Data Management Plan

Washington Soil Health Initiative: State of the Soils Assessment

Jadey Ryan

Leslie Michel

Dani Gelardi

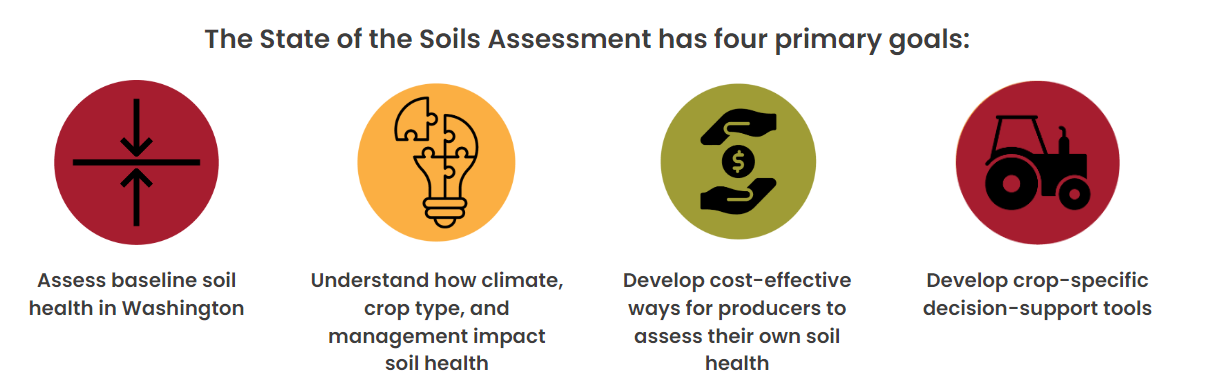
2023-11-30

Table of contents

# 1. Overview

The [Washington Soil Health Initiative](https://washingtonsoilhealthinitiative.com/) (WaSHI) is a partnership between the Washington State Department of Agriculture (WSDA), Washington State University (WSU), and the State Conservation Commission. WaSHI establishes a coordinated approach to healthy soil in Washington.

To date, nearly 1,000 soil samples and management surveys across 50 different cropping systems have been collected as a part of the [state of the soils assessment](https://washingtonsoilhealthinitiative.com/state-of-the-soils/) (SoSA). WSDA and WSU lead this project with support from staff, students, conservation districts, and agricultural professionals throughout Washington.



## 1.1 What is data management?

Effective data management involves properly documenting, storing, and sharing our data and the information we derive from the data. If the data aren’t useable by researchers, policymakers, or growers, the hours spent in the field collecting soil samples and the hours spent in the lab analyzing them may be wasted.

The guidelines detailed in this DMP will help us achieve the above data-driven goals, while also optimizing the value of the data by supporting information sharing and innovation. Our data management strategies attempt to follow **FAIR** (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) principles while also maintaining data privacy (Wilkinson et al. 2016).

## 1.2 Data life cycle

The U.S. Fish and Wildlife Service developed a great illustration of the data life cycle that shows the elements of data management and emphasizes the importance of data quality at every step (U.S. Fish & Wildlife Service 2023). Each step within the data life cycle requires careful intention to ensure transparency, quality, and integrity.

|  |
| --- |
|  |

Our adaptation of this data life cycle is outlined below.

|  |  |
| --- | --- |
|  | **Plan**  Each sampling year presents an opportunity to consider what worked and what could be improved from the previous year. Planning involves making decisions about data acquisition, management, and quality control. For example, each year we provide a spreadsheet template with our requested column headers to Soiltest lab to ensure the measurements are reported with correct units and in the format we use. Special projects that deviate from our standard operating procedures require additional planning. |
|  | **Acquire**  We acquire data by collecting and analyzing new samples, deriving new insights from existing data, or accepting data sets from collaborators. |
|  | **Maintain**  Maintenance involves processing data for aggregation, analyses, and reporting. We create metadata that facilitates interpretation of the data and ensure the data are in a non-proprietary format that is accessible to our collaborators and future selves. |
|  | **Access**  Access refers to data storage, publication, and security. Raw and processed data with accompanying metadata should be stored, backed up, and available for information sharing with our partners. With PI approval, anonymized and aggregated data that does not compromise growers’ personally identifiable information (PII) should be made publicly available in a data repository or data product/decision-support tool. |
|  | **Evaluate**  We evaluate data while processing and analyzing it to maximize accuracy and productivity, while minimizing costs associated with errors or tedious data cleaning labor. Evaluation workflows should be efficient, well-documented, and reproducible. Our evaluated data help us better understand how factors and management decisions impact soil health. |
|  | **Archive**  Properly archiving our results supports the long-term storage and usefulness of our data. While similar to the Access element of the life cycle, archiving focuses on preserving data for long-term/historical retention that aren’t needed for immediate access. For example, we archive each year’s raw data for long-term storage and set those files to Read-Only. |
|  | **Quality Assurance / Quality Control (QA/QC)**  Data quality management prevents data defects that hinder our ability to apply data towards our science-based conservation efforts. Defects include incorrectly entered data, invalid data, and missing or lost data. QA/QC processes should be incorporated in every element of the data life cycle. |

## 1.3 DMP overview

This living document will be continually reviewed and improved based on lessons learned, new information, and collaborator feedback. Subsequent chapters provide more technical details.

### Roles and responsibilities

All WaSHI personnel who will be interacting with SoSA data must familiarize themselves with the contents of this document. Following chapters with technical details will be referenced when relevant. If all collaborators are not consistently implementing this DMP, then the benefits of effective data management are lost.

The WSDA Data Scientist, supported by the project Principal Investigators (PIs), is responsible for providing guidance to WaSHI staff working with SoSA data and ensuring the implementation of the DMP. The Data Scientist is also responsible for reviewing and updating this document annually and as needed. Upon updates, the Data Scientist will distribute this document to WaSHI staff and commit the source code to the [GitHub repository](https://github.com/WA-Department-of-Agriculture/washi-dmp).

Current roles as of November 2023

| Role | Affiliation | Name |
| --- | --- | --- |
| Data Scientist | WSDA | Jadey Ryan |
| Co-PI | WSDA | Dani Gelardi |
| Co-PI | WSU | Deirdre Griffin LaHue |
| Data Stewards | WaSHI personnel |  |

## 1.4 Data sharing and public access

SoSA relies on growers’ willingness to volunteer their fields for sampling and participate in the required management survey. Their willingness depends on their trust in WaSHI to protect their privacy. Only aggregated and anonymized results will be publicly available or shared. The below data privacy statement may be shared with potential participants:

### Data privacy statement

Data will be aggregated and reported in a way which mitigates personal identification of growers. Information will be used to understand broad impacts and characterize trends in soil health and production practices across regions. Results will not be reported in a way that makes individuals identifiable. Information collected in this survey may be subject to release in accordance with RCW 42.56 (Public Records Act).

Procedures for anonymizing data are detailed in [Section 8.3](#sec-maintain-confidentiality).

### Acknowledging WaSHI data in publications

All research partially or completely funded by WaSHI must include acknowledgements to the State of Washington. The following text should be included in all publications resulting from this funding:

Data was in part provided by the Washington Soil Initiative, which is supported by the State of Washington and administered by the Washington State Department of Agriculture, Washington State Conservation Commission, and Washington State University.

If WaSHI staff make [substantial scientific contributions](https://www.pnas.org/doi/10.1073/pnas.1715374115) to the manuscript, discuss the possibility of co-authorship credit.

## 1.5 Acknowledgements

This DMP was adapted from the R.J. Cook Agronomy Farm Long-term Agroecological Research Site DMP (Carlson 2021), U.S. Fish and Wildlife Service data management life cycle (U.S. Fish & Wildlife Service 2023), Harvard Medical School Longwood Research Data Management DMP guidelines (Harvard Medical School 2023), and the Data Management in Large-Scale Education Research book (Lewis 2023).

# 2. Formats and standards

## 2.1 Data formats

Data generated from or integrated into WaSHI can be non-digital or digital.

### Non-digital data

Non-digital data, such as field forms, management surveys, and chain of custody forms, are manually recorded on paper forms. Paper forms must be transcribed or converted to digital file formats and then stored in the WaSHI filing cabinet in the Natural Resources Building in Olympia.

### Digital data

Digital data include tabular, spatial, and binary data, such as lab results, sample locations, and field photos. Non-conventional data also include code, algorithms, tools, and workflows.

**Tabular data** include comma separated values (csv), tab separated values (tsv), Microsoft Excel open XML spreadsheet (xlsx), and portable document format (pdf).

**Spatial data** include file geodatabases (gdb), vector shapefiles (zipped folder containing multiple file extensions), keyhole markup language (kml or kmz). Tabular data may also contain spatial data as longitude and latitude.

**Binary data** include photos (jpeg, png, gif, tiff), videos (mp4), code (R, py, js), and object-oriented data files (RDS, Rdata, parquet, arrow).

**Proprietary data formats** include Microsoft Excel, Word, and Powerpoint files (xlsx, docx, pptx). RDS and Rdata files are an example of an application-specific data format that can only be opened using the R programming language or RStudio IDE. These types of files should be saved in conjunction with a copy of the data in a non-proprietary and open-standard format, such as csv, to maintain accessibility for those who do not have Microsoft Office or do not use R.

**Written documents and presentations** are in formats including Microsoft Word and Powerpoint (docx and pptx), hypertext markup language (HTML), and pdf.

**Notebooks** combine text with executable code to generate written documents and presentations in docx, pptx, HTML, or pdf formats. These notebooks are stored in formats depending on the programming language: a few examples include R markdown (rmd), Quarto (qmd), and Jupyter notebook (ipynb).

The list below is not exhaustive and will continue to grow as additional useful data sources are discovered.

| **Type** | **Source** | **Formats** |
| --- | --- | --- |
| **Lab results** | Provided by the lab analyzing the soil sample, principal investigator of a study, or grower | csv, xlsx, pdf, xml, json, RDS, RData |
| **Management surveys** | Collected through interviews with grower | csv, xlsx, RDS, RData, paper form (to be digitized) |
| **Field forms** | Completed in the field during/immediately after sampling | pdf, paper form (to be digitized), csv, xlsx |
| **Sample locations** | Identified prior to sampling and may be edited during sampling using ArcGIS Online, Collector, Field Maps or Google Maps | ArcGIS feature layer, shp, kmz, csv, xlsx |
| **Chain of custody forms** | Completed prior to shipping or dropping off samples | pdf, paper form (to be digitized) |
| **Climate data** | [OSU PRISM](https://prism.oregonstate.edu/), [NOAA](https://www.ncdc.noaa.gov/cdo-web/datasets), [Esri Living Atlas](https://livingatlas.arcgis.com/en/home/) | csv, shp, netCDF, tiff, gdb |
| **Soil data** | [NRCS Web Soil Survey](https://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm), [NRCS WA gSSURGO](K:\NRAS\Arc_Data\Soil_Health\NRCS_StatewideSoils) | gdb, accdb |
| **Strata classification** | Provided by Soil Health Institute in 2021 as a [lyr file](K:\NRAS\Arc_Data\Soil_Health\SoilHealthInstitute_Strata) | lyr |
| **Images** | Logos, icons, photos taken in the field | jpeg, png, gif, tiff, svg |
| **Videos** | Recordings of meetings, training videos | mp4 |
| **Documents** | Reports, manuscripts, SOP, QAPP, factsheets, brochures | docx, txt, html, pdf |
| **Presentations** | Powerpoints, slide decks | pptx, html, pdf |
| **Code** | Scripts for wrangling, processing, analyzing data; markdown for producing documents and presentations; style sheets for html outputs | R, py, ipynb, js, yml, rmd, qmd, css, scss |

## 2.2 Data standards

**Date** will be expressed as YYYY-MM-DD according to the [ISO 8601 standard](https://www.iso.org/iso-8601-date-and-time-format.html).

**Date with time** will be expressed as YYYY-MM-DD**T**HH:MM:SS**Z**.

* **T** separates date from time. The **Z** indicates the date-time is using the Universal Time Coordinated (UTC) with no offset.
* Pacific Standard Time (PST) has a UTC-8:00 offset and Pacific Daylight Time (PDT) has a UTC-7:00 offset and would be expressed as YYYY-MM-DD**T**HH:MM:SS-8:00. The **Z** has been replaced with the offset.
* Example: 2023-11-28T14:55:56-08:00.

|  |
| --- |
| "ISO 8601" from Randall Munroe's xkcd |

**Geospatial** data will be accompanied by metadata that abides by the [ISO 19115 standard](https://www.iso.org/standard/53798.html) by following Esri’s [documentation](https://pro.arcgis.com/en/pro-app/latest/help/metadata/create-iso-19115-and-iso-19139-metadata.htm) when working in ArcGIS Pro. Metadata contains information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.

**Code** will follow the style guide in [Chapter 9](#sec-code-style-guide).

# 3. Organization

## 3.1 Storage

Non-digital data, such as paper forms, must be transcribed or converted to digital file formats and then stored in the WaSHI filing cabinet in the Natural Resources Building in Olympia.

All digital data are stored in the WSDA shared drives, and other locations listed below.

**WSDA shared drives:**

* Agency files: <Y:\NRAS\Soil_Health_Initiative>
* GIS: <K:\NRAS\Arc_Data\Soil_Health> (requires permissions from IT)

**Esri products and services:**

* ArcGIS Online [Soil Health - WSDA Internal Group](https://nras.maps.arcgis.com/home/group.html?id=17a60cfa644c4c60ab622fdd84500f8f#overview)
* WSDA GIS on-premise [ArcGIS REST Services Directory](https://fortress.wa.gov/agr/gis/wsdagis/rest/services/NRAS) (only Jadey, Perry, and Joel can publish to this server; Ed Thompson is the contact for getting access)

**Database for lab results and management data:**

* WISKI, but very likely will migrate to SQL Server or a less water-focused database

**GitHub organizations for code-based projects:**

* [WSDA](https://github.com/WA-Department-of-Agriculture)
* [WaSHI](https://github.com/WASoilHealth)

**Microsoft Teams for data sharing between WSDA and WSU:**

* WSDA and WSU Teams WaSHI channels

**Local workstations:**

* Should NOT be the only place data are stored!

### Read-Only raw data

Raw data, such as lab results from Soiltest or exports from ArcGIS Online, should be immediately set to Read-Only. Right click the file > click on Properties > check the Read-Only attribute box. The file should then be copied over to a Working folder for any processing or analyses. The final data set should be saved in a separate descriptively titled Clean folder. Keeping a README.txt to document your processing and analysis steps is good practice.

## 3.2 Naming conventions

When naming folders and files, we want consistent and clear names that are findable and understandabe by both humans and computers. From only a file name, we should immediately know what the file contains and which file is the most recent version.

|  |
| --- |
| “Documents” from Randall Munroe’s xkcd |

### Best practices

#### Convey meaning with delimiters and capitalization

Deliberately use underscores, hyphens, and uppercase letters so we can easily understand the contents and programmatically parse file and folder names.

* Use underscores to delineate metadata elements (i.e. name from version Name\_Version).
* Use hyphens to separate parts of one metadata element (i.e. date YYYY-MM-DD).
* Different conventions may work better for different purposes (folder, file, variable name). See [Section 3.2.2](#sec-naming-guidelines) for which conventions to use for different file sets.
  + **PascalCase**: capitalize the first letter of each word without spaces or delimiters.
  + **lowerCamelCase**: the first word is all lowercase and subsequent words have the first letter capitalized without spaces or delimiters.
  + **snake\_case**: all lowercase with underscores separating words
  + **kebab-case**: all lowercase with hyphens separating words
  + **Train-Case**: each word begins with an uppercase letter with hyphens separating words

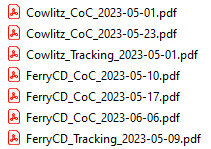
Avoid spaces and special characters (only use underscores and hyphens). Characters like / () ! ? % + " ' have special meaning to computers and can break file paths and URLs.

#### Character length matters

Computers are unable to read file paths and file names that surpass a certain character length. Be concise AND descriptive. Windows path limit is 260 characters.

#### Group and sort files by name

Consider how folders and files should be grouped and sorted. Put that piece of metadata in the beginning of the file name. For example, if we want all chain of custodies grouped by conservation district and then sorted by date, we would use DistrictName\_CoC\_YYYY-MM-DD, which looks like Cowlitz\_CoC\_2023-05-01.pdf. If we wanted them grouped by date, we would put the date at the beginning.



#### Use version numbers

If not using the date to version, or to keep multiple drafts from the same date, add version information to the end of the file name. Think about how many possible versions there could be. If there may be more than 10, use leading zeros before single digit numbers so the file name always has the same length. V1 through V15 will not sort the same way as V01 through V15. For example: 2023 SampleRequestForm\_Lewis\_V3 or 2023 SampleRequestForm\_Lewis\_V03.

#### Collaborating on files

Add your initials at the end of the file name when “saving as” a file that multiple people are working on. This ensures a version is kept as a backup. Alternatively, use [Track Changes](https://support.microsoft.com/en-au/office/track-changes-in-word-197ba630-0f5f-4a8e-9a77-3712475e806a) if working in a MS Word document.

### Naming guidelines

Many files and folders in our shared drive do not follow the above best practices or below naming conventions. We are learning and improving as we go. Going forward, please follow the conventions listed below.

|  | Naming convention | Examples |
| --- | --- | --- |
| **Folders** | YYYY\_PascalCase  optional: year | 2024\_Sampling  DataManagement |
| **Dynamic files**  (include metadata in name) | YYYY-MM-DD\_PascalCase\_V1 | 2023-11-15\_SurveyPerennial.xlsx  2023\_WSDASoilHealth\_V1.xlsx  2023\_WADE-WaSHI-WaCSE\_JR.pptx |
| **Static files** | snake\_case | washi\_logo\_no\_background.png |
| **Domains**[[1]](#footnote-99) | kebab-case | washi-dmp |
| **Scripts and syntax**[[2]](#footnote-101) | snake\_case  lowerCamelCase | create\_soils.R (script)  washi\_flextable (function)  sampleId (variable) |

## 3.3 Folder structure

We organize our folders into a hierarchical structure to clearly delineate segments of our projects, improve searchability, and ensure reproducibility across years.

### Hierarchy

We strive for a balance between a deep and shallow structure. If too shallow, there are too many files in one folder and they are hard to sort through. If too deep, we have to click too many times to get to a file and specific files can be difficult to find.

Y:\NRAS\Soil\_Health\_Initiative is the parent folder for all WaSHI content.

Within the Sampling sub-folder, we use a combination of **date-** (each year has its own sub-folder) and **categorical-** based (data set and documentation that span across years) folder structure .

Y:\NRAS\Soil\_Health\_Initiative\Sampling  
├── \_completeDataset  
├── 2019\_SCBG  
├── 2021\_Sampling  
├── 2022\_PartnershipsInSoilHealth  
├── 2023\_Sampling  
├── 2024\_Sampling  
├── Maps  
├── Projects  
├── QAPP  
├── SOPs  
├── TrainingVideos  
├── ArchivedSampleInventory.xlsx  
├── EquipmentInventory.xlsx  
└── SOSImpacts.xlsx

Within the 2023 sub-folder, we have sub-folders for planning, forms, data, and processes. This structure helps maintain a reproducible workflow year after year.

Y:\NRAS\Soil\_Health\_Initiative\Sampling\2023\_Sampling  
├── Applications  
├── CoCs  
├── Equipment  
├── FieldForms  
├── Forms  
├── GIS  
├── LabData  
├── Labels  
├── ManagementSurveys  
├── PublicDocs  
├── Purchases  
├── Reports  
├── SampleIDAssignments  
├── Scripts  
├── 2023\_DataTracking.xlsx  
└── PostSeasonWrapUp\_2023.docx

As mentioned in [Section 3.1.1](#sec-raw-data), it’s good practice to maintain the raw data. We use additional sub-folders for the LabData folder. Everything in Raw has been set as Read-Only.

Y:\NRAS\Soil\_Health\_Initiative\Sampling\2023\_Sampling\LabData  
├── 2023\_DataTemplateSoiltest.xlsx  
├── Clean  
├── QC  
├── Raw  
└── Working

Soil\_Health\_Initative > Sampling > 2023\_Sampling > LabData > Clean has five levels of nesting. We wouldn’t want to add too many more levels before the hierarchy becomes difficult to manage.

### Archive folders

When too many drafts or versions begin to clutter a sub-folder, create a new folder with the structure Archive\_FileTitleOrDescription and place the old drafts there. Leave the most current, accurate file in the main folder.

For example, the final, current SOP is listed in the main folder and previous working drafts were moved to the Archive\_SOP folder.

Y:\NRAS\Soil\_Health\_Initiative\Sampling\SOPs\Sampling  
├── 2023\_WSDA\_NRAS\_SOP\_Soil\_Health\_Monitoring.docx  
├── 2023\_WSDA\_NRAS\_SOP\_Soil\_Health\_Monitoring\_WEB.docx  
├── 923-NRAS-SoilHealthSOP.pdf  
├── 923-NRAS-SoilHealthSOP\_WEB.pdf  
├── Archive\_SOP  
│ ├── 2021  
│ ├── 2022  
│ ├── 2023  
│ │ ├── 2023\_WSDA\_NRAS\_SOP\_Soil\_Health\_Monitoring\_LMedits.docx  
│ │ └── 2023\_WSDA\_NRAS\_SOP\_Soil\_Health\_Monitoring\_LM\_JRedits.docx

# 4. Documentation

## 4.1 Metadata

## 4.2 Data dictionary

# 5. Data flows

## 5.1 Project structure

## 5.2 Original data

## 5.3 Working data

## 5.4 Clean data

# 6. Data quality

## 6.1 Quality assurance and quality control

Link to QC SOP

QC Codes

## 6.2 External Data Sources

Rate data quality

Maintain metadata

Standards for accepting data

Parameters for what we will or won’t use (soils data with management metadata only)

List example places to look

# 7. Data preservation and versioning

## 7.1 Retention schedule

## 7.2 Archive

Non-code data should be archived in WSDA shared drive

## 7.3 Version with GitHub

Code should be versioned and archived using GitHub

Include helpful resources for how-to use

## 7.4 Staff turnover

* WSDA GitHub Organization
* Zenodo
* Database credentials

# 8. Data sharing

## 8.1 Public repositories

* GitHub
* Zenodo
* data.gov

## 8.2 Understand WaTech data categorization

<https://watech.wa.gov/Categorizing-Data-State-Agency>

## 8.3 Maintain confidentiality

Anonymize and aggregate

## 8.4 Timeline

When to share data (after publication?), how long to share data?

# 9. Code Style Guide

## 9.1

Template snippet

Section break snippet

.R and .qmd file names

function names

argument names

variable names

# References

Carlson, Bryan. 2021. “Data Management Plan for the R.J. Cook Agronomy Farm Long-Term Agroecological Research Site.”

Harvard Medical School. 2023. “Data Management Plans.” <https://datamanagement.hms.harvard.edu/plan-design/data-management-plans>.

Lewis, Crystal. 2023. *Data Management in Large-Scale Education Research [in Preparation]*. <https://datamgmtinedresearch.com/>.

U.S. Fish & Wildlife Service. 2023. “Data Management Life Cycle.” <https://www.fws.gov/data/life-cycle>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.

1. Domains include RStudio project names and GitHub repositories since they become a URL. We use kebab-case because hyphens are better than underscores for URLs, [as recommended by Google](https://developers.google.com/search/docs/crawling-indexing/url-structure). [↑](#footnote-ref-99)
2. See [Chapter 9](#sec-code-style-guide) for the code style guide. [↑](#footnote-ref-101)