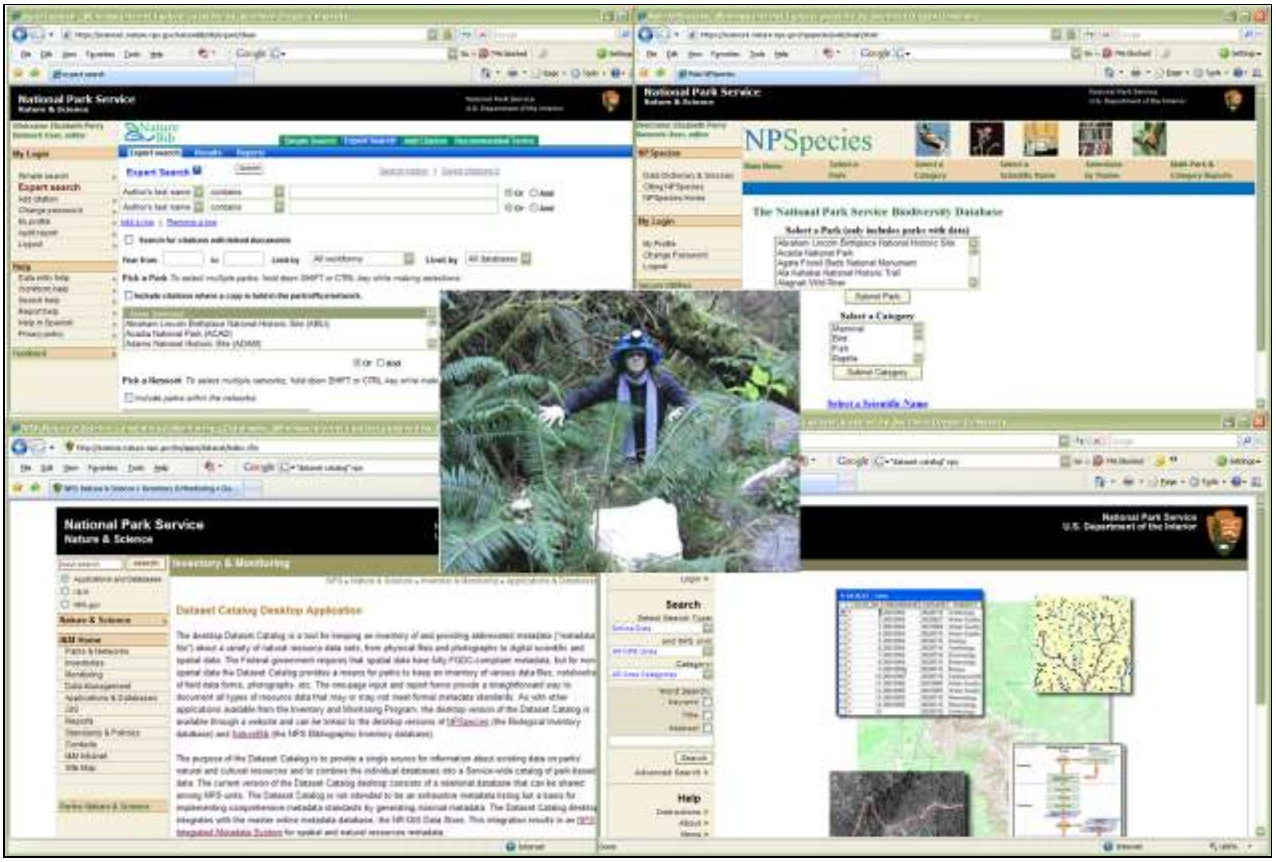
National Park Service
U.S. Department of the Interior

Natural Resource Program Center

Klamath Network Data Mining

2001-2007

Natural Resource Report NPS/KLMN/NRR—2009/156



ON THE COVER
Collage of databases related to Klamath Network data mining, and, at center, former data miner Beverly Shaw “discovering” files at Redwood National Park. Collage by Elizabeth E. Perry; center photograph by Laura Bridy

Klamath Network Data Mining

2001-2007

Natural Resource Report NPS/KLMN/NRR—2009/156

Elizabeth E. Perry
National Park Service
Klamath Network Office
1250 Siskiyou Boulevard
Ashland, Oregon 97520-5011

Robert E. Truitt
National Park Service
Mojave Network Office
Lake Mead National Recreation Area
601 Nevada Way
Boulder City, NV 89005

October 2009

U.S. Department of the Interior
National Park Service
Natural Resource Program Center
Fort Collins, Colorado

The National Park Service, Natural Resource Program Center publishes a range of reports that address natural resource topics of interest and applicability to a broad audience in the National Park Service and others in natural resource management, including scientists, conservation and environmental constituencies, and the public.

The Natural Resource Report Series is used to disseminate high-priority, current natural resource management information with managerial application. The series targets a general, diverse audience, and may contain NPS policy considerations or address sensitive issues of management applicability.

All manuscripts in the series receive the appropriate level of peer review to ensure that the information is scientifically credible, technically accurate, appropriately written for the intended audience, and designed and published in a professional manner. This report received informal peer review by subject-matter experts who were not directly involved in the collection, analysis, or reporting of the data.

Views, statements, findings, conclusions, recommendations, and data in this report are those of the author(s) and do not necessarily reflect views and policies of the National Park Service, U.S. Department of the Interior. Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the National Park Service.

This report is available from the Klamath Inventory and Monitoring Network's web site: (http://www.nature.nps.gov/im/units/KLMN) and the Natural Resource Publications Management web site (http://www.nature.nps.gov/publications/NRPM).

Please cite this publication as:

Perry, E. E., and R. E. Truitt. 2009. Klamath Network data mining: 2001-2007. Natural Resource Report NPS/KLMN/NRR—2009/156. National Park Service, Fort Collins, Colorado.

NPS 963/100531, October 2009

Contents

Figures and Tables

Appendixes

Page

Acknowledgements

Many people have provided tremendous help and support to the Klamath Network's data mining effort. Daniel Sarr, Klamath Network Coordinator, was instrumental in creating and maintaining the program. Sean Mohren, current Klamath Network Data Manager, provided guidance, support, and oversight throughout the project. All of the park staff were incredibly helpful with the data mining project, especially the Point of Contact at each park: Michael Murray (CRLA), Nancy Nordensten (LAVO), Barney Stoffel (LABE), David Larson (LABE), Deana DeWire (ORCA), John Roth (ORCA), Bow O'Barr (REDW), Leonel Arguello (REDW), Windy Bunn (WHIS), and Jennifer Gibson (WHIS). Over the years of the project, the Klamath Network Data Mining Team consisted of Laura Bridy, Tim Shepherd, Beverly Shaw, Carol Powell, Sean Smith, Hanna Waterstrat, Jules Filipski, Rebecca Miller, and Susan Colbert, all of whom relished capturing the most remote and obscure information. We would also like to extend a big thank you to the Inventory and Monitoring Program's database development team in Fort Collins, who patiently helped us with the use of NatureBib, NPSpecies, and Dataset Catalog. In particular, Allison Loar, Wendy Schumacher, Simon Kingston, Lisa Nelson, Kathy Dratch, Willene Hendon, Mark Wotawa, and Chris Dietrich answered our countless questions and continuously worked with us to make the data mining effort successful.

Introduction

Building on past recommendations and studies calling for park management of natural resources to be more scientifically based, the National Park Service (NPS) implemented the Natural Resources Inventory and Monitoring (I&M) Program. This program meets the demand for scientific information to guide resource management decisions (Leopold et al. 1963 and Robbins et al. 1963), as laid out in past NPS initiatives including the Natural Resource Challenge (NPS 1999) and NPS-75 (NPS 1992). In order for park staff to access natural resource information to make sound management decisions, the information needs to be structured, accessible, in standardized digital formats, and comprehensive. This allows park staff and researchers to efficiently review what has already been accomplished, build upon past studies, and adapt future management needs, without reinventing the wheel. The Klamath Network (KLMN) supported a data mining team to glean past natural resource-associated information from each park, to consolidate it for park management needs, and to help support I&M vital signs efforts.

There were a number of driving reasons why the Network decided to develop and implement data mining across the network of parks. The predominate reasons and some of the underlying drivers were the guidance on gathering information on the 12 Basic Inventories and the shift to treating data and information as a resource. Out of this, the I&M program developed the I&M applications (the databases) and recognized the need to populate them with information.

From a Network perspective, several reasons deemed it imperative that the KLMN assist parks in implementing the I&M applications. If the parks were expected to use the newly developed national applications, then those applications would require a minimum effort to populate them with relevant data for use. Park staff did not have the time, money, or staff to accomplish these tasks.

The KLMN Data Mining Team (DMT) sought out information tucked away in filing cabinets, boxed up in storage, written as digital files, and kept at outside institutions. The focus of gathering the information was centered around the 12 Basic Inventories, which is a set of core natural resource inventory data needed to effectively manage a park's resources (developed by the I&M Program). This set of 12 distinct categories, first identified in Appendix A of the Natural Resource Inventory and Monitoring Guidelines (NPS 1992), and additionally supported through NPS Inventory and Monitoring web sites (NPS 2003), includes the following abiotic and biotic ecosystem components:

1) Natural Resource Bibliographies
2) Comprehensive Vascular and Vertebrate Species List
3) Species Occurrence and Distribution
4) Vegetation Inventory
5) Base Cartography Data
6) Soils Resources Inventory
7) Geology Resources Inventory
8) Water Bodies Location and Classification
9) Baseline Water Quality Data
10) Air Quality Data

11) Air Quality Related Values
12) Climate Inventory

Each of the 32 NPS I&M networks was tasked with discovering more information on these 12 categories and synthesizing the information (new and historic) currently being held in the parks. To accomplish this goal, the Klamath Network undertook many field and library inventory projects, including the development of the DMT. All the information that the DMT gathered pertained to references, datasets, and voucher specimens of these 12 Basic Inventories.

NPS staff has recognized that the preponderance of research data and information collected on public lands over the past century is as valuable as the resource itself. This recognition helps with avoiding redundancy of past research, reducing needless collection of additional specimens, and potentially predicting status and trends of park resources sooner by incorporating historic and legacy data. These data require good data management, including documentation, long-term archiving, and accessibility. Data mining efforts to capture, document, and preserve this information increase sound data management, as well as fulfill the NPS mission, to preserve "unimpaired the natural and cultural resources and the values of the national park system for the enjoyment, education, and inspiration of this and future generations" (National Park Service Organic Act, 16 U.S.C.1.).

KLMN data mining accomplished the objective of searching out and documenting the existing natural resource information located in the KLMN park units. Not only does this help the Network and other researchers avoid unnecessary redundancy and provide them the opportunity to build upon the research of others, but it also helps preserve institutional memory, passing the information on to present and future park staff and researchers inside and outside the organization, in a searchable and accessible manner.

The long-term survival of information depends upon documenting what has been done (Smith et al. 2005). Without documenting the who/what/where/when/how of a dataset, a substantial amount of knowledge can be lost. Figure 1A displays this loss over time when typical information management practices are followed. However, data details (and thus, institutional memory) are retained over time when proper data documentation is followed (Figure 1B).

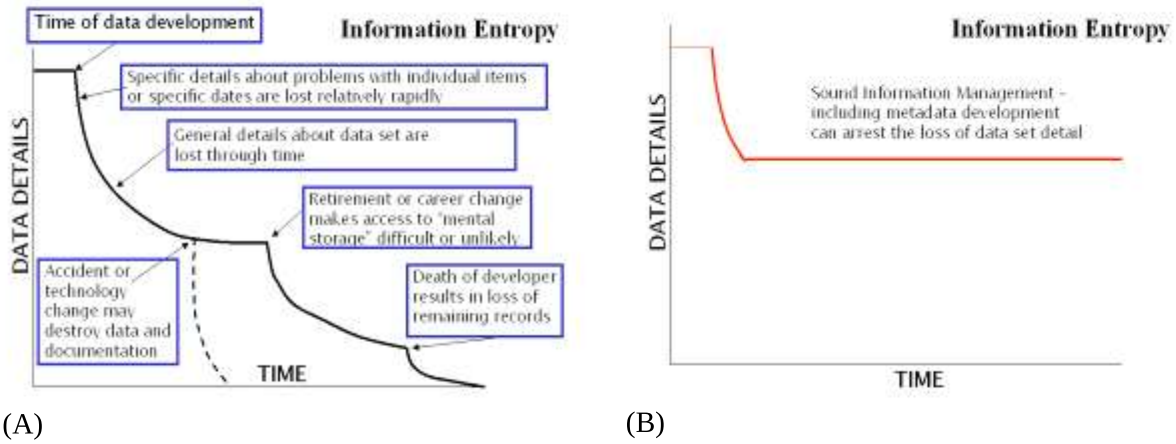


Figure 1. Information entropy (A) without metadata and (B) with metadata.

At the start of this project, the KLMN was one of the first networks to perform data mining and embarked upon it at a scale and funding level unmatched amongst the networks. The data mining project was funded from fiscal year 2001 to 2007 with KLMN monitoring and water quality appropriations, and through agreements with Southern Oregon University, Oregon State University, and the United States Geological Survey.

Over the course of the project, the data miners captured information about the 12 Basic Inventories at each of the KLMN's six parks: Crater Lake National Park, Lassen Volcanic National Park, Lava Beds National Monument, Oregon Caves National Monument, Redwood National and State Parks, and Whiskeytown National Recreation Area. The information found on the 12 Basic Inventories was diverse, but the methods for capturing relevant details were uniform across the six parks, in order to aid in metadata documentation, relevance, standardization, and longevity.

Being the first I&M network to conduct data mining on this scale, the KLMN found some unexpected short-cuts and challenges. The purpose of this report is to share Klamath Network data mining methods, accomplishments, lessons learned, and suggestions for the future.

Methods

Initially, the guidance from the Washington Support Office (WASO) of the I&M Program was to document bibliographic information and at least 90% of the evidence of vertebrate and vascular plant species for each park. Not long into the KLMN data mining effort, it became evident that this task was too large to accomplish in one session in a standardized and efficient manner. Therefore, the decision was made to split the work into three phases, with voucher mining occuring first, then the highest priority information captured in Phase I and the remaining data captured in Phase II.

The voucher mining work was completed before starting Phase I. In Phase I, data miners focused on capturing hardcopy references pertaining to vascular plants and vertebrates. A secondary focus during this phase was hardcopy information about water quality. During Phase II, a team of data miners finished cataloging the remaining Phase I information and then expanded the focus to catalog raw non-GIS (i.e., tabular) data relevant to the 12 Basic Inventories. A secondary focus was to document hardcopy and digital references associated with the remaining 12 Basic Inventories (e.g., geology, soils, weather, climate, air, non-vascular plants, and invertebrates).

Data Mining Team Staffing

Data mining was accomplished in three phases: 1) voucher mining, 2) Phase I data mining, and 3) Phase II data mining. The voucher mining was accomplished between July 2001 and December 2003 by the Network's Data Manager and with assistance from park staff. The voucher mining was later amended with the use of a Southern Oregon University (SOU) graduate student (Julies Filipski) and the SOU DMT lead. Phase I data mining was carried out from January 2004 to February 2005 by a team of eight network staff: six 1 year temporary NPS employees, an SOU cooperator as team lead, and the Network Data Manager. The Phase II DMT consisted of three term NPS employees and the Network Data Manager, who worked on the data mining project from May 2005 until its completion in October 2007.

Protocol Development

In order to have a consistent, standardized data mining effort, the DMT underwent an initial training at the beginning of the program. The data miners first read the national guidance and justification for the I&M Program, including NPS-75. This familiarized them with the program's background, the Network's structure, the 12 Basic Inventories, and the importance of the data mining project. During this time, the data miners received their general NPS accounts (email, credit cards, etc.) and logins for the national databases they were to use. In implementing Phase II, the DMT also read the Klamath Network Phase I Data Mining Protocol (Smith et al. 2005), which outline work completed and lessons learned during the Phase I data mining. The protocol contains information on methods used to search for and enter the data. These methods were further discussed in the training, which specifically described what record or information type does and does not warrant capture and how to enter information on the documents, datasets, and vouchers found. Data miners developed a Phase II-specific protocol based upon the methods detailed in the Phase I protocol, in order to standardize data entry of this different type of information (mainly datasets and digital references). Both sets of KLMN protocols (Phase I and

Phase II) are available on the KLMN web site at:
http://science.nature.nps.gov/im/units/klmn/Inventories/Basic_Inventories/INV_Bibliography.cfm

Point of Contact

Each park designated a Point of Contact (POC) to work with the DMT to tailor the data mining efforts to meet the park's needs and priorities. In addition, the POC kept in communication with the KLMN Data Manager to ensure everyone was current on the data mining efforts. Upon arriving at a park, data miners met with the POC to become oriented with the park's data locations, obtain background information, discuss procedures and protocols, and develop an overall plan. The POC coordinated the logistics (e.g., housing, office space) and scope of the data mining project. Further park-specific training was conducted, as necessary, on the filing structure and park priorities.

Databases (I&M Applications)

Three databases (NatureBib, NPSpecies, and Dataset Catalog) developed by the National I&M Program were used by the data miners to document and organize the reference information, species-specific information, and datasets. NatureBib and NPSpecies have both online and desktop functionalities, while Dataset Catalog is a desktop application that has since been replaced by the NPS Metadata Tools and Editor.

NatureBib

The NPS developed NatureBib as one of the 12 Basic Inventories. NatureBib is an online natural resource bibliography that houses information on references, datasets, and other materials related to the NPS. The database is populated with citations of works that are found not only in parks, but also in other collections (e.g., museums, universities). NatureBib is primarily used to catalog and manage reports, articles, conference proceedings, theses and dissertations, gray literature, and other documents containing information on park natural resources. Currently, there are over 300,000 references in NatureBib. The official site, containing many more details and links (including login requests, the main online database, and background information) is:
http://science.nature.nps.gov/im/apps/nrbib/index.cfm

NPSpecies

Along with NatureBib, the NPS developed a database that houses information on species presence/absence within the national park system. This database, NPSpecies, contains links for references and datasets mentioning scientific names. All references linked to NPSpecies are species-specific (or the lowest level of scientific identification possible). Linkages between NPSpecies and NatureBib connect the citation and species information for a reference. The taxonomic names list comes from the Integrated Taxonomic Information System (ITIS) (http://www.itis.gov/index), an authority on taxonomy. The official NPSpecies web site contains all the information on the database's background, login requests, and links to the online and desktop NPSpecies applications. This site is found at: http://science.nature.nps.gov/im/apps/npspp/

Dataset Catalog

Dataset Catalog (a desktop application) is an I&M Program tool for inventorying and providing abbreviated metadata ("metadata lite") about natural resource datasets. It provides a means for parks to inventory physical and digital files, spatial and non-spatial data files, notebooks of field data forms, photographs, etc. Dataset Catalog is not intended to be an exhaustive metadata listing, but rather a basis for implementing comprehensive metadata standards by generating minimal metadata. The one-page input and report forms provide a straightforward way to document resource data that may or may not meet formal metadata standards. Dataset Catalog can be linked to the desktop versions of NPSpecies and NatureBib.

While Dataset Catalog is still supported in terms of fixing bugs and providing back-end conversion support to users, no additional features or versions will be created. Dataset Catalog is being phased out, as its applications are being integrated into the NPS Metadata Tools and Editor and the NPS Data Store. Full information on Dataset Catalog, including links for downloading the application, may be found at: http://science.nature.nps.gov/im/apps/datacat/index.cfm

Database Training

NPS staff from the National I&M Program trained the data miners on NatureBib, NPSpecies, and Dataset Catalog. Training consisted of learning the databases' desktop and online versions and the appropriate data entry protocols. These trainings included directions on what should be entered, how to enter it, and what to do if a problem arose. In the initial training, the focus was on Phase I material (i.e., hardcopy references pertaining to vertebrates and vascular plants).

Later in the project, the DMT received formal training on creating metadata for datasets related to the 12 Basic Inventories. Metadata would be captured in Dataset Catalog and, for qualifying datasets, exported to the NPS Metadata Tools and Editor and then uploaded to the NPS Data Store (previously called the NR-GIS Metadata and Data Store). The Klamath Network Data Manager gave the Dataset Catalog training, which covered understanding various dataset formats, entering metadata into the Dataset Catalog, parsing the metadata, adding taxonomic hierarchies to metadata containing scientific names, and uploading the complete metadata to the NPS Data Store. The DMT tested the database in the office and then took this knowledge to the parks, developing a protocol and Metadata Interview (Appendix A) for its use.

Finding and Documenting Data and Information

The Klamath Network DMT's purpose was to locate and document information on the natural resources held within the KLMN parks. As part of this effort, the data miners created maps and task lists to organize and prioritize the work (Bridy et al. 2004). After discussions with park personnel and a preliminary resources inventory, the data miners formed a plan for mining that specific park. For the main information types, the documentation methods are listed below.

Vouchers

The KLMN captured voucher evidence records to document vascular plant species lists in each park. Vouchers are kept in many areas outside of the park in which they were collected, so the first task was to locate these voucher specimens. This work was conducted from 2001-2003 by Robert Truitt, KLMN Data Manager, and Jules Filipski, a Southern Oregon University graduate student (Smith et al. 2005).

To find vouchers relevant to the KLMN parks and retained in various museums, universities, and private collections, a request was sent out to all known major collection facilities. Specific focus was on those institutions with known collections of vertebrates and vascular plants. Records from all areas were combined, false records discarded, and the information entered in a project-specific database containing voucher metadata. After whittling down the records, a voucher species list for each park was formed. Each park then further verified its respective list. The lists were then submitted to NPSpecies, where they were uploaded with appropriate notes on the park status of the species. These initial lists formed the basis of the NPSpecies voucher records for each park. The voucher information was also provided to park curation staff.

As a result of the voucher mining work, three particular noteworthy recommendations became evident: 1) request all the voucher records for a category (e.g., all fish specimen records); 2) eliminate records, removing those that are documented to have been collected outside of your area of interest; and 3) request all records directly from the facility and not via an Internet search. The reason for these recommendations became evident when we first tried the logical approach of filtering web voucher records (based upon various criteria such as park name, state, county, etc.) from a facility. First, a number of facilities did not have web access to their collections, but they did have the records in an alternate electronic form and were able to provide them in a spreadsheet format. Second, for those collections where we were able to obtain web downloads of vouchers, upon comparison of records both requested and obtained through a web filter, we found a large number of park relevant records that did not have populated fields in the attributes we had filtered for. This was also helpful in consolidating the list of all records obtained into a list of those relevant for a park unit, by eliminating those we were positive were not collected from the park. We then collaborated with long-term park staff familiar with historical location names to identify a number of vouchers in the reduced voucher records.

In a separate but related project, vascular plant vouchers from Whiskeytown, housed at the Shasta Community College herbarium, were checked for identification accuracy by Windy Bunn, a Whiskeytown employee. This project resulted in finding numerous misidentifications. All of these were documented and updated in NPSpecies, strengthening the park's species list. Details on this project, including the report resulting from this work, may be obtained at: http://science.nature.nps.gov/im/units/klmn/Inventories/Vouchers/Bunn.cfm.

Hardcopy

Based on the POC's input, data miners picked an appropriate location to start searching and proceeded methodically through file drawers and bookshelves of each park for paper document records. If a reference contained information about vertebrates, vascular plants, or water quality, that information was entered or updated in NatureBib. Data miners linked the scientific names in references to appropriate entities in NPSpecies. References containing only species' common names were entered in NatureBib but not in NPSpecies. At most parks, the NatureBib number was written on the document or on a sticky note placed on the document. An Excel spreadsheet of hardcopy locations was continuously updated as data miners progressed through the areas (Appendix B). This shared workbook allowed all data miners to see where the others were working, which areas needed further work, and any details on issues with specific files. All drawers, folders, or files that looked like they might contain sensitive personnel information were skipped.

Digital

Data mining a park's digital files took considerable planning. The DMT used Directory Printer (http://www.galcott.com/dp.htm) to export a complete file structure of the drive to Excel (Appendix C). This workbook mimicked the DMT's workbook for hardcopy files and was similarly updated for specific digital folder/file locations with progress notes. Further notes were made on specific files if necessary in this workbook (e.g., folders containing duplicate, partial, or corrupt files). A useful tool was The File Extension Source web site (http://www.filext.com), which lists the corresponding current and historic programs associated with different file extension codes. This was extremely useful, not only to open files possibly containing important resources, but also to efficiently skip unimportant files. No folders possibly containing sensitive personnel information were opened.

Park-specific Methods and Outcomes

Although the data mining protocols (Smith et al. 2005, Bridy et al. 2006) provided guidance for most cases, entry methods were still tailored to each park. This was done due to the differing amounts and types of information at each park, preferences of park personnel, and time constraints. Each KLMN park is listed below, along with notes on the methods performed there and whether or not any deviations from the normal protocol existed.

Crater Lake National Park

The DMT discovered natural resource references for Crater Lake National Park (CRLA) in the main park buildings. In 2004, the DMT mined these locations for Phase I references. They systematically moved from the Natural Resources Building (Rat Hall) to the Ranger Station (Canfield Building) to the Visitor Center (Steel Building). Of note in Rat Hall were three large filing cabinets at the center of the building, the attic, and the Terrestrial Ecologist's office. The Canfield Building contained water quality information, especially in the Fish Biologist's and Aquatic Ecologist's offices. Finally, the Steel Building houses the park's library. The Phase I materials found in these locations were entered, with the exception of one box due to safety concerns and some microfiche that could not be read due to a broken machine. Sticky notes with the NatureBib number were attached to the processed references.

In 2006-2007, the DMT searched for Phase II information. Ample information was found in the same areas as in Phase I, plus the Science and Learning Center. The computer shared drives and hard drives in these locations were also data mined. During Phase I, Phase II information was not noted; during Phase II, all areas were quickly re-searched for Phase II information, new Phase I references, and Phase I references that needed additional data (e.g., change of location). In this examination, the second DMT found many references and datasets, especially concerning geology and geothermal studies. Other datasets captured focused on vegetation and wildlife. The DMT did not search personal offices for Phase II information.

During the Phase I effort, copies of digital files were moved by park and KLMN staff to a common data mining folder for capture. Only a few references were captured in this manner; most were copied from the common drive and entered from the KLMN office. During Phase II, the DMT worked directly on the CRLA shared drive to capture the digital references and datasets. In doing so, 90-95% of the information on the shared digital drive was mined and pertinent references and datasets captured in the appropriate databases.

Lassen Volcanic National Park

Park staff involvement at Lassen Volcanic National Park (LAVO) greatly aided the data mining effort. At the start of Phase I, the POC developed a spreadsheet of locations containing relevant information, encompassing the Science Center, Interpretive Library, Manzanita Lake Loomis Museum, and the Manzanita Lake Discover Center. The DMT noted on this spreadsheet their progress, the areas with Phase II information, and the procedures being followed. The majority of the Phase I effort focused on the Science Center. LAVO staff marked priority references for NatureBib entry. After entering these, the DMT returned to the Phase I protocol (Smith et al. 2005), including capturing references stored at Redwood National Park's archives.

A method unique to LAVO was that, after a document had been entered and the NatureBib number written on it, it was moved to the Research Library, per the request of the park POC. References were not moved from the Natural Resource Manager's office, Manzanita Lake, Interpretive Building Library, main office bookshelves, or personal shelves. To standardize the Research Library entries, "Research Library" was stamped across the document's top edge and a label was placed along the left edge with the author's last name and title's first word.

LAVO's involvement in their NPSpecies list also had an additional level of detail. In other parks, all species names in a reference were entered into NPSpecies, with the intention of park staff later refining this list. Since LAVO's NPSpecies list was already in good standing, data miners did not enter new names without park approval. Instead, they checked with park staff about species found in references but not on LAVO's list. If the species had a synonym on the park list, it was entered using that name. If no synonym was found, the park ecologists decided whether or not to add the species.

Some information was not captured, including lake data in the Research Library (mainly non-priority partial copies). The decision by park staff to only enter references of scientific merit (e.g., no newspaper articles) meant some lesser gray literature was not captured. LAVO retained relatively little of this low priority information, so excluding it did not cause undue delays of separating these low priority documents from the ones of merit.

The Phase II DMT entered hardcopy references concerning air quality, geology, and geothermal studies. After capturing this hardcopy information, the DMT entered LAVO's Phase II information and hardcopy natural resource maps. The LAVO digital drive was mined remotely while working in other parks, after the DMT received permission from LAVO. None of the personnel folders were mined, nor were any personal computers.

Capturing datasets at LAVO went smoothly; park staff worked closely with the DMT to ensure proper data entry. First, staff listed the priority datasets. Air quality datasets were not entered, as this data has been captured in a national database and park staff felt entering it at the park level would duplicate work. Second, they split the datasets and the corresponding staff member completed a Metadata Interview form (Appendix A) for each. Third, the DMT used these interviews to enter the datasets into Dataset Catalog. These collaborative steps greatly streamlined the data mining process.

The vast majority of relevant information found was captured in the appropriate databases. LAVO Resource Management staff consistently enters new materials and new park additions.

Lava Beds National Monument

Data mining at Lava Beds National Monument (LABE) concentrated on entering information in the Natural Resources office, Visitor Center Library, and Fire Ecologist's office. At first the

DMT wrote the NatureBib number on a sticky note, but later the LABE staff decided to write the number directly on the document; all references were then updated.

The DMT used an Excel spreadsheet (Appendix B), noting relevant information on the documents, their locations, and data mining progress. This spreadsheet was helpful when finding dataset locations in Phase II, since the park's filing structure did not change much during the data mining effort.

During Phase I, the DMT captured all of the digital references on vertebrates and vascular plants. A CD of the park's shared drive was made; the DMT used it to enter the pertinent documents. Also, park staff had converted 5 inch floppy disks, in DOS format, to Word versions and put this information on CDs, which were also data mined. In Phases I and II, species lists that did not have basic identifying information were not entered. However, lists with an identifying characteristic (e.g., clearly part of a LABE study) were entered into NPSpecies.

In 2006, data miners systematically went through the filing cabinets, capturing all relevant Phase II, and any new Phase I, information. After completing the hardcopy files, the DMT mined the remaining digital files. The area's geology, including the caves and soils, was detailed in many references and datasets. The shared drive's folders relevant to the 12 Basic Inventories were fairly well organized and navigable, which made finding digital documents and datasets efficient.

The DMT located complete hardcopy and digital datasets. Metadata for hundreds of non-GIS datasets were captured in Dataset Catalog, with the majority being hardcopy, older datasets. After these were entered, the DMT completed metadata on current projects. In particular, weather and air quality data, wildlife sightings, bat outflight information, cave conditions, and vegetation responses to fire were among the major categories of dataset information captured.

Due to the sensitive nature of some of the park resources, certain files were excluded from the data mining effort. These in-house park resource documents (e.g., detailed sensitive cave information) were not mined, per the request of park staff.

The park staff was very receptive to the idea of using the I&M databases to capture LABE references and datasets. Staff interest and participation in database upkeep has helped keep the LABE collection current; they actively update the databases with new information.

Oregon Caves National Monument

Oregon Caves National Monument (ORCA) was the first park in the KLMN to be data mined. As such, procedures were developed and revised here. Natural resource information was found in the Natural Resources Library, Natural Resource Management office, and the history file cabinet drawers. Due to an insufficient Internet connection in Phase I, NatureBib and NPSpecies' desktop versions were used, with entries later sent to the National I&M staff and posted online.

The Resource Chief requested some hardcopy file reorganization. References in the Natural Resources Library's park-specific filing cabinet were arranged alphabetically by author and chronologically by date, consolidating multiple copies of the same document. Most of the folders were organized in this manner, but no folders were moved or documents removed.

Phase II references and datasets were marked with a sticky note and listed in a spreadsheet (Appendix B). The Phase II DMT utilized this spreadsheet of location data. The second DMT finished entering almost all hardcopy references pertaining to the 12 Basic Inventories, with about 1% undone in the library/publication collection. Datasets entered in Dataset Catalog focused mainly on ORCA's vegetation, cave resources, and wildlife sightings.

At the request of the Chief of Natural Resources, geology and invertebrate information was the highest priority in the Phase II effort. The Chief of Natural Resources moved all pertinent digital data from the hard drives to the shared drive to facilitate data mining, immensely aiding the data mining effort. Phase II digital data mining ocurred both at ORCA and through remote access at other parks. Furthermore, ORCA material was captured at multiple other parks (i.e., information related to ORCA was entered with holdings in other parks). Through remote data mining and multiple site visits, all Phase I and Phase II materials on the shared drive were captured.

Redwood National and State Parks

Redwood National and State Parks (REDW), which have supported extensive natural resource and research programs, contained a huge amount of scientific data and information. Due to the vast quantity of documents and datasets available at REDW, the protocol for this park was altered more than at any other park.

The DMT's first step at REDW was to outline information in each location. In this outline, the DMT included a description of the type of documents and datasets that were in each place. Prioritization codes, set by the DMT and park staff, were given to each location based on the topics that location contained: 1. Phase I materials; 2. Phase II materials; and 3. Materials beyond the scope of the data mining effort.

After outlining REDW's locations, the data miners began with the priority 1 areas. The DMT entered references containing information from areas within and adjacent to the park. Aquatic information about any portion of Redwood Creek or its watershed was also entered, as were any references that dealt with anadromous fish in the Smith River or Klamath River.

Starting in September 2004, the protocol for entering correspondences was altered, due to the sheer volume of these filed documents. With the switch, the DMT entered only correspondences that were: in report format, containing a scientific name, or referencing information with scientific backing. This change focused the project on entering more references containing substantial information.

To avoid duplicate work in Phase II, Phase I data miners linking vertebrate and vascular plant species in a document to NPSpecies also linked any invertebrates and non-vascular plants in that document at the same time. Also, the protocol for entering only references containing scientific names in NPSpecies was altered for REDW staff's bird species lists. If these contained common names, the scientific names were found in the American Ornithologist's Union checklist (http://www.aou.org/checklist/index.php3) and entered into NPSpecies.

The DMT wrote NatureBib numbers on the hardcopy documents. They also kept spreadsheets detailing their work (Appendix D), immensely streamlining the process for determining which areas were done and where certain information (e.g., datasets, invertebrate references) was located. They did not reorganize any files at REDW. For the most part, all areas were already organized by park personnel. Multiple copies of the same reference were often encountered in different areas; these files were not consolidated, but rather the additional holdings were added in NatureBib.

REDW was first data mined in 1998 by Jeanne Pendergast of the NPS Pacific West Regional Office, working in conjunction with the REDW Archivist, Bow O'Barr. Records were entered into the Procite database, NRBib, and later uploaded to NatureBib. Therefore, many files pre-1998 already had a NatureBib citation. However, new citation details were added by the KLMN DMT (e.g., changes in file location, more complete biodiversity information, NatureBib numbers, and NPSpecies links).

The initial Phase II focus was on capturing hardcopy documents and digital data. After these goals had been met, the DMT shifted focus to hardcopy data. All hardcopy documents through the date of the DMT's effort were captured, minus a few cabinets in the Geology Library in Arcata and a couple low priority locations. The other hardcopy hydrology, geology, and soils documents in the Geology Library were entered and digital datasets noted on the DMT's progress spreadsheet but not entered into Dataset Catalog.

In Phase II, park personnel completed Metadata Interviews (Appendix A) on priority datasets, working with the DMT to capture relevant datasets in Dataset Catalog. As this was the first park where the DMT used Dataset Catalog, the process for entry was developed and refined here. The data miners worked to capture as many datasets as possible, focusing on current and recent digital datasets. All priority datasets from the Wildlife Management and Vegetation branches and some of secondary importance from the Geology branch were captured in Dataset Catalog.
As at other parks, sections of the shared digital drive were in various states of organization. The Phase II DMT captured as much information as possible, adding the NatureBib number to the document's File Properties. Some files were converted into the latest version of Word to make them readable. However, the DMT skipped folders containing all unknown file types or corrupted files. The DMT systematically went through the folders on the shared drive (minus folders related to administration), tracking progress and relevant information on a shared spreadsheet (Appendix C).

Whiskeytown National Recreation Area

The Natural Resource Building, Building 318, and the Whiskeytown National Recreation Area's (WHIS) Library housed natural resource information to be data mined.

The Natural Resource Building had many areas containing information related to vertebrates and vascular plants. All of these hardcopy references were captured in Phase I. Datasets were noted on the DMT's spreadsheet that may have Phase II information. The Natural Resource Chief requested that some file reorganization occur in this location; some non-resource management files were thus discarded and disparate folders on the same subject matter were combined.

All of the relevant information in Building 318 was captured during Phase I. Although searched again during Phase II, no new information was found. It does not seem that any additional information is being added to this location; the DMT considered this location completely mined.

The WHIS Library contained references in a large bookcase. All pertinent references here were entered, located with the Library's finding aid. It was checked again in Phase II, but all of the relevant information housed in the Library had already been entered.

In Phase II, the majority of the effort was spent at the Natural Resources Building. The hardcopy files were again mined for Phase II information and some previously undiscovered files and new files related to the 12 Basic Inventories were captured. As all hardcopy documents have been captured, and park personnel enter new documents into NatureBib and NPSpecies during the winter season, the data mining effort at WHIS is effectively complete.

WHIS' datasets were in the Natural Resources Building and digital shared drive. Primarily, datasets on water quality monitoring, vegetation studies in fire plots, and the geologic resources and soils of WHIS were entered in Dataset Catalog. Very few datasets on wildlife were located.

As in other parks, WHIS' shared drive was first given a cursory examination. This was done by examining the folder structure and talking to park personnel, who stated that all relevant information was on the shared drive and none was on individual computers. The Phase II DMT systematically went through the shared drive's file structure, eliminating folders for entry with non-relevant information (including administrative information). Files with 12 Basic Inventories information were entered into NatureBib and NPSpecies. The NatureBib number was not written into the document's File Properties, per the request of park personnel, but was marked on the digital progress spreadsheet (Appendix E). Most of the digital information available has been captured, with particular focus on the vegetation-related files.

Database Results

NatureBib

There are now about 15,000 references linked to Klamath Network parks in NatureBib. Table 1 shows the details of park numbers and staff responsible for entering the references. References linked to the Klamath Network are references that deal with an Inventory and Monitoring project, are related to inventory and monitoring efforts in the Network, are not park-specific but housed at the Klamath Network office, or concern multiple parks in the Network. From the data miner's detailed records and known dates of record entry, the number of references entered by each group has been deduced.

Table 1. Number of references linked to NatureBib for each Klamath Network park unit showing the responsible contributors. DMT = Data Mining Team.

Park	Pre-2004 Entries*	Phase I DMT Entries	Phase II DMT Entries	Park Personnel Entries**	TOTAL
Crater Lake National Park	1200	406	490	60	2156
Lassen Volcanic National Park	1440	404	171	98	2113
Lava Beds National Monument	770	444	480	37	1731
Oregon Caves National Monument	249	529	156	85	1016
Redwood National and State Parks	2512	1551	1796	0	5859
Whiskeytown National Recreation Area	603	191	163	63	1020
Klamath Network	65	1284	120	31	1500

**Pre-2004 entries include both records uploaded to NatureBib from other databases (e.g., Procite) and entries by others directly in NatureBib before the Klamath Network data mining effort began.*

***Entries by park personnel are records created in NatureBib by park staff during the Klamath Network data mining effort.*

Table 2 shows the NatureBib references edited by the second DMT. This number of edits is known but the number of edits by past personnel is unknown. Edits to a NatureBib record usually involved adding to the holdings location when copies of the reference were found and adding to NatureBib categories (e.g., Biodiversity section) that may not have previously been part of the entry process. Many times, the number of edits depended upon the number of duplicate files. For example, the same files were found repeatedly at REDW. Other parks, such as LAVO, with a stricter filing system, did not have nearly as many duplicate references.

Table 2. The second Data Mining Team's NatureBib edits for each Klamath Network park unit.

Park	NatureBib Edits
Crater Lake National Park	280
Lassen Volcanic National Park	65
Lava Beds National Monument	557
Oregon Caves National Monument	108
Redwood National and State Parks	4039
Whiskeytown National Recreation Area	146
Klamath Network	153

NPSpecies

The KLMN DMT linked any reference containing a scientific name to NPSpecies. Some references only contained one and others contained hundreds. All references were first linked to NatureBib. The DMT carefully checked the documents' species lists to make sure that all species linked to the park were actually present in the park, not just mentioned in the document for another reason (e.g., species found on nearby lands, species closely related to park species, or species that might have historically occurred in the park). Table 3 below displays the number of references linked to each park and the group that linked the reference, through the end date of the project. Unlike references linked to the Klamath Network in NatureBib, no references are linked to it in NPSpecies since each species must specifically occur in a park.

Table 3. Number of references linked to NPSpecies for each Klamath Network park unit, showing the responsible contributors. DMT = Data Mining Team.

Park	Pre-2004 Entries*	Phase I DMT Entries	Phase II DMT Entries	Park Personnel Entries**	TOTAL
Crater Lake National Park	9	241	157	27	434
Lassen Volcanic National Park	25	95	45	130	295
Lava Beds National Monument	0	209	122	14	345
Oregon Caves National Monument	4	55	34	16	109
Redwood National and State Parks	26	971	833	4	1834
Whiskeytown National Recreation Area	1	103	63	60	227

**Pre-2004 entries include both records uploaded to NatureBib from other databases (e.g., Procite) and entries by others directly in NatureBib before the Klamath Network data mining effort began.*
***Entries by park personnel are records created in NatureBib by park staff during the Klamath Network data mining effort.*

Each park's species list is based on park research, vouchers, species lists confirmed by an expert, and references the DMT entered. The general method for entering a new species to these lists was to enter the species name exactly as it was written in the document (citing the source document on the record), even if this name was not a valid recognized scientific name or was an identifiable misspelling. The reasoning was that through verification and validation, as well as through periodic list maintenance, examinations and changes could be made to any of the names and linkages between misspelled and properly spelled names could be made, ensuring the name in NPSpecies matched the name in the reference. Table 4 displays the total number of scientific names that were linked to each of the parks as of the end of the project.

Table 4. Scientific names linked to NPSpecies for each Klamath Network park unit at the end of the Data Mining Project.

Park	NPSpecies Scientific Names
Crater Lake National Park	3666
Lassen Volcanic National Park	2047
Lava Beds National Monument	1435
Oregon Caves National Monument	2254
Redwood National and State Parks	6625
Whiskeytown National Recreation Area	2255

Dataset Catalog

One of the primary goals of the Phase II data mining was to capture park-related datasets pertaining to one of the 12 Basic Inventories, the second Data Mining Team entered hundreds of datasets into Dataset Catalog (Appendix F-K). The first goal was to capture current datasets since park personnel related to these studies were still at the park and the DMT could glean as much information as possible about these datasets. The Metadata Interview (Appendix A) was helpful in ranking priority datasets. Past datasets, where staff working on the studies may no longer be at the park, were secondary priority. Further details on entry methods used for Dataset Catalog are in the Klamath Network Data Mining Phase II Protocols (Bridy et al. 2006).

Table 5 depicts the total number of datasets (i.e., with minimal or full metadata) entered into Dataset Catalog for each park by the DMT. At all parks, metadata development reached the level of Dataset Catalog. Time limitations prevented the appending of species information and upload of metadata to the NPS Data Store.

Table 5. Dataset Catalog entries made by data miners for each Klamath Network park unit.

Park	Dataset Catalog Entries
Crater Lake National Park	91
Lassen Volcanic National Park	77
Lava Beds National Monument	129
Oregon Caves National Monument	52
Redwood National and State Parks	277
Whiskeytown National Recreation Area	230

Discussion

The KLMN was one of the first networks to perform data mining as an integral part of its inventory and monitoring program, and it did at a scale and funding level unmatched amongst the other networks. Altogether, the data mining project lasted over six years with varying levels of staffing. The Klamath Network's commitment to the project and its funding is now complete and further updates will be joint park/Network ventures or based in-park. Data entry into each of the three databases (NatureBib, NPSpecies, and Dataset Catalog) brought each of the park's records current to the time of the latest data mining activity. To that end, the data mining project was successful in bringing the parks current in their knowledge and access to critical natural resource datasets. Depending upon the degree and rigor of data management activities in the parks, data mining may need to be periodically implemented to ensure that existing and new staff have access to and familiarity past natural resource efforts.

For future data mining efforts in the KLMN and elsewhere in the I&M Program, we present here some key "lessons learned" and recommendations to further strengthen and streamline the process.

First, data mining is greatly assisted when a park already has an archiving system in place. A rule set and protocols for archiving documents and datasets that have been entered in the databases are highly useful, especially when in conjunction with the park's collections strategy. Since parks with archival programs have all of their information stored in a standardized system, data miners can quickly locate pertinent documents and datasets. Not only would this archival program make a library of useful products, but it would also lessen the problematic issue of holdings locations changing in the databases with personnel turnover or office reorganization. Hence, fewer updates would be needed to records already in the databases and the quality of the records contained in the databases is strengthened. A good example is the resource management documents library at LAVO, which has a permanent location, set filing structure, and process for entering new documents.

For parks that do not already have an archiving system in place, it is imperative that they are given sufficient lead time before a data mining project is implemented, to best organize the park's files and determine the park's data mining needs. Communication with the park POC is invaluable at this juncture, in order to set park priorities for data entry. Consolidating the park's important files (from personal computers, network drives, and physical locations) in a systematic way would be ideal, benefiting both the project and the park. While a complete reorganization and change of structure is usually not practical or even feasible, it would be helpful to have some initial rules set in place so that not only can the data miners efficiently access pertinent files, but also so the park personnel better understand what resources they have in their parks. Although all parks accumulate files and records from a variety of activities, the investment of working with an archivist or other records management specialist to develop such a system will yield dividends in ease of access, lack of lost data, and facilitation of future data mining efforts.

Similarly, when data mining is complete at a park, park personnel should have a system in place to continue the effort of updating the databases. This ensures that the databases are useful and relevant to park staff. Once the data miners have completed entering the backlog of information, park staff trained on the databases could enter new documents and datasets as they develop. In-park presentations of the databases; their functions, design, and products; how to set up an account; and how to enter information while the data miners are in the park would be a valuable way to ensure that park personnel are familiar with and skilled in using natural resource databases.

Where feasible, data mining should encompass outside facilities, such as universities and museum collections, which are likely to hold substantial data and information about park resources. Many parks have had research completed by non-park personnel, and with staff turnover through time, information held in other locations may become lost to the park. A defined search of outside institutions in the region would very likely result in new park knowledge and improve the functionality of the NPS databases. For example, University of California, Berkeley has completed many studies at LABE and the university may contain documents pertinent to park resource management.

In addition, there are some factors that make the data mining process go more smoothly and increase data miner retention, as discovered by the Klamath Network effort. These suggestions for other parks and I&M networks to consider are listed below.

If possible, station individual data miners at one park for a substantial period of time, rather than traveling to different parks in short time periods (e.g., weekly or every few months). Familiarity with a park's natural resources, filing and library systems, and interested staff takes time but increases efficiency. If park permission is granted, an option may be to have the DMT remotely data mine other parks, allowing them to remain in one location longer and not take up valuable office space in smaller parks. This reduces costs and turnover while increasing work time, continuity, consistency, and cooperative relationships with the parks. Having continuity in a park also increases morale and allows the development of familiarity between data miners and park staff.

Scheduling data mining to occur during the research off-season allows researchers to better assist the data miners, which is particularly important when cataloging datasets (which are often incomprehensible without explanation). Also, park housing and office space are often more available during the off-season, for times when data mining remotely is not practical or possible. Moreover, seasonal staff may be easier to hire during the off-season than when fieldwork is in full swing.

Data miners should keep track of their database entries in an Excel spreadsheet (Appendixes B, C, D, and E). This is an easy way to track work completed in case of database problems and would facilitate better validation of files uploaded to the national web page (for data miners using the desktop version). Keeping current individual and group spreadsheets with detailed progress information was imperative to the success of the DMT. These spreadsheets kept the project well organized and allowed for systematic entry of all information in each park.

One key reason for the success of the Klamath Network's data mining effort was its flexibility to work with each park. This increased communication and tailoring to the specific needs of each park helped make the products of the project more useful and transparent. Through the course of the project, many lessons were learned as to the needs of each park, the I&M's effort to create a natural resource bibliography, and what partnerships were possible to continue collaboration. Hopefully, the I&M databases available will become indispensible to the parks and incorporated into daily data management routines.

In conclusion, the Klamath Network data mining effort was hugely successful in cataloging a wealth of the information available at the six parks and in making this information readily available through the I&M databases. The effort not only provided a comprehensive natural resource bibliography to support the 12 Basic Inventories of the I&M Program, but also provided a means for familiarizing park staff with pertinent park-focused natural resources information and how to efficiently access it. This intensive effort was indispensible for helping the parks gain access to years of accumulated data and information. With continued park engagement and partnership with the I&M Program and other research institutions, these data mining efforts will provide an invaluable base of knowledge and a data management system to support park management for future generations.

Literature Cited

Bridy, L., E. Perry, T. Shepherd, and R. Truitt. 2006. Klamath Network data mining phase II protocols. Natural Resource Report NPS/PWR/KLMN/NRR—2007/002. National Park Service, Oakland, California. Available at (http://science.nature.nps.gov/im/units/klmn/Inventories/Basic_Inventories/Documents/Data_Mining/PhaseII_Protocols_FINAL_20061121.pdf). Accessed 8 December 2008.

Bridy, L., R. Miller, C. Powell, B. Shaw, S. Smith, and H. Waterstrat. 2004. Klamath Network data mining final report for FY 2004. Klamath Inventory and Monitoring Network, National Park Service.

Leopold, A. S., S. A. Cain, C. M. Cottam, I. N. Gabrielson, and T. L. Kimball. 1963. Wildlife management in the national parks. Available at (http://www.nps.gov/history/history/online_books/leopold/leopold.htm). Accessed 11 September 2009.

National Park Service. 1992. NPS-75: Natural resources inventory and monitoring guidelines. U.S. Department of Interior, National Park Service, Washington, D.C. Available at (http://science.nature.nps.gov/im/monitor/docs/nps75.pdf). Accessed 25 August 2009.

National Park Service. 1999. Natural resource challenge. National Park Service. U.S. Department of Interior, National Park Service, Washington, D.C.

National Park Service. 2003. NPS: Nature & science: Inventory and monitoring of park natural resources. Available at (http://www.nature.nps.gov/protectingrestoring/IM/resourceinventories.cfm). Accessed 8 December 2008.

Robbins, W. J., E. A. Ackerman, M. Bates, S. A. Cain, F. D. Darling, J. M. Fogg, T. Gill, J. M. Gillson, E. R. Hall, C. L. Hubbs, and C. J. S. Durham. 1963. A report by The Advisory Committee to the National Park Service on research, of the National Academy of Sciences – National Research Council. Available at (http://www.nps.gov/history/history/online_books/robbins/robbins.htm). Accessed 11 September 2009.

Smith, S. B., R. E. Truitt, L. Bridy, E. Perry, and T. Shepherd. 2005. Klamath Network data mining phase I protocols. Natural Resource Report NPS/PWR/KLMN/NRR—2007/001. National Park Service, Oakland, California. Available at (http://science.nature.nps.gov/im/units/klmn/Inventories/Basic_Inventories/Documents/Data_Mining/KLMN_PhaseI_Protocols_082506_DMT_NRR_v3.pdf). Accessed 8 December 2008.

Appendix A: Klamath Network Metadata Interview

Metadata Project

by the Klamath Network

"As part of the Service's efforts to 'improve park management through greater reliance on scientific knowledge,' a primary purpose of the Inventory and Monitoring Program is to develop, organize, and make available natural resource data and to contribute to the Service's institutional knowledge by facilitating the transformation of data into information through analysis, synthesis, and modeling" (I&M Program, NPS 2004).

To achieve this goal, we (the Klamath Network Data Mining Team) will catalog natural resource data collected at each of the six national park units associated with the network (Crater Lake NP, Lassen Volcanic NP, Lava Beds NM, Oregon Caves NM, Redwood NSP, and Whiskeytown NRA). We are placing this information into an Access database and creating a catalog of key descriptive information about the data. Without the catalog, crucial facts about the data may be lost. We are including all datasets because each one may have current or future relevance to the I&M Vital Signs Monitoring Program. The cataloged information will be made available both to the park personnel and the public. Each park will be supplied with a full copy of the completed Dataset Catalog. The public will have limited access through the uploading of the Dataset Catalog to the NR-GIS Data Store by the network's data manger. The level of availability will depend on the completeness of the metadata, your preference, and the sensitivity of the information (e.g., T&E species). Also, the data or dataset will not be posted at the NR-GIS Data Store; only information "about" the data and who to contact if further information about the dataset is desired. The decision whether or not to make the dataset available will rest with the park and researcher(s).

We need your help identifying the most important data to catalog. Also, we would appreciate examining associated summaries and reports (progress reports, published reports, etc.) that go with each set of data. Please let us know the locations of these documents.

To capture essential descriptions of crucial elements of the data, please fill out the questionnaire on the following pages.

Appendix A. Klamath Network Metadata Interview (continued).

Metadata Interview Questionnaire

Please direct us to the most important datasets first.* If the answers to some of the questions can be answered with the dataset itself, or existing supporting documentation such as summaries and reports, please direct us to that information and skip to the next question.

Name______________________________________

1. Where is the dataset (e.g., Network Drive S:/Team/Veg/Sillett_tree_physiology.mdb)?

2. What is the **location** of the study? What areas of Redwood National and State Parks? Please include the names of any other **National Parks** in which the study was conducted.

3. What is the **title** of the dataset?

4. Who is the **author** of the dataset?

5. What is the **date** of the latest version (published or made available for release)?

6. What is the **permit number** (e.g., NPS Research Permit and Reporting System)?

7. What is the **project name**?

8. What is the **project number** (e.g., PMIS or RMP project number)?

9. Are there any **keywords** that you especially want to describe the dataset (e.g., northern spotted owl, Strix occidentalis caurina, abundance, Bald Hills)?

10. Please give a brief **abstract** of the dataset. Include information such as the project from which the data are derived, who was involved, general methodology used (#sites, sampling frequency, protocol, equipment), and references to concurrent or related data. If this information is available in a supporting document, please direct us to it, list the location below, and skip the rest of this question. If writing an abstract here, use extra pages if necessary.

Appendix A. Klamath Network Metadata Interview (continued).

11. Briefly, what is the **purpose** of the dataset (i.e., why was the data collected, and what use(s) or information will the data provide)?

12. Please indicate the **timeframe** of data collection in the most exact date possible:

__Single date: on ______________________
__Span of dates: from ___________________ to _____________________
__Multiple dates: (please list) _____________________________________

13. How **frequently** is the dataset **updated** (i.e., the interval at which new data are appended to the dataset)? Pick one:

__Continually __Daily __Weekly __Monthly __Annually __Biannually __As needed __Irregular __None planned __Unknown

14. What is the **status** of the dataset? Pick one:

__New a dataset in the planning, implementation, or collection stage
__Active data are still being added to the dataset periodically
__Inactive data is no longer being collected but may have future updates
__Legacy the data were collected by previous projects or personnel that needs validation and documentation
__Partial a dataset that is in work or that has not been completed per data gathering protocols/specifications
__Historic data without planned updates from historical natural resource activities
__Other

15. What **progress** has been made on this dataset?

__Planned __In work __Complete

17. Please briefly explain the **quality** of the data. Specifically, please address any issues that would affect data quality; whether or not the data has been verified, validated, and/or critically reviewed; and if the data were created according to a set standard.

Appendix A. Klamath Network Metadata Interview (continued).

18. Does this dataset contain **sensitive** information? ___Yes ___No

If **yes**, please specify who should be able to view this metadata:

__Public access denied __Federal only __NPS only __Park only

19. Please describe the **data fields** and give meanings for abbreviations (ex: gen = genus, Mamu = marbled murrelet). If this information is located in the data or elsewhere, please direct us to it, and skip this question.

20. Please briefly list any **additional** important information about the dataset that has not been captured elsewhere on this form.

*"The most important datasets" are those that are the most important to share with other researchers, and the most important to carry on to the next generation in perpetuity.

Citations

I&M Program, National Park Service. (August 3, 2004, Draft). I & M Data Management Vision and Framework. Inventory & Monitoring Program. Available online. (http://science.nature.nps.gov/im/datamgmt.htm). Accessed 1 November 2005.

Appendix B: Example of a Park Location Excel Spreadsheet

Room/Office	Section	Phase I Progress/Notes (see DMT#1 docs for names)	Phase II Progress/Notes	General Notes	Datasets	Digital Maps	Metadata	Observations	Proposals	Digitally Formatted Items
HEADQUARTERS:										
Resource Management Office										
File Cabinet 1	Drawer 1		Tim done	Mostly admin stuff; nothing to enter						
	Drawer 2		Tim done	Archeological Information: compliance and clearances						
	Drawer 3	done (EAs, plants, and animals)	Tim done	Folder: "N419 Bird Banding" (3 CDs of photos); "N1617 2004 Spring Plant Survey K. Riebeling" (aerial photos with plot locations, also sheet with UTM location (no projection)); "N1621 Misc Bald Eagle Info" (Photographs)						5” floppy disks; unknown content
	Drawer 4	done except Air Quality for 2nd round, IARs pre-1980 are done	Bess done	IARs entered and Air Quality captured						
	Drawer 5	done (pests)	Bess done		Data but not park-specific; general info					

Appendix C: Example of a Digital Location Tracking Excel Spreadsheet

S:\team\Veg Directory tree (structure)	Progress	Datasets	GIS/Maps	Notes/Comments
Exotics	Laura Done	\Species & Control Info\CYSC\crosstab.doc, CYSC, = Initial Scotch Broom Burn Response in Dolason and Elk Camp Prairies, crosstabulation. data entry forms for weed control. RXCYSC.doc = treatment sites and plant characteristics, not much explanation.		
Forestry: 2nd Growth Mngmt	Bess done!!!	...\2003 Forest Sampling (all files, incl. subdirectories = data in Excel files, huge datasets); ...\GIBSON (complete); ...\Lostman - Holterridge\... (datasets throughout folder and all subdirectories, enough for complete); …\stand reduction hr (varied Excel files w/data and graphs, maybe not good enough for even lite?); ...\Xowannutuk Plots\'03 SG Forest plots\'04 Post-burn data (Excel files only, not enough for complete); .../MCdata.dbf (Access table, not sure what it relates to)	...\Lostman - Holterridge\... (GIS projection files throughout); ...\Maps; ...\NRCS (Excel files have UTMs etc. & all soil survey sheets have coordinate info etc., but it's not specifically GIS formatted & there are no maps)	IMAGES: ...\2nd growth scoping pictures; ...\2003 Forest Sampling\Photos; ...\2003 Forest Sampling\thinning pictures; ...\2004-2005 Westside Cruise; ...\Lostman - Holterridge\... (images throughout subdirectories); …\Whiskey-40 Cruise 2005 (images and movies throughout); ...\Xowannutuk Plots\... (images and Excel-formatted explanations throughout); ...
Forestry: Acres	Tim done	Two excel files with acres, one by vegetation community, the other differences based on current and past descipt.		
Forestry: Mill Creek '04-'05 survey\11-24yr Stand info	Directory EMPTY			
From shdata\Veggie D\DAVISON\WETLANDS	Tim done			WETFIELD.WP (Species, symbols and indicator codes of vascular plants observed at the Davison Ranch, Redwood Creek estuary, Freshwater Lagoon, and South Operations Center vicinity.)
From shdata\Veggie D\DAVISON\WETLANDS\ELEV	Tim done	*.RAW (input files for programs to determine elevation); *.DAT (output files); not sure what other files are for (.GRF, .PLT, .OUT)		
From shdata\Veggie D\LICHEN	Bess done	it looks like this folder is a complete, stand-alone dataset!		

Appendix D: Example of a Data Miner’s Hardcopy Progress Excel Spreadsheet

Folder structure			Phase II Progress/Notes	General Notes	Datasets	Digital Maps	Images	Metadata
Natural Resources	16 files	5,803,687 bytes	root done					
BAER_05	6 files	21,787 bytes	n/a	Photopoint sheet.xls and photopoints.dbf? Should open to see if relevant		GIS files		
Reveg	48 files	35,357 bytes	n/a	all GIS files				
Photopoints	0 files	0 bytes	n/a	folder may contain a photo dataset, but hard to tell without further information			veg pics	
Oaks	2 files	7,074 bytes	done	shrub oaks.txt and tree oaks.txt = plant ID				
Plant_Communities	1 file	92,672 bytes	done	"WHIS_Communities.doc" just contains general spp to look for in each community type and doesn't have any specifics, also very draft version, didn't enter				
Arnica	2 files	38,912 bytes	done		Arnica venosa survey, entered			
Orchard_2004	5 files	179,200 bytes	done		orchard dataset, entered			
Orchard_2005	2 files	50,176 bytes	done	added Needs Assessment as a related doc in DataCat but didn't enter in NB				
BAR	0 files	0 bytes	done	root empty				
Biotech2000	0 files	0 bytes	didn't open	looks personal				
Camden Ea	11 files	4,430,336 bytes	done	each file is a separate section				
draft_ea_internal_comments	0 files	0 bytes	n/a	empty				
Weeds	23 files	729,088 bytes	n/a	budget				
Exotics	15 files	3,642,368 bytes	done					
Maps 3-05	3 files	3,325,573 bytes	n/a			topo maps		
Exotics Management Plan	1 file	39,936 bytes	done					

Appendix E: Example of a Data Miner’s Digital Progress Excel Spreadsheet

Entry Date:	Nature Bib#	Author	Date	Title	NP SPP?	Location	Notes
2/6/06	**593286**	AU	1996	Second growth forest …	N	S:\team\Veg\Forestry\2nd Growth Mngmt\2ndgrowth.pln\Old\2NDGRFR.MAY	updated
2/7/06	**609945**	RNSP	2003	Map unit symbol …	N	S:\team\Veg\Forestry\2nd Growth Mngmt\NRCS\Converleg.doc	no bd info
	609947	RNSP	2003	No title [Explanation…	N	S:\team\Veg\Forestry\2nd Growth Mngmt\NRCS\leg2003.doc	entered entire folder as one holding; didn't mark NB# on the individual documents in that folder
	609956	RNP	1998	Folder: 1998 burn plans	N	S:\team\Veg\Forestry\RxFire\1998 Burn Plans	
2/8/06	**609963**	RNP	2000	Folder: 2000 burn plans	N	S:\team\Veg\Forestry\RxFire\2000 Burn Plans	
	609997	Hooke	1997	Fire behavior and weather …	N	S:\team\Veg\Forestry\RxFire\RXFM burn summaries\1997\countsnar0997,doc and \countsobs0997.doc \\ S:\team\Veg\Forestry\RxFire\RXFM burn summaries\1998\counts98.doc	no bd info; two documents in one holding, then one final doc in second complete holding
	592342	RNSP	ND	Redwood Currents	Y	S:\team\Veg\Forestry\RxFire\OCTNOV~1.PDF	Updated; only one newsletter; NB# for entire chunk of them
2/9/06	**610007**	Childers	1998	Upper Lyons…	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\UPPER LYONS PRESCRIBED BURN NARRATIVE.doc	no bd info
	610008	Childers	1998	Upper Dolason prescribed …	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\UPPER DOLASON PRESCRIBED BURN NARRATIVE.doc	no bd info
	610009	Underwood	1998	Burn boss narrative …	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\98lowerdolason.doc	no bd info
	610010	Underwood	1998	Burn boss narrative: …	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\98mainstem.doc	
	610011	Underwood	1998	Mammal plot narrative	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\98mammal.doc	no bd info
	610013	Underwood	1998	Burn boss narrative:…	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\98midbasinE.doc	
	610014	Underwood	1998	Burn boss narrative: Basin…	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\98midbasinW.doc	
	610015	RNSP	2000	Boyes Prairie prescribe…	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\Boyes Prairie.doc	no bd info
	610017	Arguello	1998	Narrative for Coyote…	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\COYONAR.doc	
	610020	LaBanca	1998	Burn narrative: Schoolhouse…	N	S:\team\Veg\Forestry\RxFire\Rxnarratives\SCLNAR98.DOC	changed from COYO.NAR to COYONAR.doc to make more stable; lots of imported files are like this
	578603	RNP	1994	Draft fire management plan…	Y	S:\team\Veg\From shdata\Veggie D\FIREPLAN	Updated; entered entire folder as one holding; didn't mark NB# on the individual documents in that folder
	610049	Reed	1992	Letter to Jim Agee	Y	S:\team\Veg\From shdata\Veggie D\LOISREED\OAKMGMT	
	23667	Reed	1986	Child's Hill Prairie oak …	Y	S:\team\Veg\From shdata\Veggie D\LOISREED\OAKMGMT\Dfgirdlg.wpd	updated

Appendixes F-K: Datasets Documented Using Dataset Catalog at the Six Parks Where Data Mining Occurred

Appendixes F-K provides a list of documented datasets and associated metadata for each park. Many of these datasets contain sensitive information about rare and/or endangered species. Copies of these appendixes have been provided to the park and can be obtained by contacting the natural resource staff at each park. Appendixes are as follows:

Appendix F: Crater Lake National Park Datasets Documented During the KLMN Data Mining Project

Appendix G: Lava Beds National Monument Datasets Documented During the KLMN Data Mining Project

Appendix H: Lassen Volcanic National Park Datasets Documented During the KLMN Data Mining Project

Appendix I: Oregon Caves National Monument Documented During the KLMN Data Mining Project

Appendix J: Redwood National and State Parks Datasets Documented During the KLMN Data Mining Project

Appendix K: Whiskeytown National Recreation Area Datasets Documented During the KLMN Data Mining Project

The Department of the Interior protects and manages the nation’s natural resources and cultural heritage; provides scientific and other information about those resources; and honors its special responsibilities to American Indians, Alaska Natives, and affiliated Island Communities.

NPS D-963/100531, October 2009

National Park Service
U.S. Department of the Interior

Natural Resource Program Center
1201 Oakridge Drive, Suite 150
Fort Collins, CO 80525

www.nature.nps.gov