracked by PCS and IFD.

The EPA Drinking Water Supplies database identifies locations of drinking water supply intakes. This database contains data for 850 supplies which serve more than 25,000 people, and 6,800 supplies serving between 1,000 and 25,000 people.

The Water Gages (including stream, lake, estuary, well, spring, climate, or other) database originates primarily with the USGS and copies are maintained on the EPA mainframe computer for ease of integration with other EPA national data systems. Although other agency's water gages, as well as some artificial gages, may appear in database, the vast majority of entries are stream gages belonging to the USGS. The database contains approximately 36,000 records for both active and inactive gaging stations throughout the country.

The Water Impoundment database was originally compiled by the U.S. Army Corps of Engineers in response to a Congressional inquiry on dam safety hazards. The EPA subsequently modified the database for use in water quality investigations. Of the 68,155 dams in the database, 2,125 are considered large (impounding 10,000 acre feet or more at normal pool volume). While not containing information for every dam in the country, the Water Impoundment database does include entries for 66,030 smaller dams.

The River Reach File (RF3) data system is a hydrologic database of surface water features across the U.S. (excluding, at present, Idaho, Oregon and Washington, which currently operate a different system - although these data are expected to be converted to RF3 soon, and Alaska). RF3 was created primarily from 1:100,000 scale USGS Digital Line Graph data. RF3 is made up of over 3,000,000 individual "reaches". A reach is generally defined as a portion of surface water between two confluences. The linework underlying RF3 contains over 95,000,000 coordinate points. RF3 is designed to facilitate hydrologic routing, identifying upstream and downstream elements, and specifying the exact location of any point on a stream network. RF3 data exists as a series of traces with associated attributes.

Working With Legacy Data

Harnessing the data sources described above to characterize baseline water quality conditions at a park was challenging. The primary data source for the Project reports is STORET, the EPA's national water quality database. STORET's greatest strength is also its greatest weakness: the system is open to any public entity for storing water quality data. As a consequence, STORET contains a phenomenal amount of water quality data. Unfortunately, due to its "openness", STORET must be considered a "user-beware" water quality database system. While there has been some rudimentary edit (bounds) checking of data entered in STORET since November 1983, users are basically free to enter their own data. The EPA does not verify, validate, or certify any data in STORET. Beyond data entry errors, the possibility of inaccurate data entering the system due to inappropriate measurement techniques, sample mistreatment, and other reasons is a serious concern. Not even an agency like the USGS, with its rigorous quality assurance/quality control procedures, is beyond uploading erroneous observations to STORET. In the course of producing BWQ Reports, the NPS Water Resources Division has, for example, found USGS-collected pH measurements in excess of 14 (e.g. a pH of 200 in Yellowstone National Park). Consequently, before rendering any substantiative decisions based on STORET data, it behooves the user to investigate who collected the data and what analytical methodologies were employed.

In using STORET data to prepare Baseline Water Quality reports, a number of automated screening procedures were employed to cope with both the quantity and quality of the water quality data. The automated screening criteria dealing with the "quality" of the water quality data serve as a means to validate, at least in a limited sense, the retrieved data. The automated "quantity" screening criteria were

intended to limit the scope of the effort, provide the most meaningful graphics and statistics, and avoid "data drowning".

Screening Methodologies and Procedures

Three general groups of screening criteria were applied to the data downloaded from STORET: (1) screens that apply to stations; (2) screens that apply to certain parameters at stations; and/or (3) screens that apply only to particular observations of parameters at stations. The first two screening criteria deal with the quantity of data; the last screening criteria group focuses on validating the quality of the data. Since the quantity of the data were site and report specific, the quantity screens that were employed will be described briefly following the quality screens.

Quality Screens

One of the first screens employed were the STORET Edit Criteria. As mentioned previously, STORET is a "user-beware" data system. As the EPA doesn't certify any data in STORET, public agencies enter and are responsible for the quality of their own data. Only data entered since November 1983 have been subjected to any realistic edit/bounds checking. Agencies entering data since this date can override the edit/bounds checking, if desired, when entering water quality data in STORET. USGS water quality data is entered into STORET without any EPA edit/bounds checking to ensure data integrity between USGS water quality data systems and STORET. In order to eliminate as much "bad" data (observations) as possible from the BWQ Reports, all water quality data downloaded from STORET were subjected to automatic edit/bounds checking for the 190 most common water quality parameters using the STORET Edit Criteria. Some examples of the STORET Edit Criteria are given in the table below. These criteria were established by water quality professionals to encompass the most likely range of values encountered in natural systems. Observations falling outside the STORET Edit Criteria were identified and then retained or discarded from the park's water quality database (and by consequence the report tables and graphs) based on whether the value was judged as being in the realm of possibility according to the Water Resources Division's professional judgment.

Example STORET Edit Criteria

STORET Code	STORET Parameter Description	High Value	Low Value
00010	TEMPERATURE, WATER (°C)	37.0	-2.0
00020	TEMPERATURE, AIR (°C)	52.0	-40.0
00095	SPECIFIC CONDUCTANCE (UMHOS/CM @ 25C)	60000.0	1.0
00300	OXYGEN, DISSOLVED (MG/L)	30.0	0.0
00400	PH (STANDARD UNITS)	12.0	0.9
00410	ALKALINITY, TOTAL (MG/L AS CACO3)	1000.0	0.0
00600	NITROGEN, TOTAL (MG/L AS N)	100.0	0.0
00665	PHOSPHORUS, TOTAL (MG/L AS P)	10.0	0.0
39516	PCBS IN WHOLE WATER SAMPLE (UG/L)	20.0	0.0

Another screen that dealt with individual observations of parameters at stations was the Date Screen. Every water quality observation in STORET must have a sampling date associated with it. Unfortunately, STORET does not prevent users from entering incorrect dates. Consequently, any

downloaded water quality observation with an incorrect and/or suspect date (e.g. a month greater than 12; a day greater than 31; or a sample date later than the STORET retrieval date) was discarded.

STORET enables the agency collecting water quality samples to provide a qualifying remark code for each parameter observation. These remarks provide additional information about the measured or observed value entered into STORET. Using each observation's remark code, the observation was either eliminated or modified and included in analyses. Data that were eliminated carried remark codes that indicated either less confidence in the observed values or that the data were for nominal or categorical parameters that didn't lend themselves to the statistical analyses and graphics contained in the report. Observations containing these remark codes comprised a very small fraction of the data. Although statistical analyses weren't undertaken on any eliminated data, all water quality observations, regardless of remark code, are included on disk accompanying each park's report. Water quality observations that were modified before inclusion in report tables and graphics included those observations carrying a remark code indicating the value was recorded as below the detection limit. These values were halved prior to inclusion in period of record, annual, and seasonal descriptive statistics and graphics. The common water quality data analysis convention for these remark codes is to use one-half the detection limit in statistical analyses. Although this is a somewhat defensible treatment of observations below the detection limit, the statistics that may be computed using these halved values may not be defensible. Consequently, any statistics in inventory, annual, or seasonal report tables that were computed using 50% or more below detection limit observations are flagged. This provides the report reader with some caution in using and interpreting these results. Water quality data included on disk with the report include the original entry (detection limit) and remark code.

Sometimes data entered in STORET represent something other than a single measurement at one location at one point in time. These samples are typically referred to as composite samples due to the fact that they vary temporally and/or spatially. Consequently, the observation entered into STORET for composite data is typically a computed value that summarizes the data over time and/or space. These data complicate statistical and graphical analyses and must be handled separately. Such treatment was beyond the scope of the BWQ Project; although composite values typically represent only a fraction of STORET observations. The composite type screen eliminated all composite observations from statistical and graphical analyses, except those with a composite type code of "A" (average) that have a one day or less sampling period. All water quality observations, regardless of composite type code, are included on disk accompanying each park's report.

Quantity Screens

In addition to the screens that validated the data used to produce BWQ Reports, several screens were employed to limit the quantity of data retrieved and analyzed in the reports. These screens define the project's scope, and although these screens are project-specific, in any data analysis effort it is important that such screens be explicitly defined.

STORET contains data from a wide variety of stations classified by the type of waterbody in which the samples were collected. As the BWQ Project's purpose was to inventory and analyze surface-water quality, the following surface-water station types were retrieved: (1) Stream; (2) Canal; (3) Lake; (4) Reservoir; (5) Spring; (6) Fresh Water Wetland; (7) Salt Water Wetland; (8) Estuary; and (9) Ocean. Stations not of these types were excluded from the database and report.

Water quality samples can be taken in a variety of aqueous media. Water quality data were retrieved from STORET only if the media were WATER or VERT (vertically integrated). WATER and VERT samples comprise the overwhelming majority of samples in STORET.

Nearly all water quality parameters associated with each station type listed above were retrieved. The only exceptions to this were the exclusion of most STORET administrative parameters and other data not suitable for statistical analysis. These data are often germane only to the collecting agency. These parameters are included on disk with each park's report.

In producing BWQ Reports for parks, many graphics and statistical tables were produced. To limit the number of graphics and tables, sliding criteria were developed to only generate graphics (time series, seasonal, and annual plots) and tables (seasonal and annual) when there was a sufficient quantity of data to warrant their production. Additionally, to focus plotting resources on water quality parameters, plots were never produced for water temperature, stage, discharge, and meteorological measurements. Without this restriction, most plots would have been for these parameters.

Data Management

BWQ Reports have now been completed for approximately 50 national park units. In the course of producing these reports, tremendous quantities of data have been compiled. Datasets for each park include: (1) water quality parameter data retrieved from STORET; (2) water quality station locations from STORET; (3) permitted discharges from the IFD database; (4) water gages from the EPA's GAGES database; (5) drinking water supply intakes from the EPA's DRINKS database; (6) water impoundments from the EPA's Water Impoundment database; (7) River Reach File hydrography; (8) all the statistical tables, graphics, and report sections; and (9) several digital (GIS-compatible) cartographic coverages including, at minimum, park boundary, states, counties, hydrologic units, roads, interstates, cities, places, quadrangle outlines, study area, and hydrography. The quantity of data for each park averages approximately 9 megabytes.

Data management is simplified in that the original datasets enumerated as one through seven above are owned and maintained by the EPA. A limited printing run of 10 BWQ Reports are produced for any given park. Two copies of the report are sent to the park; two copies remain with the Water Resources Division; three copies are sent to the appropriate System Support Office (for the I & M, GIS, and Water Resources Program Managers, respectively); and one copy is kept by the Servicewide I & M Program. The remaining two copies are sent to the NPS's Technical Information Center and the Department of Commerce's National Technical Information Service to provide distribution, for a fee, to anyone else who desires a copy of a particular park's report. A series of diskettes are included with each report. Using these diskettes, anyone can reprint all of the report (except the cover). Additionally, all the databases enumerated above are contained in compressed (ZIP) format on the diskettes included with the report, except the digital cartographic coverages which are only provided upon request. All the databases sent on disk and used in report production are documented in each report's appendix.

Although this strategy decentralizes data management (and storage), the source document and data files must still be managed. These data are kept on computer in Fort Collins and backed up whenever reports are completed for parks. Eventually the data (and by consequence the entire report - including the digital cartographic coverages) for a park will be available on an NPS web or FTP site.

Contact

For additional information on the Baseline Water Quality Data Inventory and Analysis Project, contact Dean Tucker, NPS Water Resources Division, 970-225-3516.

BYE Backup Program and Software*

The BYE routine is a DOS batch file implementation of a backup system developed at Shenandoah National Park. The original BYE routine was written for and by the Shenandoah I&M Unit to solve two problems: a) not always knowing what new or modified files needed to be backed up each day; and b) having to remember all of the available backup options and commands. Semi-automatic daily and/or weekly backups can be made via typing "BYE" at a DOS prompt with the included software. One may optionally choose to search all of the files on the drive for daily backups or just the .ZIP files for weekly backups. The included software includes a shareware version of PKZIP (including WINZIP for Windows) and the BYE program files for use with both DOS and Windows. Anyone using the shareware programs for an extended time should register them with the respective vendors.

In general, the BYE applications use PKZip to assemble an editable list of new and modified files which are displayed using the MS-DOS EDIT program. The file list shows the full path of files that are new or modified since a user specified date (files created or changed on a specified date according to the system clock). Files can be removed from the backup list by scrolling to the file to be deleted and typing Ctrl-Y or by highlighting with the mouse and pressing Delete. Help for using the EDIT program is available by typing F1. Once the backup list has been edited, select Exit from the File menu. If changes to the list were made, the user is prompted to Save them. Next, the BYE application searches for a .ZIP file with the same file name and presents alternate backup options if one is encountered. Daily and/or weekly backup .ZIP files may then be copied to a diskette, tape, or other media (or re-ZIPped to multiple diskettes if necessary) for safe storage.

All of the BYE batch application functions discussed above are available through current versions of MS DOS and PKZip. The functions may be applied manually if desired, and the included batch files may be edited and adapted to customize individual backup strategies.

Instructions for installing the BYE backup programs and printouts of the BYE-DATE.BAT and BYE-WEEK.BAT batch files are included below*. Although the BYE programs are intended as customizable examples without comprehensive support, some technical assistance may be obtained from:

Joe Gregson (Joe_Gregson@nps.gov)
Inventory and Monitoring Program
1201 Oak Ridge Drive, Suite 350
Fort Collins, CO 80525

Phone: 970-225-3559 Fax:970-225-3585

* Note that the BYE programs were not programmed for and have not been tested with Windows 95 or Windows NT.

BYE Backup Program Installation Instructions

1. In the root directory of the hard disk drive (e.g., C:\>), make a directory called BACKUP.

    -> C: (substitute another drive letter if necessary)
    -> cd\
    -> md backup

2. Copy the program files to C:\BACKUP or other directory (see note below)
    (files: BYE-PROG.EXE, PKZSHARE.EXE,BYE-READ.TXT).

Note: The PKZIP files may be put in an existing directory such as BATCH or UTILITY if one already exists, but they must reside in a directory in the system's default PATH.

    -> copy A:*.* C:\BACKUP*.* (substitute other drive letters if necessary)

3. Change to the \BACKUP or other directory, and unzip the self-extracting files.

    -> cd \BACKUP
    -> BYE-PROG.EXE
    -> PKZSHARE.EXE

4. After the files are copied to the \BACKUP directory, edit the AUTOEXEC.BAT file in the root directory and add C:\BACKUP to the PATH statement. Whenever the PATH statement is edited, the computer must be rebooted by pressing the Reset button, pressing <Ctrl><Alt><Delete> at the same time, or turning the machine off and back on.

5. Enter BYE or BYE-READ at the DOS prompt to get program instructions.

Installing BYE for MS Windows.

1. Install the program for DOS as above.

2. Start Windows and open the program group on the desk top where you want the BYE icon.

3. Click on File -> New -> Program Item -> OK.

4. Type BACKUP in the Description box.

5. Type in the PATH (e.g., C:\BACKUP\) and BYE-WIN.BAT in the Command Line box.

6. Type C:\BACKUP in the Working Directory box.

7. Click on Change Icon and choose the BYE icon or Browse for one in the Program Manager.

8. Click on OK and double check your entries; click OK again.

9. Try starting BYE with the new icon.

BYE-DATE.BAT Batch File Printout

```
@echo off
rem BYE-DATE.BAT
rem Syntax: BYE_DATE mmddyy  (mm= number of month, dd=day of month, yy=last two digits of
year)
rem The program uses PKZip to create a list of files with dates equal or greater than the date entered
rem on the command line.  If no date is specified, an usage message is generated.
rem After the dated file list is created, the DOS Edit program is called to edit the list.
rem After the list is edited, PKZip creates a backup .ZIP file,
rem if one of a similar name does not already exits.
rem Caveat1:  PKZip and the DOS Edit programs must be in the default path.
rem Caveat2:  The directory C:\BACKUP must exist.

rem Check for missing date command usage; issue a message and end if detected.
if not "%1"=="" goto List
cls
echo.
echo This program backs up files from a specified date forward to today.
echo.
echo.
echo Usage: BYE-DATE mmddyy     mm is numberical month (e.g., 01 to 12)
echo                        dd is day of month     (e.g., 01 to 31)
echo                        yy is last two digits of year
echo.
pause
goto end

:List
rem Switch to the root directory, scan the drive for dated files, and save file list to a file named LIST.
cd\
pkzip -t%1 -r -@backup/bk%1.lst dummy.zip

rem Clear screen, display results of file search and edit instructions on screen, and pause for reading.
cls
echo.
echo.
echo A list of files with dates from %1 has been created.
echo.
echo Delete unwanted backup files from the list when the DOS editor starts.
echo                (Type F1 in the DOS Editor for Help.)
echo.
pause

rem Call DOS Edit program to edit the file list.
edit backup\bk%1.lst

rem Check if a .ZIP file for this date already exists; if so, skip PKZip and
rem go to the choice menu. Otherwise make a new backup .ZIP file from the
rem edited BK%1.LST file, and go to ending message of the batch file.

if exist backup\bk%1.zip goto choicemenu
goto choice2
```

```
:choicemenu
cls
echo.
echo The file *** BK%1.ZIP *** already exists!
echo.
echo                PRE-EXISTING FILE OPTIONS
echo.
echo    1. Update Previous Backup with New and Modified Files
echo    2. Overwrite the Previous Backup
echo    3. Cancel Backup Operation
echo.
choice /c:123   Choose an Option:
if errorlevel 3 goto choice3
if errorlevel 2 goto choice2
if errorlevel 1 goto choice1

:choice1
copy backup\bk%1.zip backup\bk%1.bak
pkzip -u backup\bk%1.zip @backup\bk%1.lst
goto message

:choice2
if exist backup\bk%1.zip copy backup\bk%1.zip backup\bk%1.bak
if exist backup\bk%1.zip del backup\bk%1.zip
pkzip backup\bk%1.zip @backup\bk%1.lst
goto message

:choice3
echo.
echo.
echo To run PKZip manually type: PKZIP PATH\FILENAME.ZIP @BACKUP\BK%1.LST
echo.
echo    *****   Type PKZIP without parameters to get help.   *****
echo.
echo.
goto end

:message
cls
echo.
echo The backup .ZIP file has been written to C:\BACKUP\BK%1.ZIP
echo NOTE: Any pre-existing BK%1.ZIP was copied to BK%1.BAK.
echo.
call dir c:\backup\bk%1.zip
echo.
echo Check the size of BK%1.ZIP to determine the number of diskettes
echo (or other media) needed to save the backup.
echo.
echo To do a multiple diskette backup with PKZip use this syntax:
echo.
echo          PKZIP -& A:BK%1.ZIP @LIST
echo.
echo    *****   Type PKZIP without parameters to get help.   *****
:end
```

BYE-WEEK.BAT Batch File Printout

```
@echo off
rem BYE-WEEK.BAT
rem Syntax: BYE_WEEK mmddyy  (mm= number of month, dd=day of month, yy=last two
digits of year)
rem The program uses PKZip to create a list of .ZIP files with dates equal or
rem greater than the date entered on the command line.
rem If no date is specified, the usage message is generated.
rem After the dated file list is created, the DOS Edit program is called to
rem edit the list. After the list is edited, PKZip creates a backup .ZIP file
rem if one of a similar name does not already exist.

rem Caveat1:  PKZip and the DOS Edit programs must be in the default path.
rem Caveat2:  The directory C:\BACKUP must exist.

rem Check for missing date command usage; issue a message and end if detected.
if not "%1"=="" goto List
cls
echo.
echo This program backs up daily .ZIP files from the specified date forward
echo to the current system date.
echo.
echo.
echo Usage: BYE-WEEK mmddyy     mm is numberical month (e.g., 01 to 12)
echo                    dd is day of month     (e.g., 01 to 31)
echo                    yy is last two digits of year
echo.
goto end

:List
rem Switch to the root directory, scan the backup dir for .ZIP files,
rem and save file list to a file named LIST.
cd\
pkzip -t%1 -@list list c:\backup\*.zip
cls
echo.
echo.
echo.
echo A list of BACKUP .ZIP files modified since %1 has been created.
echo.
echo Remove unwanted backup files when the DOS editor starts.
echo.          (Type F1 in the DOS Editor for Help.)
echo.
pause
edit list
```

```
rem Check to see if a previous backup file with the same name exists.
rem If the file already exists switch execution to gosub1;
rem otherwise, zip up the new file and post instructive messages.
if exist backup\wk%1.zip goto gosub1
pkzip backup\wk%1.zip @list
cls
echo.
echo.
echo The backup .ZIP file has been written to C:\BACKUP\WK%1.ZIP
echo.
call dir c:\backup\wk%1.zip
echo.
echo Check the size of WK%1.ZIP to determine the number of diskettes
echo (or other media) needed to save the backup.
echo.
echo To do a multiple diskette backup with PKZip use this syntax:
echo.
echo          PKZIP -& A:WK%1.ZIP @LIST
echo.
echo.
goto end

rem The suggested file name already exists, so end job with error message.
:gosub1
cls
echo.
echo.
echo.
echo The file *** WK%1.ZIP *** already exists!
echo.
echo To run PKZIP manually use: PKZIP PATH\FILENAME.ZIP @LIST
echo.
echo.
:end
```

Inventory and Monitoring Program Contact Information

Gary Williams	I&M Program Manager	970-225-3539
Steve Fancy	Monitoring Specialist	970-225-3571
Joe Gregson	Information Management Specialist	970-225-3559

Inventory and Monitoring Program
Natural Resource Information Division
National Park Service
1201 Oak Ridge Drive, Suite 350
Fort Collins, CO 80525