The National Park Service
Alaska Region
Inventory & Monitoring Program

Specifications for Electronic Storage of Reports for the Inventory and Monitoring Program for the Alaska Region

National Park Service
Southwest Alaska Network
Inventory & Monitoring Program
250 West 5th Avenue
Anchorage, Alaska 99501

Version: May 2004
DRAFT

File Name: AKRO_2004_ReportElectronicStorage_0405.doc

Recommended Citation:
Alaska Regional Office. 2004. Specifications for electronic storage of reports for the Inventory and Monitoring Program, Alaska Region. National Park Service. Anchorage, AK. 13 pg.

Topic(s):
Administrative, Interdisciplinary

Theme Keywords:
Reports, protocols, guidelines

Placename Keywords:
Alaska, Arctic Network, Central Alaska Network, Southeast Alaska Network, Southwest Alaska Network

Initial Distribution:
Guidelines Website:
Inventory and Monitoring Program, Southwest Alaska Network
http://www.nature.nps.gov/im/units/swan/index.cfm?theme=SWAN_specifications

Table of Contents

Background 4
Review of report 4
Report format 4
Report naming 5
Electronic storage of report 6
Data server organization 7
Post processing of report 7
Acknowledgements 8
References 8
Appendix A: Scanning Specifications 9
Appendix B: Scanning In-house vs. Outsourcing 10
Appendix C: Considerations When Digitizing Reports for Text-Searching Functionality 12
Appendix D: Digitizing Reports 13

Background

Partially in response to the Electronic Freedom of Information Act requirements and partially due to good management practices, the National Park Service (NPS) in recent years has built on its infrastructure to provide information over the Internet. This allows any user to access the information desired. In particular, two functions are becoming part of NPS standard business:

1) the ability to search for information through ANSI Z39.50 compliant or individual project search engines; and
2) the ability to retrieve information independent of format such as documents, maps, and photographs.

NPS staff use these tools every day and have come to expect full functionality.

The Inventory and Monitoring (I&M) Program's Natural Resource Bibliography (NatureBib) centralizes bibliographic information about the natural resource documents within the parks. This has helped in the ability to search on natural resource topics. Each NatureBib bibliography record provides the "metadata" for each report and all the necessary information to link to the electronic copy of the report. These electronic reports are stored in a central on-line library where they may be immediately retrieved on-line from NatureBib and may be linked directly to any website.

Inventory and Monitoring reports should be stored on a stable, secure and manageable Internet image server. These images could be retrieved through either NPS Focus (the NPS Z39.50 compliant search engine), NatureBib, or the individual project on-line database.

Review of report

Inventory and Monitoring Program reports should comply with the I&M guidelines. Once the report has met these guidelines, it is ready to be electronically stored.

Report format

Most reports that are created today are generated on a computer using software like Microsoft Word. These files are easy to convert to .pdf or other formats. Hardcopy reports, however, will need to be scanned, saved as an uncompressed .TIF file, and converted to .pdf. Please see Appendix A, B and C for specifications on scanning reports.

Report naming

Report names guidelines are provided to assist in linking electronic documents to bibliographic records. Bibliographic records can generally be sorted by author name and hence, making the link easier to manage. These guidelines also assist in "parsing" the file names to be used elsewhere, such as in a database. By use of the underscore, a data manager may parse author from year and so on.

Report names should follow these guidelines:
<Author last name><Author first init>_<publication year>_<document short name>_<BibKeyID>.doc (or .pdf)

Using underscores or hyphens as delimiters.

For example:
MillerJ_2003_ALAG_FreshFishInventory_561863.doc

Where,
Author Last Name (primary) = Miller
Author first Initial = J
Year Published = 2003
Park Code = ALAG
Document Short Name = FreshFishInventory
BibKeyID* (from NatureBib) = 561863

* If the document will not be entered into NatureBib, use the date as version control. For example: MillerJ_2003_KATM_FreshFishInventory_0306.doc

Document short name
Most titles of natural resource reports are too long to use as the electronic file name. When selecting the <document short name>, select the significant nouns that best describes the report. It is recommended to precede this with the park code.

Tip!
While a document is being written and is in draft form, the primary author may be unknown and the publication year uncertain. It is recommended the following format be used for this interim version:

<document short name>_<YearMonthDay>.doc

For example: ALAG_FreshFishInventory_0304.doc

The author name and year may be added during the finalizing of the report. This technique may be useful in distinguishing early draft reports from nearly final reports. When the BibKeyID is added and replaces the version date, this distinguishes it as the final report.

Electronic storage of report

The final report should be provided in its raw format, such as Word, to the Network Data Manager or other assigned staff. Process the report as follows:

- Enter into NatureBib
- Rename to appropriate name, if necessary, adding the BIBKEYID as part of the name.
- Convert to .pdf format
- PDF and originals filed in appropriate Network/Region directory
- Submit PDF format to WASO I&M data server
- Link report to NatureBib
- Link report to appropriate website (Network or park specific)
- Send notice to Technical Committee and park staff via e-mail that the reports are available

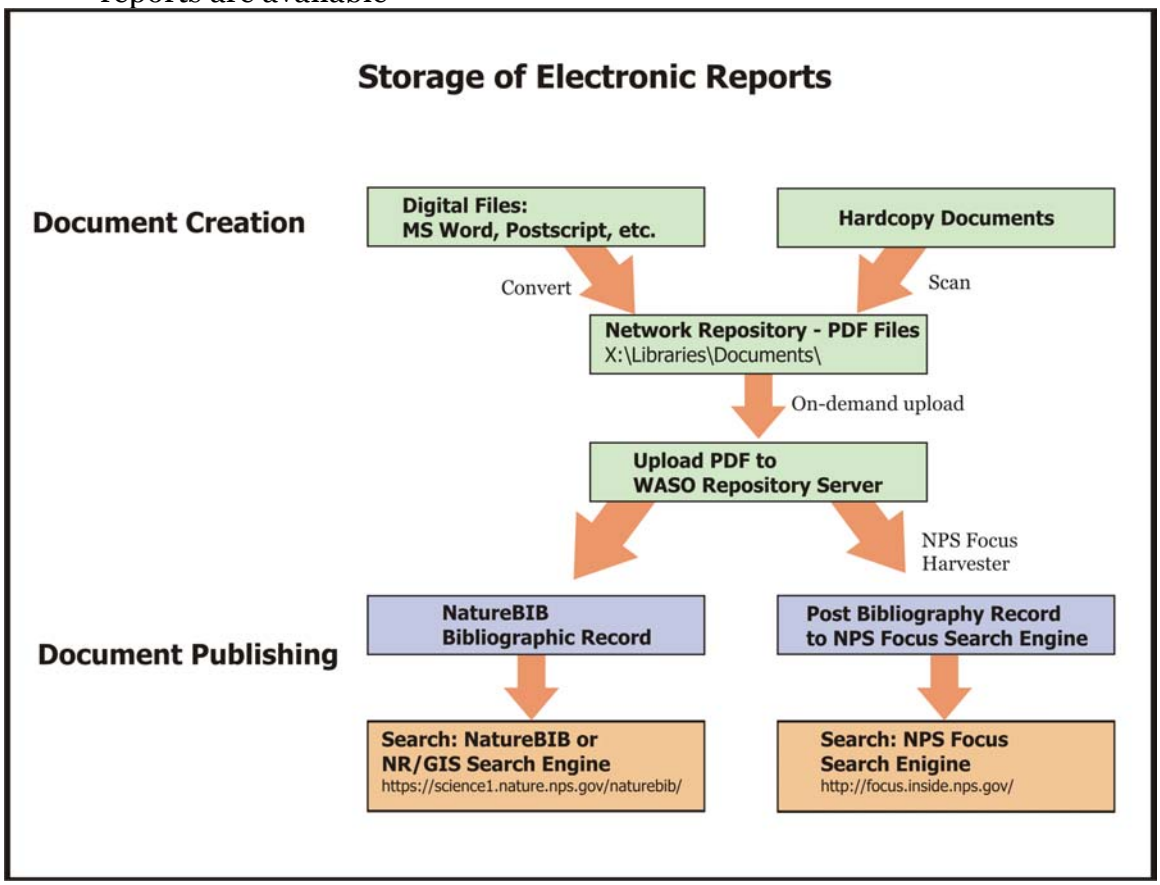

Diagram 1: Storage of electronic reports.

Data server organization

The Alaska regional and WASO I&M data server will be organized in the same manner. This organizational file structure is as follows:

/Network
- /Network – for network-wide documents
- /Park1 – Documents specific to this park
- /Park2 – Documents specific to this park
- /Park3 – Documents specific to this park
- /MISC – Miscellaneous documents that are not network-wide, but may be related to the network. For example, a report on salmon productivity within the Alaska Peninsula area produced by Alaska Department of Fish and Game.

For example:

/SWAN
- /ALAG
- /ANIA
- /KATM
- /KEFJ
- /LACL
- /MISC
- /SWAN

The Network Data Manager, or other assigned staff, will manage these files.

Post processing of report

The report will be further processed to ensure it is accessible via the Alaska Region's distribution process and NPS Focus, see Diagram 1.

Similar to how the GIS data is centralized at the regional office and distributed to all of the parks, the reports will be centralized at the regional office, then distributed to all of the parks. This provides a "local" archive of the reports.

Reports stored on the WASO I&M repository server and recorded in NatureBib will be processed to ensure NPS Focus can find these records. This will allow the user to search through NPS Focus, and retrieve the same information.

Acknowledgements

These guidelines were developed from a series of monthly discussions with a work group from the Inventory and Monitoring Program. Members of this work group are Lisa L. Nelson (WASO), Wendy Schumacher (WASO), Brent Frakes (ROMN), and myself. Kass Evans, from the NPS Focus Digital Library, helped tremendously in providing key information and by testing sample reports for the Inventory and Monitoring Program. I appreciate all of their help in finding a stable repository for NPS reports.

References

Alaska Regional Office. 2003. Interim peer review guidelines for the Inventory and Monitoring Program, Alaska Region. USDI National Park Service. Anchorage, AK. 10 pg.

Alaska Regional Office. 2003. Specifications for annual progress reports and final reports submitted to the Alaska Inventory and Monitoring Program, Alaska Region. USDI National Park Service. Anchorage, AK. 9 pg.

Evans, Kass. 2003. National Park Service NPS Focus Image Creation and Management Guide. Draft. Retrieved January 7, 2004 from the World Wide Web: http://focus.nps.gov/doc/techdoc/ImagingGuide.html

Frakes, Brent. 2003. Using Adobe digital products to enhance data mining. Inventory and Monitoring Data Managers Meeting. Phoenix, AZ.

Inventory and Monitoring Program. 2003. Image server inventory and monitoring program needs statement. National Park Service. Ft. Collins, CO. 3 pg.

Appendix A: Scanning Specifications

Extracted from the NPS Focus Image Creation and Management Guide, by Kass Evans on the NPS Focus Digital Library and Research Station website:

Letter-size documents on typewritten/printed paper being scanned:

- Scan at a resolution of 300 dpi or 400 dpi if the text has very small print
- Use black & white bitmap (do not scan as greyscale or color)
- Save as uncompressed TIF file

Color Pages:

- Choose target size the same as the original
- For color photo, choose resolution 300 dpi, 24 bit color (do NOT use 32 or 48 bit color)
- For black & white photo, choose resolution 300 dpi, 8 bit grayscale (do NOT use 16bit grayscale)
- Save as uncompressed TIF file
- Scan to yield file size approximately: yields file size approximately 4200 x 3000 pixels, file size 20-25MB for color

Appendix B: Scanning In-house vs. Outsourcing

Extracted from the NPS Focus Image Creation and Management Guide, by Kass Evans on the NPS Focus Digital Library and Research Station website:

Scanning In-house vs Outsourcing

The decision on whether to scan resources in-house or outsource the work to a scanning vendor will be an individual decision for each project based on a variety of factors such as:

- availability scanning equipment of appropriate quality for content being scanned
- trained scanning staff (e.g. do they know how to calibrate a monitor?)
- time table - is a quick turn-around time needed?
- physical condition of information resources - does the physical condition require professional quality equipment and services (e.g. the slides are pink, the reports are bound)
- quantity of information resources - the larger the quantity
- funding

Below are some of the issues to consider for Unbound Text, Bound Text, and 35mm slides and negatives.

Unbound Text

Unbound Text: this includes content that can be separated into separate sheets of paper which can be placed face down on a flatbed scanner.

This is the easiest type of material to scan in-house. Since text is generally scanned at 300 dpi, an inexpensive scanner is all that is needed and the staff doing the scanning do not need special training.

However, it should be kept in mind that scanning vendors use high end equipment with sheet feeders that can scan hundreds or thousands of pages a day. Because of this scanning costs can be as low as $ .10 per page. It may be cheaper and faster to use a vendor than doing the same work in-house.

Bound Text

Bound text: this includes all books, journal, and technical reports that have been bound together at the top or left side. Pages cannot be scanned on a flatbed scanner unless the binding is removed or the "book" is forced open and smashed down onto the scanner.

It is highly recommended for this type of material, that the project contract with a vendor that uses an overhead scanner with a book cradle. With this equipment, the book/report is place in a book cradle and the cover is opened only as far as is possible without destroying the binding. The overhead scanner uses archival lighting and each page is scanned in 1-2 seconds so this is the safest method for delicate materials. Special software is used to correct the perspective for materials that are scanned from an angle because of the binding. There are vendors throughout the country that use this equipment but the NPS has had a good experience with Crowley Micrographics who are on GSA schedule.

Do not use a digital camera to take images of pages of typed text because the quality and angle of the image will make it difficult to effectively run OCR (optical character recognition) software on the page in order to index each word for full-text searching.

Appendix C: Considerations When Digitizing Reports for Text-Searching Functionality

To consider whether digitizing in-house is feasible, there are a number of tools that are required:

- High end photocopier – The newer models of photocopies also have the capability to be scanners if they have a TWAIN driver. The photocopies document feeders are far more efficient than any of the desktop scanners. It should be emphasized that the largest bottle-neck of the scanning process is the speed of the form-feeder; a faster form-feeder will be cost-effective.
- Adobe Acrobat 6.0 – It is easy to scan any document, whether on a desktop scanner or photocopier) using the Acrobat interface. Acrobat pulls the scanned images in as one complete PDF document.

Optical character recognition (OCR) is a powerful tool because it can allow a text search through any of the scanned documents. In addition, a word index (i.e., word database) of all documents can also be created and maintained through time. Here are some considerations when using the OCR capability (aka "Capture") functionality in Adobe Acrobat 6.0.

- Documents must be scanned at 300 dpi and in B&W bitmap for Adobe Capture to work efficiently.
- Capture is advantageous because it retains the original format of the document, yet is good at compressing the file sizes.
- Adobe Capture is only about 95% accurate. Some common pitfalls include:
    - Adding spaces between letters,
    - Confusing commas for "l"s
    - Confusing quotes for "ll"s

Appendix D: Digitizing Reports

From the presentation "Using Adobe Digital Products to Enhance Data Mining," by Brent Frakes during the NPS Inventory & Monitoring Program Data Managers Meeting in Phoenix, Arizona in April 2003:

Using Adobe Digital Products To Enhance Data Mining

Brent Frakes
HTLN

Presentation Goals

- Demonstrate Need For Digital Data Mining
- List Tools
- Outline Process
- List Pitfalls and Successes

Need for Digital Data Mining

- Preserve single-copies
- Text search capability
- Reduce storage space
- Disseminate information on web
- Ensure that information can be found/Preserve continuity

Tools

- High-end photocopier that can be used as a scanner
- Adobe Acrobat 5.0 ($27)
- Adobe Capture 3.0 ($700-$2500)

Process

- Digitizing
- Optical Character Recognition (OCR)
- Storage
- Dissemination

Digitizing

- High end Photocopier used as scanner
- Up to 3000 pages per hour
- Controlled by Acrobat
- 300 dpi

OCR

- Adobe Capture
- Completely automated
- Customizable (OCR, Compression, etc.)
- Can run in the background
- ~95% accurate
- Maintains original format
- Text searchable

Storage

- All as Portable Document Format (pdf)
- Can include metadata
- Author.Year.Title naming convention
- Word database/index using Adobe Acrobat 5.0

Dissemination

- Web based
- Three levels of security
  - Internet
  - Intranet
  - Intranet (password protected)
- Includes OCR disclaimer

Recommendations

- Excellent to preserve and organize documents
- Parks have been extremely receptive
- Bottle-neck is scanning – GET A GOOD Photocopier
- Coordinate with region or servicewide for OCR, get GS-5 or 7 to do scanning

Pitfalls and Problems

- Document Bindings
- Copyright Issues
- Cost-Benefit of Information Being Scanned
- Assumption that OCR is perfect, confuses:
  - Justified text split→ S P L I T
  - Quotes, commmas → "Hi" "Hill
  - Small or fuzzy fonts
- Capture is not intuitive