National Park Service
U.S. Department of the Interior

Natural Resource Program Center

Data Management Plan – Southeast Alaska Network

Natural Resource Report NPS/SEAN/NRR—2008/058

ON THE COVER
Sunset at Bartlett Cove, Glacier Bay National Park and Preserve.
Photograph by: Craig Murdoch, NPS.

Data Management Plan – Southeast Alaska Network

Natural Resource Report NPS/SEAN/NRR—2008/058

William F. Johnson and Brendan J. Moynahan
National Park Service
Southeast Alaska Inventory and Monitoring Network
3100 National Park Road
Juneau, Alaska 99801

September 2008

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

The Natural Resource Publication series addresses natural resource topics that are of interest and applicability to a broad readership in the National Park Service and to others in the management of natural resources, including the scientific community, the public, and the NPS conservation and environmental constituencies. Manuscripts are peer-reviewed to ensure that the information is scientifically credible, technically accurate, appropriately written for the intended audience, and is designed and published in a professional manner.

Natural Resource Reports are the designated medium for disseminating 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. Examples of the diverse array of reports published in this series include vital signs monitoring plans; monitoring protocols; "how to" resource management papers; proceedings of resource management workshops or conferences; annual reports of resource programs or divisions of the Natural Resource Program Center; resource action plans; fact sheets; and regularly-published newsletters.

Printed copies of reports in these series may be produced in a limited quantity and they are only available as long as the supply lasts. This report is also available from the Natural Resource Publications Management website (http://www.nature.nps.gov/publications/NRPM) and the Southeast Alaska I&M Network website (http://www.nature.nps.gov/im/units/SEAN) on the Internet, or by sending a request to the address on the back cover.

Please cite this publication as:

Johnson, W. F., and B. J. Moynahan. 2008. Data management plan: Southeast Alaska network. Natural Resource Report NPS/SEAN/NRR—2008/058. National Park Service, Fort Collins, Colorado.

NPS D-170, September 2008

Acronyms

DMP	Data Management Plan
DOI	Department of the Interior
GIS	Geographic Information System
GLBA	Glacier Bay National Park and Preserve
I&M	Inventory and Monitoring Program
IRMA	Integration of Resource Management Applications
KLGO	Klondike Gold Rush National Historical Park
NPS	National Park Service
NRDT	Natural Resource Database Template
SEAN	Southeast Alaska Network
SITK	Sitka National Historical Park
SOP	Standard Operating Procedure
VSMP	Vital Signs Monitoring Plan
WASO	Washington Support Office, Fort Collins, CO

Document Change History

Version Date	Author	Nature of Change
1/8/2008	Bill Johnson	Initial outline
2/16/2008	Bill Johnson	Outline overlaid with narrative
2/25/2008	Bill Johnson	Draft for VSMP
3/10/2008	Bill Johnson	Appendix A added
4/15/2008	Bill Johnson	Appendix B added
9/30/2008	Bill Johnson	Revised following peer review
10/15/2008	Bill Johnson	Formatted for publication

Contents

Page

EXECUTIVE SUMMARY ..... **xiii**

CHAPTER 1. INTRODUCTION **1**

1.1 PLAN CONTEXT 1
1.2 PLAN PURPOSE 1
1.3 PLAN AUDIENCE 1
1.4 ROLE OF INVENTORY AND MONITORING 1
1.5 MISSION OF THE SEAN 2
1.6 THE SEAN DOMAIN 2
1.7 THE SEAN CUSTOMERS 3
1.8 DATA MANAGEMENT NETWORK DESIGN PHILOSOPHY 5
1.9 PLAN REVISIONS 9

CHAPTER 2. INFRASTRUCTURE AND SYSTEMS ARCHITECTURE **11**

2.1 THE SEAN OVERALL ARCHITECTURE 11
2.2 STANDARD COMMERCIAL APPLICATIONS 12
2.3 ENTERPRISE APPLICATIONS 13
2.4 IP NETWORK AND WEB SERVER INFRASTRUCTURE 13
2.5 DATABASE SERVERS 16
2.6 MID-TIER 16
2.7 FILE STORAGE 16
2.8 INTER-NETWORK SHARED RESOURCES 16

CHAPTER 3. PROJECT MANAGEMENT AND DATA PROCESSING WORKFLOW **19**

3.1 PROJECT MANAGEMENT 19
3.2 DATA PROCESSING WORKFLOW CONCEPTS 20
3.3 WORKFLOWS DIFFER BY PHASE 21
3.4 THE SEAN DELIVERY WORKFLOW 22
3.5 THE SEAN ACQUISITION WORKFLOW 23

CHAPTER 4. DATA MANAGEMENT ROLES AND RESPONSIBILITIES **25**

4.1 GROUP ROLES 25
4.2 DEVELOPMENT ROLES DIFFER MARKEDLY FROM OPERATIONAL ROLES 26
4.3 SPECIFIC STAFF ROLES 26
4.4 SPECIFIC RESPONSIBILITIES OF PARTICIPANTS 27

CHAPTER 5. DATABASES **31**

5.1 DEFINING "DATABASE" FROM A HOLISTIC PERSPECTIVE 31
5.2 DATABASE CONTENT IS STORED IN SPECIFIC REPOSITORIES 33
5.3 IMPLEMENTING THE AUXILIARY REPOSITORY 33
5.4 REPOSITORY ISSUES FACED IN EACH CORE FUNCTION 35
5.4.1 Core Function 1: Information Dissemination 35
5.4.2 Core Function 2: Information Repository 36
5.4.3 Core Function 3: Information Certification 37
5.4.4 Core Function 4: Information Acquisition 38
5.5 DATABASE CREATION WORKFLOW 39
5.6 INTER-ALASKA DATABASE COOPERATION OFFERS BENEFITS 39

CHAPTER 6. DATA ACQUISITION, PROCESSING, AND REPORTING **41**

6.1 THE SEAN'S DATA FLOW 41

6.2 A Distinct Development Phase Precedes the Operations Phase 42
6.3 The Development Phase 42
6.4 The Operations Phase 43
6.5 Legacy Data Acquisition 46
6.6 Data Analysis and Reporting 47

CHAPTER 7. QUALITY ASSURANCE AND QUALITY CONTROL 49

7.1 Practical Quality Assurance 49
7.1.1 Quality Assurance for Paper-Based Tabular Data Acquisition 49
7.1.2 Quality Assurance for Computer-based Data Acquisition 49
7.1.3 Quality Assurance Considerations for Using Data from Multiple Sources in One Project Database 50
7.2 Quality Control 50
7.2.1 Effective Quality Control is Based in Validation Criteria 50
7.2.2 Quality Control for Tabular Time-Series Data Products 51
7.2.3 Quality Control for Non-tabular Products 52

CHAPTER 8. DATASET DOCUMENTATION 53

8.1 Tabular Data Will Comply with the 8-Level Metadata Standard 53
8.2 External Data Considerations 53
8.3 Data Definitions May Be Proscribed by a Partner 53
8.4 Metadata for Non-tabular Products 53

CHAPTER 9. DATA OWNERSHIP AND SHARING 55

CHAPTER 10. DATA DISSEMINATION 57

10.1 Dissemination Is a High Priority Core Function 57
10.2 The Development Phase 63
10.2.1 Create Web Site Objects for Basic Information Delivery 63
10.2.2 Deploy Objects to Production Web Site 64
10.3 The Operations Phase 64

CHAPTER 11. RECORDS MANAGEMENT, DATA MAINTENANCE, AND ARCHIVING 67

11.1 Deliverable Data Products Must Be in Digital Form 67
11.2 Directives Designate Some Project Records as Mission-Critical 67
11.3 Preserving the SEAN Repositories 67
11.4 Local Physical Archiving 68
11.5 "Version Obsolescence" Must be Continually Guarded Against 68
11.6 Archiving Technologies Are Undergoing Rapid Changes 68
11.7 The Value of Monolithic Databases for Housing Repositories 69

CHAPTER 12. PROJECT TRACKING AND DOCUMENTATION 71

12.1 Scope of Project Tracking Responsibility 71
12.2 Tracking is by Spreadsheet 71
12.3 Current Tracking Sheet Format 71

CHAPTER 13. IMPLEMENTATION 73

13.1 Implementation Requires Multiple Passes to Complete 73
13.2 Step One – Establish the Core Structure using Six Vital Signs 73
13.3 Step Two –Establish Mature Data Management Structure 74
13.4 Step Three –Enhance Repository and Dissemination Core Functions 75
13.5 Step Four - Ongoing Operations 75
13.6 Estimated Workforce Composition for Implementation 75

LITERATURE CITED 77

APPENDIX A – DATA FLOW DIAGRAMS A-1

APPENDIX B – STANDARD OPERATING PROCEDURES B-1

Tables

Page

Table 1.1. The SEAN customers and their interests in using the network..... 3

Table 2.1. Facilities required to support the system architecture, by location..... 12

Table 2.2. NPS applications available to store and disseminate vital sign-related information. 13

Table 4.1. Staff requirements by Core Function. Besides the four Core Functions identified in the SEAN's underlying model, management and development functions are also included. 27

Table 4.2. Task responsibilities for each required staff role..... 28

Table 5.1. Target repositories for specific deliverables..... 33

Table 5.2. Possible auxiliary repository architectures considered, in order of desirability. 34

Table 5.3. Specific file types acceptable to the SEAN auxiliary repository..... 37

Table 6.1. Development tasks necessary to fulfill information repository, certification, and acquisition for each project. 42

Table 6.2. Operations tasks needed to fulfill information Acquisition, Certification, and Repository Core Functions. 44

Figures

Page

Figure 1.1. Satellite image illustrating SEAN's geographic domain. 2

Figure 1.2. Map showing SEAN in relation to other networks in Alaska. 3

FIGURE 1.3. CONTEXT LEVEL DATA FLOW DIAGRAM. ARROWS POINTING INTO A BOX INDICATE THAT THE GROUP ACCEPTS INFORMATION FROM THE NETWORK. ARROWS POINTING OUT OF A BOX INDICATE THE GROUP PROVIDES INFORMATION TO THE NETWORK FOR SUBSEQUENT STORAGE AND/OR ACTIONS. 5

Figure 1.4. The SEAN Core Functional Model overview. Dissemination services (red) provide all data products to all customers using Internet web servers. Repositories (yellow) store the certified inventory and monitoring products called for in the protocols. Certification processes (green) assure repositories have the highest quality data and that sensitive items are restricted to authorized users. Data acquisition processes (blue) led by park staff and cooperators include a wide array of tasks, ranging from designing field forms to producing talks on long-term trends. 7

Figure 2.1. The SEAN overall architecture. In operating order, data acquisition (blue) provides deliverables to validation (green). Certified products are installed in the appropriate repository (yellow). Customers find the data using web browsers and servers (red)..... 11

Figure 2.2. Details of the specific equipment and objects the SEAN directly employs to manage its data. Associated functional areas are denoted using the four colors consistently employed throughout the plan. Items in gray, however, have been added to represent development tools. Development tools are essential but are not associated with one single functional area. 15

Figure 3.1. The four stages of SEAN project management within the context of two phases: development and operations. 19

Figure 3.2. The SEAN delivery architecture detail; a partial decomposition of Figure 2.1..... 23

Figure 3.3. The SEAN acquisition architecture detail; a partial decomposition of Figure 2.1. 23

Figure 4.1. The SEAN provider roles are applied to Core Functions. Providers are Partners, I&M WASO, Data Manager, and Project Leaders. 25

Figure 5.1. Depiction of the SEAN's "database" in the broad sense. Also shown is the relationship of the customers to the database. Red indicates the Dissemination Core Function and yellow corresponds to the Repository Core Function..... 32

Figure 6.1. Level 1 data flow diagram illustrating participants, processes, and stores. Participating roles are shown in boxes. The specific data they generate or require are identified by labeled directional arrows. The circles denote high level processes that connect the consumer to the information. The "Data Stores" symbol in the center holds the SEAN broadly-defined database of Chapter 5. 41

Figure 6.2. Sequence of steps involved in the Certification Core Function. 46

Figure 10.1. The SEAN home page content mock-up, illustrating the planned scope. 59

Figure 10.2. Vital sign main page concept example. 60

Figure 10.3. Example of an I&M enterprise application that may be used to provide web access to SEAN information and data: NPS Data Store. 61

Figure 10.4. Example of a possible partner's repository for the SEAN information and data. Shown is a typical web dissemination page maintained by Western Region Climate Center..... 62

Figure 10.5. Results of a localized Google search for SEAN background documents..... 63

Figure 13.1. Estimated SEAN development skill set requirements..... 76

Figure 13.2. Estimated SEAN operations skill set requirements..... 76

Executive Summary

The Southeast Alaska Network (SEAN) conducts and coordinates long-term ecological monitoring for and with Glacier Bay National Park and Preserve (GLBA), Klondike Gold Rush National Historical Park (KLGO), and Sitka National Historical Park (SITK). The SEAN has made a fundamental commitment to rigorous management of monitoring data, so data are available and accessible to a wide array of park and non-park audiences. This Data Management Plan (DMP) is a companion to both the SEAN Vital Sign Monitoring Plan (VSMP [Moynahan and Johnson 2008]) and the Inventory & Monitoring (I&M) National Data Management Plan (National Park Service. 2008). The SEAN VSMP prescribes the approach to long-term monitoring of specific Vital Signs. The national DMP provides and justifies overarching strategy and direction for data management components of the Vital Signs Inventory and Monitoring Program. The SEAN DMP, including the standard operating procedures in Appendix B, provides detailed specifications for establishing and fulfilling the network's information management role.

Introduction

Chapter 1 provides the background of the SEAN, its data management needs, and its foundational model for information management. The plan calls for first determining exactly what data products are needed to meet NPS goals, and then developing processes to deliver those products. It is essential that project protocols specify the data deliverables in advance and in detail.

The SEAN's conceptual foundation for its data management responsibilities is an information model that identifies four Core Functions. Core Function 1, "Information Dissemination," provides all network products to all customers using Internet web servers. Core Function 2, "Information Repository," securely stores the data products defined in the protocols. Core Function 3, "Information Certification," assures repositories have the highest quality data, with sensitive data secured. Core Function 4, "Information Acquisition," is led by park staff and covers everything from designing data collection forms to producing talks on long-term trends. Two organizing principles for the DMP are (1) all certified vital signs data will be web-accessible, and (2) vital signs data will be managed under an "authoritative source" concept.

Infrastructure and Systems Architecture

Chapter 2 describes the process by which an operational architecture is mapped onto the SEAN's data management philosophy. The SEAN will conform to all applications software standards required by DOI and NPS. Information dissemination services are provided to all customers by Internet web servers maintained by Washington Support Office (WASO) and, for some vital signs, external partners. Information repositories are physically located at WASO I&M, outside agency partners, curatorial collections, and the SEAN itself. The Information Certification Core Function uses the SEAN file server facilities for staging data. Information Acquisition processes use park file servers, networked workstations, and field acquisition devices. The SEAN plan details exactly how these architectural elements interact.

Project Management and Data Processing Workflow

Chapter 3 describes a project management model with two-phases: Development and Operations. We describe distinct workflows for both development and operation of each vital sign monitoring project. Development begins at defining end-products and working back to data point observation. Conversely, operation begins with data acquisition and culminates in dissemination. Thoughtful execution of these distinct yet complementary phases will ensure end products meet initial expectations, and that data collection meets the pre-defined needs.

Data Management Roles and Responsibilities

Chapter 4 identifies the roles of personnel and entities involved in the SEAN's data management process. For dissemination and associated repository operations, both WASO I&M and selected partners are used. SEAN staff is responsible for the development side of dissemination, such as building project web pages. Most Information Certification Core Function tasks, including validating that products meet specifications and installing certified data products in repositories, are done by the SEAN Data Manager. Information acquisition is the responsibility of the Project Leader, as is issuing final certification. The Data Manager and Project Leader share responsibility for certain limited tasks.

Initially, more resources are required for development and fewer for operations. Once the initial vital signs projects are established, development needs drop substantially.

Databases

Chapter 5 identifies and describes the types and locations of anticipated SEAN databases. Database contents include both tabular and non-tabular data. The main I&M repository is used to house network data whenever possible. When appropriate, data deliverables should be designed to fit within the constraints of NPSpecies, NatureBib, and other enterprise applications. Tabular data, in particular, should have specifications that allow them to be saved in the I&M's generic data repository, NPS Data Store.

Repository files are periodically refreshed and version-upgraded, so they may be readable after new versions of software get introduced. To meet data accessibility requirements over an extended period, supported file types are limited in number. The I&M NPS Data Store is used to contain all SEAN metadata, regardless of where the authoritative source data are actually housed. Even data sets entrusted to partners have their metadata recorded in NPS Data Store. Tabular data are collected in MS Access databases designed specifically to support the products of the project. The SEAN develops the database objects based on the Natural Resource Database Template (NRDT). The SEAN also programs data entry and edit screens, based on the NRDT "front end builder" tool set. Actual operation of Access databases is the responsibility of the Project Leader.

Data Acquisition, Processing, and Reporting

Data acquisition, processing, and reporting procedures are described in Chapter 6. Data acquisition refers to the collection and organization of field data and the creation of deliverable data products from them. Processing refers to steps that formalize, validate,

redact and certify the proposed products. Reporting includes installation of certified products into repositories. The remaining Core Function, Dissemination, is described in detail in Chapter 10.

Quality Assurance and Quality Control

Chapter 7 summarizes quality assurance and quality control (QA/QC) procedures. Quality assurance steps are taken to encourage accurate and complete data acquisition. These steps include thorough training of collection staff, verification, careful selection of field devices, and others.

Quality control is accomplished by formally validating every data item submitted. All validation criteria are enumerated in detail in advance in the protocol, both for tabular and non-tabular deliverables. Predefined validation criteria include mandatory and optional items. Failure to meet any mandatory criterion, after automated analysis of a submission unit, causes the entire submission to be returned to the Project Leader for corrections. Where applications permit, every row of a repository update is stamped with its date, protocol version, and the user id of the person making the changes.

Dataset Documentation

Chapter 8 outlines documentation procedures intended to attach metadata to all certified vital signs data sets. All data capable of being documented using the NPS-modified 8-level Federal Geographic Data Committee (FGDC) metadata standard have their metadata stored in NPS Data Store. This occurs regardless of whether a project includes a GIS component, or whether the data are actually stored in NPS Data Store.

The network's Internet web site is intended to provide access to considerable documentation for each vital sign including tracked deliverables, exception logs, protocols, abstracts, etc. Ample documentation is provided so data can "stand on their own."

Data Ownership and Sharing

Chapter 9 presents SEAN's principles of data ownership that are integrated into Cooperative Agreements, contracts, and data dissemination processes. Network data are the property of NPS. NPS data are not copyrighted and are in the public domain. Park Superintendents and network staff collaborate on assigning sensitive designations to specific items. Partners do not house sensitive data. The network Data Manager responds to Freedom of Information Act (FOIA) requests for validated data as directed. FOIA requests for draft and incomplete products are serviced by the Project Leader responsible for those products.

Data Dissemination

Chapter 10 presents the steps that the SEAN takes to act on its philosophy that dissemination of high quality products drives all other data management tasks for the network. All deliverables are digital and accessible through Internet web browsing. Each vital sign has its own main page offering the deliverables specified in the protocol. The most useful vital sign information is directly provided by hyperlinks set up by SEAN staff. Customers may also use SEAN's search appliance to browse through historical or

background files to find details of specific individual interest. The "push model" was replaced with a "pull model" of data access.

Data are disseminated only from designated repositories. The repositories contain one and only one copy of a deliverable, which are loaded only after candidate items pass all validations. In this way, SEAN ensures that data always meet the measureable standards defined in their protocol.

Records Management, Data Maintenance, and Archiving

Chapter 11 describes SEAN's approaches to the considerable challenges of managing and archiving data over the long term. SEAN repositories all have the capability to archive digital items. All deliverables must be provided as computer files of an acceptable format as specified in the protocols. Use of I&M enterprise application repositories are particularly attractive because their data are professionally and reliably archived. SEAN-held products are backed up using best practices, and comprehensive permanent-retention backups are made for archival purposes. Version obsolescence of digital data is continually monitored and remediated. In the future, using a comprehensive central database management system may improve the long-term viability of monitoring data.

Project Tracking and Documentation

In Chapter 12 we outline a method for using spreadsheets to track progress in building the specified products. Each project has a tracking sheet, and deliverable tracking is updated only by SEAN staff. The tracking sheet is always available for public viewing as a link on the corresponding project web page. In the future, a database approach may enhance project tracking.

Implementation

In Chapter 13, we propose a multi-step implementation that provides essential functionality first. Subsequent enhancements are added without impacting existing service. The first step develops substantial new processes and is the most time consuming. Operations begins after development is well underway. Operations will eventually become the largest consumer of resources.

The Data Manager is expected to be involved in protocol development to ensure that deliverable data products are articulated, described, and technically defined in detail. The Network Coordinator and Data Manager intend to achieve effectiveness by closely coordinating their efforts.

Acknowledgements

Of particular help to this plan were those very talented NPS data managers who developed the structure and content of what is generally known as the "National Data Management Plan." The SEAN plan also owes particular thanks to the helpful assistance of Margaret Beer, Bill Eichenlaub, Scott Miller, and Dorothy Mortensen; four experts in the field.