Interim Alaska Region Data Resource Management User Guide

Alaska Regional Data Stewardship Team

Version 2

August 30, 2019

Table of Contents

Background

The Alaska Region of the U.S. Fish and Wildlife Service is committed to improving the management of its data resources. A comprehensive, long-term plan to develop and implement a Regional Data Management System (RDMS) will begin in the winter of 2019-2020, with the expectation that it will be fully operational and in use in 3-5 years. In the interim, there is a strong desire in the Region to begin managing existing and new data resources with systems and staff capacity available now.

Data is appropriately managed when it is documented, secure, discoverable and accessible. Documentation (i.e., metadata, or data about data) provides sufficient and relevant information so that users can understand, interpret, and use all data and derived products without additional guidance. Security procedures prevent loss and ensure data integrity. Discoverable data is readily found (e.g., using a data catalog), whereas accessible data is readily obtained (e.g., downloaded from a website). Together, these qualities ensure that data resources can be effectively and efficiently used both within the Service and beyond.

This document describes the data management activities necessary to build an archive folder for a project. An archive folder represents the definitive version of a project's output. Data resources within the archive folder are the subject of metadata records and these resources, supported by their metadata records, are the authoritative versions intended for long-term storage, all subsequent workflow, analysis, products, and sharing to outside partners.

Although there are areas of overlap, this document does not specifically address best practices in managing data resources throughout the course of a project's implementation (i.e., project workflow). Project workflow should be structured to maximize data integrity from creation or collection to the final archive described in this document. Project workflow is outlined in a data management plan (DMP), covering issues like field note to digital conversion, version control, and reproducible analyses. A Regional DMP, which can be customized for specific Program and project needs, will be developed separately.

The audience for this guide are data management early adopters, and their supervisors, in Fisheries and Ecological Services (FES), Migratory Bird Management (MBM), National Wildlife Refuge System (NWRS), and the Office of Subsistence Management (OSM). **This guide is a living document (working draft) and will be revised to reflect the ongoing development of available systems and feedback obtained from users.** The procedures and processes outlined here, while following best practices, are limited by current available resources and organizational structures. It is anticipated that many activities necessary in the interim will be greatly streamlined in the final RDMS.

This document is organized into four sections. **The Big Picture: Integrating Data Management with Project Management** put data management activities into the larger context of project management and provides guidance on determining which products should undergo data management. The second section describes the **Four Fundamental Activities of Data Management**: **Establish Roles and Responsibilities**, **Quality Management**, **Security and Preservation**, and **Documentation**. Each activity is described and implementation guidance is given. **File Organization and Best Practices** describes a

recommended archive folder structure to use for every project and best practices in naming conventions to use in the file structure. Best practices for the Region's most common types of data are given special consideration.The **Sharing** section is under development, but will describe how metadata records and associated products can be shared outside of the Region's network drives. Appendices include a data management quick guide (**Appendix A**), a tree diagram of the archive folder structure (**Appendix B**), and examples of untidy and tidy data (**Appendix C**). A glossary is provided to define data management-related terms used in this document.

This guide is supplemented by the Alaska Region Metadata Guide. mdEditor (mdEditor.org) is a web-based application that will be used to create archival-quality metadata for projects and products.

The Big Picture: Integrating Data Management with Project Management

There is a distinction between project management and data management (**Figure 1**). Project management is the application of knowledge, skills, tools and techniques to oversee a project to completion in the desired time and to a specified quality. Data management is concerned with the handling of data to ensure long-lasting integrity and usability. **Figure 1** shows the relationship between the project and data management life cycles. Data management is most effective when fully integrated into project work flows, project oversight, and staff supervision (i.e., project management). The purpose of this document is to provide knowledge and guidance for the data management aspect of projects. Guidance on project management is beyond the scope of this document.

Figure 1: The Big Picture: integrating data management with project management.

<table>
<tr><th colspan="2">Project Management Life Cycle</th><th colspan="2">Data Management Life Cycle</th></tr>
<tr><td>Framing</td><td>Define the Problem, Identify Management Context, Define Objectives</td><td colspan="2" rowspan="3"></td></tr>
<tr><td>Prioritizing</td><td>Identify Priority Actions to Address the Problem (via Conceptual Model, Results Chain, VOI, Program Criteria)</td></tr>
<tr><td colspan="2">Project Idea Approved</td></tr>
<tr><td>Design</td><td>Determine Approach, Protocols, Data Model, Analysis</td><td>Plan</td><td>Adopt or Adapt the Regional Data Management Plan

Establish Roles and Responsibilities

Set Up the Archive Folder and Establish Security

Initiate the Project Metadata Record

Document Quality Assurance Procedures</td></tr>
<tr><td colspan="2">Full Proposal Approved (if appropriate)</td><td colspan="2"></td></tr>
<tr><td rowspan="3">Implement</td><td rowspan="3">Run Project

Produce Data Resources (Raw Data to Final Report and Everything in Between)</td><td>Acquire</td><td>Perform Quality Control</td></tr>
<tr><td>Process</td><td>Update Archive Folder</td></tr>
<tr><td>Analyze</td><td>Initiate Product Metadata Records and Continually Update</td></tr>
<tr><td colspan="2"></td><td>Archive</td><td>Complete Metadata Records</td></tr>
<tr><td>Learn</td><td>Evaluate Data Resources</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Take Action</td><td>Management Decision, Revisit a Life Cycle Step</td></tr>
<tr><td colspan="2" rowspan="3"></td><td colspan="2">Approve Metadata Record and Data Resources</td></tr>
<tr><td>Publish</td><td>Metadata Record to Data Catalog</td></tr>
<tr><td>Share</td><td>Data Resources in a Public Repository (if approved)</td></tr>
</table>

Definition of Project and Product (aka Data Resources)

For the purposes of the interim plan, priority will be placed on projects and their derived products. Projects are discrete efforts on a particular topic with defined objectives or goals. In the interim, it is at the program's discretion to determine which projects (completed, ongoing, or proposed) should undergo data management. Use in management decisions and partner needs may be two selection criteria managers could adopt. The Alaska Region has historical and ongoing projects that can and should undergo data management; however, it may be necessary to reconstruct data management activities for these projects.

Products (aka Data Resources) are recorded information and can be generated by experiments, models, simulations, observations, analysis, and other activities that create or synthesize data resources. In a new project, anticipated products should be identified during the planning stage when the design of sampling and analysis take place in consultation with the program's biometrician. Which products are documented is at the discretion of the principal investigator, their supervisor and any relevant program or branch policies but the decision should be guided by the principle of reuse. For example, a tabular dataset that is intended to be used again should undergo data management and documents, such as reports, presentations, and peer-reviewed papers, should be managed to be discoverable and associated with the datasets and code used to produce results.

Four Fundamental Activities of Data Management

There are four fundamental activities of data management (Establishing Roles and Responsibilities, Quality Management, Security and Preservation, and Documentation).

Establish Roles and Responsibilities

It is critical that there be an explicit agreement on individual roles and responsibilities in terms of data management and documentation. It is also important that others know who to contact to obtain more information about projects and products or the source data.

Explicitly identify a person or position in the following roles:

1) Project Manager: this may be synonymous with the project's principal investigator. Responsible for the management of the project and all associated data products. Responsible for ensuring that the project is performed as described in the project and data management plans. If a particular required role is not explicitly identified, it is the responsibility of the Project Manager to perform the necessary tasks. Responsible for project metadata creation.
2) Data Originator(s): person(s) generating/collecting data. Responsible for data they collect, author or generate. Responsible for following best practices for the data type. Responsible for product metadata creation.
3) Data Custodian: person responsible for the management of the project archive folder.
4) Data Steward: subject-matter expert(s) responsible for reviewing and approving the data products, including documentation. Responsible for building the project-product associations between metadata records.

5) Data Trustee: an upper level position in the organization that has ultimate responsibility for ensuring that adequate resources (e.g. staff and funding) are allocated to allow for completion of all aspects of data management. The trustee also has ultimate responsibility to ensure that governance policies are applied to the project and resultant data resources. The trustee will most likely not be involved with data management, but rather with data governance.

Note that one individual may fill many roles. In many cases, the Project Manager may serve in all roles with the exception of Data Trustee. It is best practice for the Project Manager and the Data Steward to be different individuals, but project staffing levels may not allow for this.

Quality Management

Data quality management is composed of quality assurance (QA) and quality control (QC). Quality assurance begins before the data are collected and are procedures used to prevent errors from entering the data. Quality control is the discovery and correction of errors in the data and generally occurs during or after data collection (e.g., detection of outliers, typographical error, a character datum where a numeric value is expected, using an incorrect species code, and etc.). Quality control should occur as soon as possible after collecting the data and before submitting data to the archive record or sharing. QA and QC procedures should be identified during the project planning phase in consultation with the program's biometrician and/or data manager. Record quality management practices (QA and QC) in the mdEditor for all documented products.

Best Practices in Quality Assurance

- Use documented protocols and standard methods
- Use high-quality instrumentation and regularly check accuracy
- Provide consistent training
- Develop standardized data collection forms (data sheet templates or computer input with data validation formats)

Best Practices in Quality Control

- Inspect data values using summary functions (tabling unique values, calculating means and variances, etc.) or by applying complex analysis algorithms.
- In Excel files, use sort and filter functions to look for data anomalies or outliers.
- Visually inspect data using scatterplots, regressions and histograms.

Security and Preservation

Establishing one archive folder for each project prevents data loss and ensures data integrity. The archive folder for a project, comprised of the recommended file structure (**Appendix B**) and described in File Organization and Best Practices section, will be placed on your program's network drive. The Data Custodian (see Establish Roles and Responsibilities) will have read/write permissions to all folders in the archive folder. All other users with access to the program's network drive will have write permissions to an "incoming" folder and read-only permissions to all other folders. The incoming folder allows anyone with access to the program's network drive to contribute products and associated metadata records to

the archive folder, the Data Steward is responsible for reviewing these files and the Data Custodian is responsible for moving them from the incoming folder to the appropriate folder in the archive folder.

Anyone with access to the program's network drive can copy the archive folder to their local hard drive and maintain the files on their hard drive in whatever manner they choose, but may only move things into the archive folder via the incoming folder.

Although the archive folder is on the network drive to allow broad access to one, definitive version of the project and its data, the network drive is not currently managed by IT as an archive. Each individual designated as a Data Trustee (see Establish Roles and Responsibilities) for at least one project will have an external hard drive at their workstation that is set up to perform daily backups of all projects that they administer. For example, the Branch Chief of a program will have an external hard drive at their workstation that backs up the archive folders of all projects in their program.

It is more broadly recommended that all workstations have an external hard drive set up to perform daily backups of its hard drive as well.

Documentation

Metadata, or data about the data, will be written using the mdEditor (mdEditor.org). mdEditor is a web-based application developed by the Alaska Data Integration Working Group (ADIwg) that allows data-generators to write archival-quality metadata without specialized technical knowledge. mdEditor JSON files will be exported to the incoming folder and ultimately saved to the project archive folder in the metadata folder.

It is highly recommended that new users first work through the mdEditor Tutorial to become acquainted with the layout and functionality of mdEditor. Step-by-step guidance on writing documentation can be found at the Alaska Region Metadata Guide and will be supported by hands-on workshops.

As a quick checklist, information to gather for projects and products includes:

- Title
- Roles and responsibilities
- Contact information for project staff
- Abstract
- Start date
- End date
- Keywords
- Focal species in the project
- Spatial extent
- Data dictionaries (for some product types)
- Protocols used in the project

Sharing

TBD (*guidance on publishing records/data to ScienceBase, ServCat or other*)

File Organization and Best Practices

There is a distinction between working folders and archive folders; development of the latter is the focus of this document. Working folders often exist on an individual's computer hard drive and are used to collect, organize, and analyze products during the course of the project. In contrast, archive folders are used to store final products that are meant to be discoverable and accessible over a long period of time. The products in the archive folder are the subject of metadata records.

Consistent archive folder organization across projects and programs allows the file creator, collaborators, supervisors, and our future colleagues to find relevant documents associated with a project quickly and understand how documents relate to one another. A tree structure describing the recommended file structure is presented in **Appendix B**.

When a new project archive folder needs to be initiated, contact your Data Stewardship Team (DST) member. They will set up the folder structure on your program's network drive and apply the appropriate read/write permissions with IT staff (see Security and Preservation section). The folder will be named using a short acronym for your program, followed by a three digit number, followed by the short title of the project (e.g., FRP_001_BarrowEider where FRP stands for Fairbanks Recovery Program). Access to this record and backup practices are discussed in the Security and Preservation section of this guide.

Best Practices in Naming Conventions

When naming the project folder (short title), files within, and even variable names (column headers in spreadsheets) there are some best practices in naming conventions to keep in mind.

- Keep names short, but meaningful.
- Use ISO date format: YYYY, YYYYMM, or YYYYMMDD. It ensures that files with the same name and different dates are sorted in order.
- When using personal names, give the family name first, followed by the first name or initials (e.g. SmithMary or SmithMC).
- Use only letters, numbers, dashes, "-", and underscores, "_". **Do not** use spaces or any other characters.
- Avoid using "draft", "version", or "final" in file names. Use date (in ISO format, see above) to distinguish versions.

There are a few very common types of data generated by Alaska Region projects. These are tabular data, databases, geospatial data, collections of similar types of data (digital images, data logger outputs, etc.), and source data (data not generated by the project but used during the course of the project. For example, data used to inform project design). Best practices for organizing these types of data for the archive folder are described below. More specific guidance and trainings will be developed for each data type in the future.

Guidance is also provided on converting data formats produced from commonly used programs (e.g., MS Office or ESRI ArcGIS) into preferred open formats. These formats are preferred for archive folder purposes because they are independent of any particular software program. That means that they can

be opened by users with a variety of applications and that they are resilient to software application upgrades or obsolescence that can degrade proprietary formats over time.

Best Practices in Tabular Data

Much of the Region's data is contained in spreadsheets (i.e., tabular data, most commonly MS Excel files). Multiple spreadsheets within the same file can contain data and derivatives of the data (tables, summaries, pivot tables, formulas, figures). There are some best practices to use when dealing with this type of data. Examples of untidy vs tidy datasets are given in **Appendix C**.

- One sheet in the file should contain only a clean version of the data. **Nothing else** should be on this sheet. This is sometimes referred to as tidy data. In tidy data:
    - The first row contains variable names, each column represents one variable. Variable names should use only letters, numbers, dashes, "-", and underscores, "_". **Do not** use spaces or any other characters. The variable name should include the unit where this is relevant (e.g., length_cm and weight_gm).
    - Each row after the first row should represent one observation.
    - Avoid formatting information in this sheet (e.g., comma in the thousands place, font settings, border lines, colors, etc). If the formatting is there to convey some information, consider adding a new variable to record that information instead.
    - For the purposes of tidy data, blank cells indicate that the data point is missing. "0" in a cell means that the data point was collected and it was "0". Deviations from this convention should be recorded in the data dictionary (see below).
    - The tidy data sheet, in addition to being part of the workbook, should also be saved in an open format (e.g. Text or CSV) using the same name as the Excel file (e.g., fish_data.xlsx and fish_data.txt) in the same archive folder as the Excel file.
- One sheet in the file should provide a brief description of each variable in the tidy data. Each row of this sheet represents one variable. This is termed the data dictionary and will be part of the metadata record for the tidy data. An example and a template for the data dictionary is available here.
- Other sheets in the file can contain summaries of the data (pivot tables), graphical representations of the data (figures), or derived quantities from the data (formulae, macros, etc). **Avoid** including metadata on these sheets.
- One sheet in the file should provide a brief description of each sheet in the file (what does it contain, any relevant information about its use). Each row of this sheet represents one sheet of the file. First column is the sheet name, the second column is the sheet description.
- Save the original workbook in the most recent format supported by the application. For example, save Excel files in .xlsx format rather than .xls format.

Best Practices in Databases

Databases are essentially a collection of tidy datasets with relationships between the tables specified. Generally, but not exclusively, databases in the Region are developed using MSAccess.

- Variable names in each table should be described (i.e., a data dictionary is available for each table). When possible, the database should be documented within the database application (e.g.

from within MS Access, add title and abstract to database properties and add description for all tables and fields)

- Constraints should be enforced on variables to promote Quality Assurance (see Quality Management section). For example, in a variable named "Gender," inputs could be constrained to the values of "Female," "Male," and "Unknown." Or, in a variable named "Length_mm," only integers between 10 and 1000 could be made allowable values.
- If possible, consider converting MSAccess databases to SQLite, an open format that will preserve the database functionality. Utilities are available to assist with this, but may require additional technical assistance. Contact your DST member if you are interested in this conversion.
- If conversion is not possible, MSAccess tables and their data dictionaries should be exported to a preferred open format (e.g. Text or CSV) and the database structure (relationships diagram in MSAccess) should be saved to a preferred open format (e.g. JPEG or PNG) throughout the duration of the project and at the completion of the project. These files can be saved in the same archive folder as the database.

Best Practices in Geospatial Data

Geospatial data are often stored in a complex proprietary format (e.g. ESRI geodatabase) that is extremely difficult to archive for long-term preservation and access. A single geodatabase can contain many individual data sets. Each individual data set, within a geodatabase, typically consists of multiple individual files (in proprietary formats) that record the spatial information, attribute information, and other essential database properties required to use the data. The complexity of the geodatabase (e.g. a few individual feature classes or many feature classes with related tables and attribute domains) will determine the best methods to use when creating open data formats for archival purposes. It is recommended that you consult with your program Data Manager and/or GIS Analyst prior to archiving geospatial data. Broad steps for managing geospatial data include:

- Fully document each individual feature class or shapefile.
- Store individual geodatabases and shapefiles in the archive folder directory structure as with other data types.
- Convert the geodatabase or shapefile to an open format (e.g. GeoPackage) and store in the archive folder.

Specific guidance relevant to this data type is currently under development.

Best Practices with Collections of Similar Types of Data

Large batches of data are often collected in a project via cameras (e.g., from a remote camera taking time-lapsed or motion-sensitive images or from an aerial survey), data loggers, etc. These batches, called a "collection" in metadata, should be saved in a single archive data folder with an accompanying text file describing relevant information for each file (location, equipment, resolution, etc.). In the case of photographic images, information for each file may be extracted from the metadata embedded in the image itself. The collection can be documented with a single metadata record.

Best Practices with Source Data

Source data refers to those data resources that were used by the project but were not created by the project. Source data is commonly used as input in the creation of new data products. Examples include base layers used in GIS processing or sensor input used to generate analytical output. If the source data is discoverable and permanently accessible through another means (e.g. USGS Streamflow data), this data does not need to be maintained in a project archive folder. However, if the source data is not readily available *in the form used by the project* it would be appropriate to save that information in the archive folder by following the best practices for each data type as described in prior sections of this guide.

In either case, a metadata record for the source data, obtained from the original source or written in the mdEditor, must be in the source data folder of the archive. The relationship between source data and products are described in the Lineage section of the product metadata record.

Appendix A: Interim Data Management Quick Guide

NOTE: For *Completed* projects, the Data Steward will often perform the actions assigned to the *Data Originator* and *Project Manager* when those roles are unable to be filled. Whenever possible, the *Data Originator* should author product metadata.

Task	Performs	Approves	Guidance
☐ Establish roles and responsibilities	Project Manager	Data Trustee	Link
☐ Create the archive folder and establish security and access controls	Data Custodian	N/A	Link
☐ Establish procedures among project members for adding products to the incoming folder, review, and accepting files to the archive	Project Manager	Data Steward	
☐ Identify anticipated products from the project and assess any sharing constraints	Project Manager	Data Steward Data Trustee	Link
☐ Create the project metadata record	Project Manager	Data Steward	Link
☐ Use data type-specific best practices to prepare products for the archive	Data Originator	Data Steward	Link
☐ Write product metadata records and export records to the incoming folder	Data Originator	Data Steward	Link
☐ Review products for best practices and metadata for correctness and completeness	Data Steward	N/A	
☐ Revise products and/or associated metadata, when necessary	Data Originator	Data Steward	Link
☐ Document QA/QC procedures used for the data products	Data Originator	Data Steward	Link
☐ Document any dependencies between source data and project-derived data	Data Originator	Data Steward	Link
☐ Move approved products to the appropriate location of the archive folder	Data Custodian	N/A	
☐ Maintain project and product metadata record associations	Data Steward	N/A	Link
☐ Confirm readiness of products for discoverability and accessibility	Data Custodian	Project Manager Data Trustee	

Appendix B: Tree Structure for File Organization of the Archive Record

Descriptions or details of the file/folder given in parentheses.

root (project name; takes the format ProgramAcronym_SequentialNumber_ShortTitle, e.g., MBMWF_011_YKDeltaNestPlot would stand for the Migratory Bird Management Waterfowl Program, YK Delta Nest Plot Survey, and it was the 11th archive record created for that program)

changelog.txt (text file for the data custodian to record activities in the archive record)

- **admin** (material related to general project administration; could be replicated each year for multiyear projects, i.e., admin2019, admin2020)
  - **contracts**
  - **correspondence** (including permits obtained for the project)
  - **proposals** (those prepared for both internal or external funding)
  - **purchasing** (significant or unique purchasing information that is deemed important to archive)
  - **training** (training materials developed for the project)
  - **travel** (significant or unique travel information that is deemed important to archive)
- **code** (R or Python scripts, for example)
- **data** (data generated from the project)
  - **raw_data** (unprocessed data as initially recorded; structure will vary by project; recommended to organize data by datatype)
    - **raw_type_of_data_x**
      - raw_data.csv
    - **raw_type_of_data_y** (repeat as needed)
      - images/text notes/ sounds/geo points or lines/etc.
  - **final_data** (data that has passed all quality control checks and are typically the data products documented in metadata records; generally data used for analysis; structure will mirror that of the raw_data folder)
    - **final_type_of_data_x**
      - raw_data.csv
    - **final_type_of_data_y** (repeat as needed)
      - images/text notes/ sounds/geo points or lines/etc.
- **documents** (materials generated by the project; these products are also documented in metadata)
  - **images** (stand-alone posters, maps, drawings, pictures)
  - **publications** (peer reviewed)
  - **reports** (not peer reviewed)
  - **talks** (presentations)
- **incoming** (holding location for materials that should be filed under one of the other branches after review by the data steward)
- **metadata** (contains the mdEditor JSON file for the project, associated products, project contacts, and data dictionaries)
- **protocols** (methods and protocol documents that guide project procedures)
  - **form_template** (data sheets, templates for data entry, etc)

source_data (data NOT generated by the project but used during the course of the project; could include data used to design the project)

- **source_type_of_data_x**
  - raw_data.csv
- **source_type_of_data_y** (repeat as needed)
  - images/text notes/sounds/geo points or lines/etc.

Appendix C: Example of untidy vs. tidy tabular data

UNTIDY DATA

AtlasGroveCOMPLETE.xls

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leaves						dry masses (kg)				
species	tree	kg	kg	kg	kg	kg		type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2		tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8		tree	PSME	135815	0	0	8338	961	145114	3.3876
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4		tree	THSE	31799	0	0	6343	864	39006	0.9105
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9		tree	ACMA	4444	0	0	925	264	5634	0.1315
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4		tree	UMCA	2921	0	0	937	273	4131	0.0964
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7		shrub	RUSP	0	0	0	1974	686	2660	0.0620
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4		fern	POMU	0	0	0	0	1271	1271	0.0296
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8		shrub	VAOV	0	0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3		shrub	COCO	0	0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2		fern	POSC	0	0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7		tree	RHPU	100	0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0		herb	OXOR	0	0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524.2	691.2		shrub	VAPA	0	0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	19104.7	954.3		tree	PISI	0	0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87.6	41.4		tree	CHLA	0	0	0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	214.1	43.8		shrub	GASH	0	0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8.7	2.5		shrub	SACA	0	0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055.2	66.3				3744390	213247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626.3	49.6										proportion
SESE	11	614.4	0.0	0.0	28.1	17.0				main trunk	reiteration	limb	branch	leaf	total	geophytic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	59	0	0	12	4	74	
SESE	25	87.7	0.0	0.0	4.1	1.3			ACMA geo	4444	0	0	925	264	5634	1.00
SESE	30	512.1	1.8	0.0	18.7	8.7			ACMA epi	0	0	0	0	0	0	

Characteristics that make this an untidy dataset

- There are three tables on this sheet
- A row is NOT one observation
- A column is NOT one variable
- Totals and percentages are calculated in the sheet

TIDY DATA

	A	B	C	D	E	F	G	H	I	J
1	Year	DecoyID	Easting	Northing	TransectType	TransectNo	NorthSouth	DistanceCat	HabitatType	Location
2	2017	D01	583280.96	7908556.96	Decoy	1	S	130-190	dry tundra	tundra
3	2017	D02	585328.6	7907168.87	Decoy	3	S	20-80	deep-arcto	shoreline
4	2017	D03	585807.74	7908935.08	Decoy	1	N	80-130	shallow-arcto	off-shore
5	2017	D04	589979.74	7908884.69	Decoy	1	S	80-130	shallow-carex	shoreline
6	2017	D05	591493.06	7909183.32	Decoy	1	N	80-130	shallow-carex	off-shore
7	2017	D06	580775.85	7907687.96	Decoy	2	S	130-190	deep-arcto	shoreline
8	2017	D07	588193.94	7907978.99	Decoy	2	S	130-190	shallow-arcto	shoreline
9	2017	D08	583045.49	7907740.22	Decoy	2	S	130-190	shallow-carex	shoreline
10	2017	D09	597465.33	7908699.94	Decoy	2	N	130-190	shallow-carex	off-shore
11	2017	D10	587330.83	7908063.19	Decoy	2	S	20-80	flooded tundra	shoreline
12	2017	D11	592117.13	7907669.82	Decoy	3	N	130-190	shallow-carex	off-shore
13	2017	D12	597645.5	7907578.94	Decoy	3	S	130-190	shallow-carex	shoreline
14	2017	D13	588262.57	7907198.67	Decoy	3	S	130-190	shallow-carex	shoreline
15	2017	D14	580980.52	7906980.62	Decoy	3	S	20-80	dry tundra	tundra
16	2017	D15	591762.97	7907404.24	Decoy	3	S	80-130	shallow-arcto	shoreline
17	2017	D16	593910.35	7907473.11	Decoy	3	S	80-130	dry tundra	tundra
18	2017	D17	589356.97	7907507.02	Decoy	3	N	80-130	ditch	off-shore
19	2017	D18	589484.33	7906763.92	Decoy	4	N	130-190	shallow-arcto	off-shore
20	2017	D19	578970.57	7905958.85	Decoy	4	S	130-190	shallow-carex	shoreline
21	2017	D20	595294.14	7907000.72	Decoy	4	N	130-190	shallow-arcto	shoreline
22	2017	D21	582452.9	7906241.03	Decoy	4	S	20-80	deep-open	shoreline
23	2017	D22	585090.64	7906456.54	Decoy	4	N	20-80	shallow-arcto	off-shore

Characteristics that make this a tidy dataset

- Each row is one observation
- Each column is one variable
- There is no special formatting, borders, derived calculations, headers with metadata information, etc. on this sheet.

Glossary

When consistent with usage in this document, definitions have been pulled from various other resources including Wikipedia, Open Data Handbook glossary, Duke Law EDRM glossary, and the Open Government Data Act codification of act definitions 44 usc 3502. It is possible to find alternative and contradictory definitions in other data management resources resources. The definitions provided here are those implied by the term's usage in this document.

Accessibility: the degree to which the resource is obtainable by an interested party. Direct access without constraint would be the most accessible (e.g. resources that may be downloaded without requiring a login), whereas resources that require third-party intervention would be less accessible.

Archive Folder: a consistent file structure with use constraints and backup schedule that houses the definitive record of a project's data resources. Products in the archive folder are the subject of metadata records and are the versions intended for use and dissemination. Contrast with working folder.

Data Catalog: database comprised of metadata allowing for the discovery of data resources.

Data Custodian: individual responsible for the storage and security of a data resource.

Data Trustee: individual having the authority to: 1) ensure resources are available to implement the complete project and data lifecycle and 2) ensure compliance with all data governance policies.

Data Dictionary: provides information on the contents of a dataset to support data quality and use. Such information includes entity (i.e., variable) definitions and allowable values. In the case of databases, or a collection of datasets, relationships between tables are also defined in the data dictionary.

Data Integrity: property describing foundational soundness of a data resource. Data with strong integrity have undergone quality control and assurance procedures throughout their lifespan, have permanence over a reasonable timeframe and changes to the data are appropriately documented.

Data Management: an administrative process that includes acquiring, validating, storing, and securing data to ensure the accessibility, integrity, and timeliness of the data for its users.

Data Management Plan: document that describes the data expected from the project, how such data will be handled throughout the project to protect data integrity, and stored at the conclusion of the project to ensure security, discoverability, and accessibility.

Data Resources: data. Recorded information, regardless of form or the media on which the data is recorded. aka Products.

Data Steward: individual responsible for reviewing the quality and metadata of a resource.

Discoverability: the degree to which information about a data resource's existence is readily obtained via searching an information system (e.g., Data.gov). Certain aspects of the metadata for the resource may be useful in enhancing discoverability, such as keywords or spatial bounds. Data catalogs can

enhance discoverability by providing a standard location for searching and organizing resources. A data resource may be discoverable (e.g. found in a search result) but not accessible (see accessibility).

ISO: the International Organization for Standardization. Entity that provides standards to ensure consistency in definitions, formats, and use.

mdEditor: a web application used to write archival-quality metadata for projects and data resources. mdeditor.org

Metadata: data that describes and provides additional information about other data to promote discoverability and proper use.

Open Format: data format that is platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information.

Project: a discrete effort on a particular topic with defined objectives or goals.

Project Management: the practice of initiating, planning, executing, controlling, and closing the work of a team to achieve specific goals and meet specific success criteria at the specified time.

Quality Assurance: preventing errors. The maintenance of a desired level of quality in a product, by means of attention to every stage of the process of acquisition, manipulation, and use

Quality Control: identifying and correcting errors. Process of review to reduce or eliminate errors made during data acquisition and manipulation.

Reproducible (analyses, workflow, or research): structuring activities so that a product (e.g., a data set, analysis result, or report) can be repeated and the same results achieved. Replication could be achieved by either the same person or team that created the original product or a different team. Documentation and scripted work flows play a key role in reproducibility.

Tidy Data: standard way of relating the structure of a dataset to its meaning. Specifically, each row represents an observation and each column represents a variable recorded on an observation.

Working Folder: a file structure used by an individual, or a group in collaboration, to store data resources under production during the course of a project's implementation. Contrast with archive folder.