



National Park Service
U.S. Department of the Interior

Natural Resource Stewardship and Science

GRSM Standard Operating Procedure: Project Data Lifecycle

Suggested Citation: National Park Service. 2016. GRSM Standard Operating Procedure: Project Data Lifecycle. Version 2016-07-20. National Park Service, Great Smoky Mountains National Park. Gatlinburg, Tennessee.

Contents

Purpose.....	3
Scope	3
Overview and Categorization of GRSM Data	3
Acronyms and Definitions.....	9
Contact Information	9
References.....	9

Version History		
Version	Update Date	Changes
20160719	20160719	Revised draft to appropriate format

Purpose

The purpose of this Standard Operating Procedure (SOP) is to establish and maintain well-documented guidance for the life cycle of project data. This SOP considers the general work flow characteristics of studies producing data and gives an overview of how those data are generated, processed, finalized and made available.

Scope

These guidelines apply to all GRSM RMS staff that create, maintain, or solicit project data.

Overview and Categorization of GRSM Data

As data are collected for the Park's vital signs and other projects, data may take different forms and be maintained in different places as they are acquired, processed and documented. All data used and maintained by GRSM, regardless of source or origin, must meet certain standards. To meet these standards, all data and data products are subjected to certain processing steps, including data quality assurance and control, data documentation, and data integration into national information management systems.

Specific details of each protocol's data lifecycle may vary depending on protocol implementation and project personnel – these details will be fully described in protocol documentation and supporting SOPs. However, several standard practices will be implemented for all GRSM projects that generate data:

- Data are classified in one of three categories – raw, provisional, and certified. All data migrate from raw to certified data following procedures set forth in park SOPs and protocols. Category descriptions are as follows:
 - o **Raw** – Data in their state of original collection, either in the form of paper or electronic field formed, electronic sensor files, or digital media (photo, video, or audio). These data generally have been collected following some quality assurance (QA) procedures, but have not yet undergone any quality control (QC) procedures.
 - o **Provisional** – Data that have been initially screened for quality to meet minimum standards for generation of provisional information products.
 - o **Certified** – Data that have undergone thorough quality assurance and screening as well as complete documentation.
 - Certified data is the only stage of data that may be distributed outside of the park;
 - Data distributed outside of the park must be listed in the NPS Data Store with appropriate metadata.
- All raw data are archived, intact in their original format.
- Working data sets are the focal point for all modification, processing, quality assurance screening and documentation of data. Working data are stored the current database in the Data/Tabular subfolder (see the Directory Structure SOP). Once the QA/QC review process is completed, that data are flagged as such within the database—this step varies depending on the specific project.
- On a cyclical basis depending on the protocol specific data management plan or upon completion of the project, data will be certified (see the Data Certification Form and SOP). Data certification means that the data steward has confirmed that the data:
 - o have undergone and passed quality assurance checks;
 - o are complete for the period of record;
 - o are fully documented and ready for analysis, archiving, posting, and distribution as appropriate.
- Certified data will be maintained in a “final” database that is write-protected to ensure data integrity over time and stored in the Data/Final subfolder (see Directory Structure SOP).

- Upon creation of the “final” database all data and related documentation will be archived to DVD and integrated into the appropriate data discovery and repository application such as the Data Store, NPSpecies and NPSTORET.
- As a general rule, information products (e.g. maps, charts, graphs, etc.) developed by GRSM will be based only on certified data. These products are also archived and made available to users through appropriate national repositories.
- Provisional information products may be developed using uncertified data in special cases, but will not be released to the public through national systems until such time as they have been certified. There may be certain exemptions from this process for data requests from within the network or with partners – and with FOIA requests. Regardless of the status of data certification, all data that are shared and information products produced, must be accompanied by metadata that describes the QA/QC processes performed on that dataset and a statement describing the level of quality of that dataset.
- Certified data should be considered static and un-editable. Any subsequent changes required to certified data sets will only be done in consultation with the GRSM Data Manager and Project Manager and must be documented in an edit log which is distributed along with the data.

Data Processing and Workflow Steps

A general schematic for the migration of all GRSM data throughout the data lifecycle is depicted below in Figure 1. The migration of raw data to certified data is shown, along with corresponding data dissemination strategies. The personnel responsible for each step are described in Table 1.

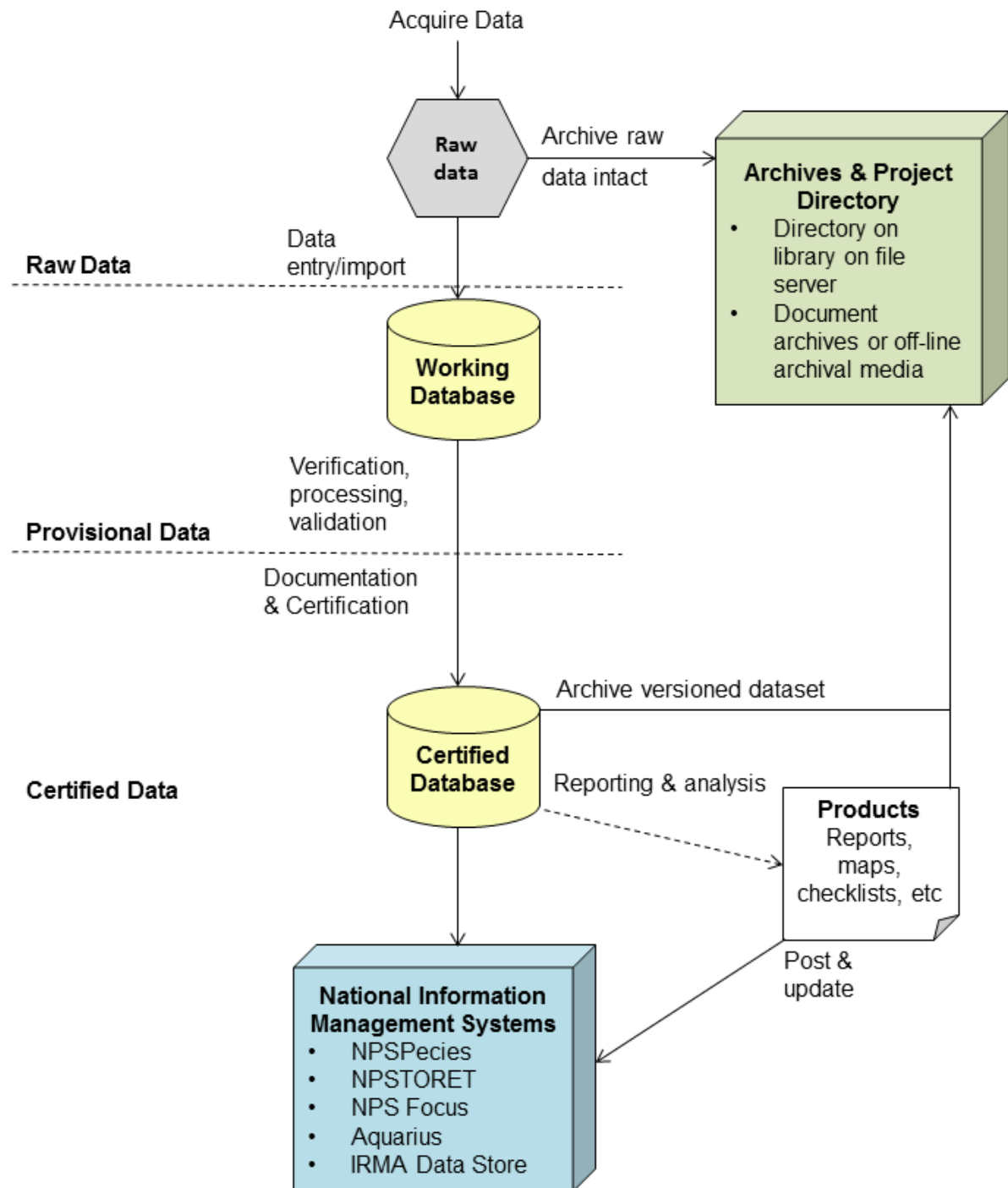


Figure 1. Schematic of the migration of data through the typical data lifecycle

The typical steps involved in data processing and lifecycle are as follows:

- 1) **Acquire data** – For data recorded by hand in the field, data forms should be reviewed regularly (at least daily) for completeness and validity in order to capture errors as close to their origin as possible. Field data collection is the responsibility of the Project Lead and/or technicians. Raw data collected by data loggers or other automated collectors will be archived as soon as is feasible in their original format and integrated into the project directory according to the structure and file naming conventions specific to the project.
- 2) **Data entry / import** – Analog data are entered manually and digital data files are uploaded to the working database. Data entry and/or import is the responsibility of the Project Lead and/or technicians in the field or as soon as possible following the return from the field for protocols using paper field forms.
- 3) **Archive raw data** – Copies of all raw data files are archived intact (prior to any modifications). Digital files are copied to the project archive directory (Data/Tabular/Archive). Hard copy forms are scanned and saved to the appropriate subfolder in the project (Documentation\Datasheets); the original field forms are archived according to the archival SOP. Archiving of hard copy field forms is the responsibility of the personnel collecting the data in the field (e.g. Project Lead, technicians or other GRSM staff).
- 4) **Verification, processing and validation** – The accuracy of raw data transcription is verified by comparison of electronic data to field data forms; data is processed to remove missing values and other data flaws; and data is validated using database queries to capture missing data, out-of-range values, and logical errors. At this point, data sets may be considered 'Provisional'. The QA/QC review of data is highly protocol specific and additional details will be provided in the data management section of the protocol or other supporting documentation. Data QA/QC is the responsibility of Project Lead and technicians as defined in the protocol document. The Project Lead will work with the Data Manager to help identify appropriate QA/QC queries which the Data Manager will then create.
- 5) **Documentation and certification** – Data are considered “certified” when the Project Lead has reviewed the verified and validated data within the read-only “final” database, ensured that the dataset is complete for the time of record, and that complete metadata and documentation have been developed. Creation of the “final” database is the responsibility of the GRSM Data Manager. The GRSM Data Manager and Project Lead utilize the Data Certification Form to ensure that certified data and metadata meet proper standards and are ready to be posted and delivered.
- 6) **GIS Data Documentation and Certification** – Data stewards will provide applicable text for the abstract, summary, and purpose sections of the FGDC metadata template. GIS metadata is in most cases automatically populated. Certified or “Final” GIS products are stored on the “X” drive in a write-restricted folder (X:\GIS_Final\data\basedata\basemap\Data\Final). The GIS office manages all certified GIS data for the park, including distribution of GIS data via approved NPS platforms.
- 7) **Archive versioned data set** – Copies of the certified data and metadata are archived according to the archival SOP. This is the responsibility of the GRSM Data Manager upon completion of the data and metadata review.
- 8) **Reporting and analysis** – Certified data are used to generate data products, analysis, and reports, including automated summary reports for monitoring projects. The analysis and reporting schedule is a key component of the protocol document and will be described in detail there. Reports and other static information products will also be stored in the project directory in the appropriate subfolders according to the directory structure SOP as part of the records management requirements of the program.
- 9) **Post products and data to national databases** – To make data available to the public, products are posted to national repositories such as the IRMA Data Store, ArcGIS Online, NPSpecies,

NPSTORET, and Aquarius as appropriate. Data products may not be posted if they contain protected information about the nature or location of rare, threatened or endangered species, or other natural resources of management concern (see Sensitive Data Handling SOP). The posting of products to national databases is the responsibility of the GRSM Data Manager and/or Project Lead as determined during the protocol development process.

- 10) **Track changes** – Certified data are considered complete and accurate – meeting the QA/QC standards set forth in GRSM protocol documents. Therefore, changes to certified data should not occur very frequently. Should changes be required, all subsequent changes to certified data are to be documented in an edit log which accompanies project data and metadata upon distribution. Significant edits will trigger reposting of the data and products to national databases and repositories. Only the GRSM Data Manager can make changes to certified data.

This sequence of events occurs in an iterative fashion for long-term monitoring projects.

Table 1. Project data lifecycle roles and responsibilities

Roles	Data Stewardship Responsibility
Field crew member	• Collect, record, and verify data • Enter data into project specific database only following thorough training and with the supervision of the field crew leader or Project Lead
Field crew leader	• Supervise data collection and verification by crew members • Enter data into the project specific database and supervise data entry by field crew members • Facilitate training needed for field crew members to effectively manage project data
GIS Specialist	• Create and maintain GIS data standards by communicating the goals of data management to the broader group • Ensure data are organized, useful, compliant, secure, and available with appropriate metadata • Oversee archival of related field documents and resultant reports as appropriate • Develop geodatabases within the National and GRSM conceptual object model framework in collaboration with the Project Lead reflecting project protocol by becoming familiar with how the data are collected, handled, and used • Create a user interface that streamlines the process of data entry, review, validation, and reporting • Develop and implement quality control and quality assurance aspects of protocols by identifying elements that can be built into the database structure to facilitate quality control, such as required fields, range limits, pick-lists, and validation rules • Work with Project Lead to facilitate data display and mapping for different end-user groups • Facilitate the data certification process in collaboration with the Project Lead • Post dataset and documentation to appropriate data discovery platform such as the IRMA, ArcGIS Online, or Github • Ensure spatial data discovery by the public via open-source protocols and Opendata.gov • Maintain records of all datasets created by and or for GRSM by utilizing the Data Store to house this information
Project Lead	• Consult with Data Manager and GIS Specialist as soon as possible when a new project is being developed • Oversee and direct data collection operations following standard operating

	procedures and protocols, including data management • Identify, justify, and document “outlier” data • Work with Data Manager to develop and implement new databases as well as quality assurance and quality control procedures • Identify training needs for staff related to data handling procedures, quality control measure, and database and GIS software use • Coordinate the design of field data forms and other user interface for the project database • Develop certified data and provide documentation to accompany the dataset for posting to the Data Store or other means of dissemination • In collaboration with the Data Manager, identify sensitive information that requires special consideration prior to distribution • Maintain proper directory structure for all project files according the directory structure standard operating procedure • Collaborate with the Data Manager to create data summary procedures to automate transforming data into meaningful information • Identify and prioritize legacy datasets for conversion to desired formats • Ensure that all datasets the Project Lead is involved in have been reported to the Data Manager so that the dataset can be tracked and secured • Increase the accessibility and interpretability of existing natural resources information
Data Manager	• Create and maintain data standards by communicating the goals of data management to the broader group • Ensure data are organized, useful, compliant, secure, and available with appropriate metadata • Oversee archival of related field documents and resultant reports as appropriate • Develop databases within the NRDT and GRSM conceptual object model framework in collaboration with the Project Lead reflecting project protocol by becoming familiar with how the data are collected, handled, and used • Create a user interface that streamlines the process of data entry, review, validation, and reporting • Ensure that database application data entry forms mirror the field data collection form • Develop and implement quality control and quality assurance aspects of protocols by identifying elements that can be built into the database structure to facilitate quality control, such as required fields, range limits, pick-lists, and validation rules • Work with Project Lead to facilitate data querying and reporting for different end-user groups • Ensure that protocol documentation is complete, complies with metadata requirements, and enhances the interpretability and longevity of data collected by and or for GRSM • Facilitate the data certification process in collaboration with the Project Lead • Post dataset and documentation to appropriate data discovery platform such as the Data Store, AGOL, STORET, or Aquarius • Identify and prioritize legacy data for conversion to desired formats • Maintain records of all datasets created by and or for GRSM by utilizing the Data Store to house this information
End Users (e.g., managers, interpreters)	• Interpret information and use information products to inform management decisions • Identify new information product needs and inform the scope and direction of science information product development

Information Technology Specialist	• Provide IT support for hardware, software, and networking
-----------------------------------	---

Keeping track of data throughout its lifecycle is the shared responsibility of everyone involved in a project that collects data within the park. This, in its essence is data stewardship. Data stewardship is a principle of mutual accountability rather than a particular job for one individual. The importance of data stewardship must not be understated—and the degree of success with which it is implemented will have direct bearing on the quality and utility of data and information products developed by GRSM.

Successful data stewardship requires that people involved in GRSM activities learn and understand the expectations for continuous information management and be accountable to perform the duties required to meet these expectation. This requirement is equally important for park staff, as well as contractors or cooperators. All personnel involved in protocol implementation receive training, briefings, materials and additional regular communication about data stewardship from supervisors, Project Leads, Data Manager, and GIS Specialist to promote the appropriate level of understanding about how these efforts relate to park management objectives, NPS and Department of Interior policies, and other federal government requirements.

Acronyms and Definitions

Quality Assurance (QA) procedures: those implemented to ensure collection of data with minimal errors

Quality Control (QC) procedures: those implemented to flag or modify data to ensure that data meet established data quality standards.

Contact Information

Great Smoky Mountains National Park
National Park Service
Gatlinburg, TN

Kendra Straub
Data manager
(865)436-1703
Kendra_Straub@nps.gov

Thomas Colson
GIS Specialist
(865)436-1701
Thomas_Colson@nps.gov

References

Cook, R. R., 2006. Sierra Nevada Network Inventory and Monitoring Program, Coordination of Data Management Roles and Responsibilities. U.S. National Park Service, Three Rivers, CA.

Wright, C. J., J. C. DeVivo, K. Funk, S. McCort, B. Blankley, E. DAVIE, m. w. Byrne, T. Curtis, D. Lagana, and N. Russell, Editors. 2008. Southeast Coast Network Information Management and Archiving Plan Appendices. Natural Resource Report NPS/SECN/NRR—2008/062. National Park Service, Fort Collins, Colorado.