

Standard Operating Procedure #10:

Data Management

Version 2.1 (1/26/2024)

Prepared by

Suzanne Sanders and Jessica Kirschbaum

National Park Service

Great Lakes Network Office

2800 Lakeshore Drive East

Ashland, Wisconsin 54806

Revision History Log

Date	Previous version	New version	Authors
2012	1.0	2.0	Suzanne Sanders and Jessica Grochowski
Figure 1 was replaced; misc. wording not relevant to vegetation was removed.			
2024	2.0	2.1	Suzanne Sanders and Jessica Kirschbaum
Revised to: • make more concise; • reflect the fact that we now use tablets with built-in cameras, rather than paper datasheets and stand-alone cameras; • note that we do not plan on collecting additional GPS coordinates at sites; • improve wording and flow of text.			

Contents

	Page
10.0 Introduction.....	1
10.1 Data Stewardship Roles and Responsibilities.....	1
10.2 Database Design.....	2
10.3 Data Management Procedures	3
Data Collection.....	3
Data Verification and Documentation.....	3
Data Certification	4
Data Distribution	4
Policy and Protected Data Standards.....	4
Data Analysis and Reporting.....	5
10.4 Literature Cited	5

10.0 Introduction

Data collected under the vegetation monitoring protocol (Sanders and Kirschbaum 2024) must be quality-checked, managed, and made available to others for a variety of purposes. This standard operating procedure outlines data stewardship responsibilities and provides specific instructions and references for managing vegetation data.

10.1 Data Stewardship Roles and Responsibilities

The purpose of data stewardship is to ensure that data are organized and compliant, among other characteristics. The demand for detailed, high-quality data and information about vegetation requires a group of people working together to ensure that data are collected using appropriate methods, and that resulting datasets, reports, maps, and other derived products are well managed.

The Great Lakes Network (GLKN) Terrestrial Ecologist serves as project manager for vegetation monitoring. The Ecologist, along with the Botanist, will supervise data collection, provide project oversight, direct on-the-ground data collections, and provide cohesive links among data collection, synthesis, interpretation, and reporting.

While the Ecologist and Botanist act as the primary stewards for the network's vegetation monitoring data, other project and GLKN personnel are also accountable for specific data management tasks. Table 1 lists stewardship responsibilities of personnel involved in the management of vegetation data. To ensure that all project data are managed properly, individuals must understand their responsibilities, communicate with one another, and assist one another as needed.

Table 1. Data stewardship responsibilities of vegetation monitoring personnel.

Personnel Role	Data Management Responsibilities Related to Vegetation Monitoring
Project crew member	Crew members collect and manage data with direction and guidance from the Botanist and Ecologist. Data collection includes obtaining GPS coordinates, taking digital photographs, and recording observations. Crew members are responsible for quality control by following data collection and recording instructions.
Project manager (Network Terrestrial Ecologist)	The project manager is responsible for all project operations and results and may also participate in field operations. The project manager ensures that data management activities are conducted according to established procedures and is responsible for data validation: approving the data content, quality, and documentation, as well as making decisions about data sensitivity and distribution. The project manager is responsible for evaluating project data at specified intervals, analyzing data for trends, and following reporting requirements.
Project crew leader (Network Botanist)	The crew leader normally performs many of the same duties as the project manager and ensures adherence to data collection and processing protocols, including data verification and documentation. The crew leader also works with the project manager on vegetation data management in the office.
Network Data Manager	The network data manager is the primary contact for database assistance, including tasks related to creating and running data packages for certification and any routine maintenance on the database.

10.2 Database Design

The vegetation monitoring database is a complex MS Access database composed of a front-end and back-end. The front-end contains a user interface portion along with various queries, forms, reports, and visual basic code for underlying function of the database as a whole and use with the user interface. The user interface consists of various forms to help the user enter and validate data, along with the ability to query, through forms, the data for various QA/QC checks and use in final reports. The back-end database consists of numerous lookup and data storage tables, which are linked to the front-end database. As of 2021, data are collected digitally in the field and imported, via a series of routines, into the production database following the completion of the season.

Digital data collection in the field is central to this protocol for ensuring data quality. Digital data collection allows quality assurance checks to occur as data are being collected, reducing the occurrence of errors and omissions. For digital data collection to be feasible, the user interface for data entry must be intuitive, flow with the typical progression of sampling, and include comprehensive quality assurance checks that ensure all required fields are complete and within acceptable ranges (where definable). Because this protocol is typically implemented by two or more teams on separate plots, the database for this protocol must also have a merge function that can accurately pull the data collected by each team into the official database.

To accommodate the need for a user-friendly field application and a merge tool, database functions for this protocol are divided between two Microsoft Access front ends, namely the Office App and the Field App, and two SQL Server instances, namely the Production SQL Server database and a local SQL Server Express instance that connects to the Field app.

The Office App is where data management tasks primarily occur, including generating field .bak files, merging field .bak files, adjudicating editable historic records, updating lookup tables, and modifying historic records. Note that the Office App was not optimized for data entry and cannot accommodate new records being added.

The Production SQL Server database (hereafter Production Database) is a cloud-based SQL Server database and is the official backend database for the protocol. Access to the Production Database must be granted through a user's Active Directory account and requires a NPS network connection. To permit data collection in the field, a field .bak file is created using the Office App. The field .bak is an exact copy of the database elements necessary for field data collection from the Production Database at the time it is generated. Field .baks allow a user to have a local instance of the backend database without requiring the special permissions and network connection needed to access the Production Database.

The Great Lakes Network will maintain the master copy of the vegetation database at the main office on a central server. Additional local copies of the database may be used by GLKN personnel stationed at parks, or teleworking, but those copies should only be used as a conduit for accessing data and not modifying the database.

10.3 Data Management Procedures

Data Collection

Data are collected on tablets in accordance with the GLKN Forest Vegetation Monitoring Protocol (see SOP 6: Field Methods and Data Collection) on data collection forms created and maintained by GLKN staff in ESRI Survey123. Crew members are responsible for following the guidelines as spelled out in the protocol (see SOP 6: Field Methods and Data Collection) and on which they were trained. For many of the variables, the data collection form will not accept out-of-bounds data, and immediate correction is mandatory. This built-in feature is the first level of safeguarding, to ensure accurate data.

Digital images of sample sites are acquired during site establishment (SOP 6: Field Methods and Data Collection) and at every visit thereafter. Photos are taken using a data collection form in ESRI Survey123; crew members are responsible for ensuring the proper settings are applied. Photos are ultimately stored on network servers, which are backed up regularly.

Because all sites have been sampled as of Fall 2022, we have a full set of GPS coordinates and do not anticipate collecting additional values.

Data Verification and Documentation

Data collected on tablets in parks will be sent to the NPS GIS Online Portal (<https://gisportal.nps.gov/portal/home>), generally at the completion of an eight-day sampling round. At the end of the field season, the data stored on the NPS GIS Online Portal are exported to MS Excel files using ESRI ArcGIS Pro via a Python Notebook. Any attachments in the data (i.e., photos taken in the ESRI Survey123 data collection forms) will be exported as part of this process. At this stage, the data goes through the first QA/QC process, which is R code developed specifically for the vegetation protocol to look for any missing or erroneous data, such as total number of plots for each of the data types, missing transects, extra transects, and the wrong sampling type. Once the data has passed the first QA/QC process, it is ready for import into the MS Access Database. There are a series of import queries (over 25) and append queries (19) built to facilitate this process, and they must be followed in a step-wise manner to ensure the import process works correctly. After data are imported into the database, a suite of QA/QC procedures exist that compare the entered data with expected formats and accepted data value ranges or domains.

Data verification starts with the QA/QC steps that are detailed in SOP #15: Quality Assurance/Quality Control.

Metadata for all Analysis Datasets will be created using the Ecological Metadata Language standard (EML). The R package EMLassemblyline will be used to automate the creation of metadata for inclusion in all published Data Packages. EMLassemblyline processes a series of CSV files that outline all the relevant metadata for each data package and generates an XML metadata file. Wherever possible, the metadata included in the source CSV files will be extracted directly from the Master database.

Data Certification

Prior to distributing the data for any type of use, the project manager certifies the master vegetation monitoring database file. The first step in this process is data verification. Verification does not indicate that the data are ready for distribution, only that the process of having the quality of the data assured and certified is underway. After data verification, a number of quality control and quality assurance measures are taken and the data are certified. Data certification is a benchmark in the project information management process that indicates that the data are complete for the period of record, they have undergone and passed the quality assurance checks, and that they are appropriately documented and in a condition for archiving, posting, and distribution as appropriate. Certification is not intended to imply that the data are completely free of errors or inconsistencies.

If this certification does not meet scheduled reporting requirements of the network, reports can include a statement explaining that results are based on data that have not been certified. The statement should include an explanation of what significance this has for using preliminary data.

Data Distribution

GLKN has a number of different avenues for data distribution depending on the audience served and the degree of analysis and customization needed by the end-user. Individual data requests can be accommodated by developing queries to produce reports of the desired information. All individuals to whom data is supplied must abide by a data use agreement, which notes, among other things, that they only use the data provided for the use specified and that the appropriate GLKN staff are acknowledged in any end product.

Policy and Protected Data Standards

According to federal policy, data collected by GLKN are to be made public. However, there are limited cases in which data may be withheld from release to protect park resources. Dissemination of certain information could put particular resources at risk, and some cultural and natural resource information is protected from public disclosure by resource confidentiality laws. These laws prohibit agencies from releasing protected information except in limited, specified circumstances.

The process of identifying protected data, including which data are protected, the degree of protection warranted, and if/how data will be disseminated was completed by all networks and signed by park superintendents in 2019.

As a general rule, GLKN provides the public with as much monitoring data as possible to facilitate communication and collaboration. GLKN therefore does not generally protect location data, such as coordinates of monitoring sites. However, a number of species that are rare in the region and/or vulnerable to poaching could potentially occur in forest plots. The locations of these species, if found, are redacted from all publicly available datasets. The redacted data are available upon request, but all requests must be reviewed by network and the affected park staff to ensure that providing these data will not result in unwanted resource impacts.

Data Analysis and Reporting

Project staff will follow the procedures in SOP #11: Data Summary and Analysis and SOP 12: Reporting.

10.4 Literature Cited

Hart, M., and U. Gafvert (editors). 2005. Data management plan: Great Lakes Inventory and Monitoring Network. GLKN/2005/20. National Park Service, Great Lakes Network Office, Ashland, Wisconsin.

Sanders, S. M., and J. Kirschbaum. 2024. Vegetation monitoring protocol, version 2.1: Great Lakes Inventory and Monitoring Network. Science Report. NPS/SR—2024/166. National Park Service. Fort Collins, Colorado. Available at: <https://doi.org/10.36967/2305281>.