Data Management Plan

Washington Soil Health Initiative: State of the Soils Assessment

Jadey Ryan

Dani Gelardi

2023-12-29

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 Chapter outline

*This Data Management Plan (DMP) is a living document will be continually reviewed and improved based on lessons learned, new information, and collaborator feedback.*

## 1.2 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 |  |

## 1.3 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. 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 policies attempt to follow **FAIR** (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) principles while also maintaining data privacy (Wilkinson et al. 2016).

## 2.1 Data life cycle

The U.S. Fish and Wildlife Service developed a great graphic to explain the elements of the data life cycle and emphasize 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 datasets 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. |

The following chapters describe our internal processes and standards to follow throughout each step in the data life cycle.

# 3. Formats & standards

## 3.1 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 |

## 3.2 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 10](#sec-code-style-guide).

# 4. 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).
* Avoid spaces and special characters (only use underscores and hyphens). Characters like / () ! ? % + " ' have special meaning to computers and can break file paths and URLs.
* Different conventions may work better for different purposes (folder, file, variable name). See [Section 4.0.2](#sec-naming-guidelines) for which conventions to use for different file sets.
  + **UpperCamelCase**: 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

|  |
| --- |
| Artwork by @allison\_horst |

#### 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.

#### Express date ‘back to front’

Remember to express date ‘back to front’ of YYYY-MM-DD according to the [ISO 8601 standard](https://www.iso.org/iso-8601-date-and-time-format.html). This maintains the chronological order of records when they are sorted alphanumerically.

| ✅ Do this | ❌ Don’t do this |
| --- | --- |
| 2020-05-28\_Agenda.pdf | 2-14-2023\_Agenda.pdf |
| 2023-01-01\_Agenda.pdf | 2023-Jan-1\_Agenda.pdf |
| 2023-02-14\_Agenda.pdf | Dec052020\_Agenda.pdf |
| 2023-12-05\_Agenda.pdf | May-28-2020\_Agenda.pdf |

#### 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.

| Sort by district | Sort by date |
| --- | --- |
| Cowlitz\_CoC\_2023-05-01.pdf | 2023-05-01\_Cowlitz\_CoC.pdf |
| Cowlitz\_CoC\_2023-05-23.pdf | 2023-05-01\_Cowlitz\_Tracking.pdf |
| Cowlitz\_Tracking\_2023-05-01.pdf | 2023-05-09\_FerryCD\_Tracking.pdf |
| FerryCD\_CoC\_2023-05-10.pdf | 2023-05-10\_FerryCD\_CoC.pdf |
| FerryCD\_CoC\_2023-05-17.pdf | 2023-05-17\_FerryCD\_CoC.pdf |
| FerryCD\_CoC\_2023-06-06.pdf | 2023-05-23\_Cowlitz\_CoC.pdf |
| FerryCD\_Tracking\_2023-05-09.pdf | 2023-06-06\_FerryCD\_CoC.pdf |

#### 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.

| ✅ Do this | ❌ Don’t do this |
| --- | --- |
| SOP\_V01.pdf | SOP\_V1.pdf |
| SOP\_V02.pdf | SOP\_V10.pdf |
| … [V03 - V09] | SOP\_V11.pdf |
| SOP\_V10.pdf | SOP\_V2.pdf |
| SOP\_V11.pdf | … [V3 - V9] |

#### Collaborating on files

Add your initials at the end of the file name when “saving as” a file that multiple people are working on (i.e., 2024\_SOP\_SoilHealthMonitoring\_LM\_JR.docx). 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.

These are guidelines and naming things is hard, so try your best. If you’re not sure how to name a folder or file, talk with Jadey. If you’re adding a bunch of files that came from somewhere else, ask Jadey to help you organize and bulk rename them.

In addition to these case naming conventions, use underscores to separate different metadata groups (date\_name\_version) and hyphens to separate parts of a date (YYYY-MM-DD) or the parts of the same metadata group.

|  | Naming convention | Examples |
| --- | --- | --- |
| **Folders** | UpperCamelCase | 2024\_Sampling  DataManagement |
| **Versioned files** | UpperCamelCase | 2023-11-15\_SurveyPerennial.xlsx  RecordsManagement\_V01.docx  PendOreilleCD\_CoC\_2023-06-05.pdf |
| **Non-versioned files** | kebab-case | washi-logo-color.png |
| **Code files**[[1]](#footnote-89) | kebab-case | washi-dmp.Rproj  01\_load-metadata.R  2024\_producer-report.qmd |
| **Syntax**[[2]](#footnote-91) | snake\_case  lowerCamelCase | assign\_quality\_codes() (function)  sampleId (variable) |

# 5. Organization

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

### Folder Structure

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 StateOfTheSoils sub-folder, we use a combination of **date-** (each year has its own sub-folder) and **categorical-** based (dataset and documentation that span across years) folder structures.

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

Within the each year sub-folder, we have sub-sub-folders for planning, forms, data, and processes. This structure helps maintain a reproducible workflow year after year. See the 2023\_Sampling for an example:

Y:\NRAS\Soil\_Health\_Initiative\StateOfTheSoils\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 6.2](#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\StateOfTheSoils\2023\_Sampling\LabData  
├── 2023\_DataTemplateSoiltest.xlsx  
├── Clean  
├── QC  
├── Raw  
└── Working

Soil\_Health\_Initative > StateOfTheSoils > 2023\_Sampling > LabData > Clean already has five levels of nesting. We wouldn’t want to add any many more levels or 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 naming convention of Archive\_FolderDescription. Place the old drafts there. Leave the most current, accurate file in the main folder.

For example, the most recent sample labels for each conservation district are listed in the top level CompletedLabels folder, and previous working drafts were moved to the Archive\_Labels folder.

Y:\NRAS\Soil\_Health\_Initiative\StateOfTheSoils\2023\_Sampling\Labels\CompletedLabels  
├── Archive\_Labels  
│ ├── CowlitzCounty\_Labels.docx  
│ ├── FerryCD\_Labels.docx  
│ ├── LewisCD\_Labels.docx  
│ └── StevensCD\_Labels.docx  
├── CowlitzCounty\_Labels\_V2.docx  
├── FerryCD\_Labels\_V2.docx  
├── KittitasCD\_Labels.docx  
├── LewisCD\_Labels\_V2.docx  
├── ...  
└── WallaWallaCD\_Labels.docx

# 6. 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> (access 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

**Box.com for external file sharing:**

* WSDA has a box.com account. Only Perry has a license; Jadey and Dani are currently editors.

**Individual devices (laptop, tablet, phone):**

* Must NOT be the only place data are stored!

## 6.1 Backup

Data must be stored in multiple locations. At a bare minimum, data on an individual computer must also be saved on the WSDA shared drive. Backing up data using version control (GitHub) or a cloud service (Microsoft OneDrive or Box.com) is strongly recommended.

## 6.2 Read-only raw data

On our shared drives, 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 dataset should be saved in a separate descriptively titled Clean folder. Keeping a readme.txt to document your processing and analysis steps is good practice, as discussed in [Section 7.2.1](#sec-readme).

## 6.3 Version control with Git and GitHub

A version control system records changes to a file or set of files over time. [Git](https://git-scm.com/) is a free and open-source distributed version control system and [GitHub](https://github.com) is the hosting site WSDA and WaSHI use to interface with this system. Git and GitHub are an important foundation of reproducible statistical and data scientific workflows (Bryan 2018).

A major benefit of using version control is ensuring changes are well documented and previous versions are accessible if any changes must be recalled. Additionally, version control makes collaboration across projects much more robust.

Version control is not just for code either! It’s useful for scripts, documents, presentations, and books (like this DMP!). Instead of saving each version of a file with a different name (i.e., Report\_V01.docx and Report\_V02.docx; for a reminder on version naming, see [Section 4.0.1.5](#sec-version-numbers)), there’s only one file Report.docx which automatically has its history and editors saved with Git and GitHub.

In the screenshot below, you can see who made a commit (which is basically a named version of changes), when that commit was made, and you can click on the commit message to view all of the files that were changed.

|  |
| --- |
|  |

After clicking on the first commit message, we see the documentation.qmd file was changed with additions highlighted in green and deletions highlighted in red.

|  |
| --- |
|  |

### Git and GitHub resources

Read Jenny Bryan’s article [*Excuse Me, Do You Have a Moment to Talk About Version Control*](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\DataManagement\Resources\Bryan_ExcuseMeDoYouHaveMomentTalkVersionControl.pdf) for a great background on Git and GitHub, why we should be using it, and a brief how to get started. For detailed instructions , please follow along with Jenny Bryan’s free online book [*Happy Git and GitHub for the useR*](https://happygitwithr.com/). Another helpful book resource is [*GitHub: A Beginner’s Guide*](https://birdscanada.github.io/BirdsCanada_GitHubGuide/), which was created by Birds Canada (avian conservation NGO) for people without a lot of programming background.

If you prefer to look through slides, check out Byron C. Jaeger’s presentation [*Happier version control with Git and GitHub (and RStudio)*](https://bcjaeger.github.io/seminar-git/)*.*

## 6.4 Staff turnover

When staff leave our group, they take their skills, institutional knowledge, and personal understanding of their file management with them. Proper offboarding is essential to ensure knowledge isn’t lost, time isn’t wasted trying to recreate workflows, and projects keep moving.

Before the employee leaves, the Senior Soil Scientist and Data Scientist ensure that:

* Folders and files are moved from the employee’s personal drive to the shared drive. They are named and organized according to [Chapter 4](#sec-naming).
* Workflows and specific processes the employee was responsible for are well documented.
* Permissions and ownerships are transferred to the appropriate remaining staff.
  + [GitHub WSDA organization](https://github.com/WA-Department-of-Agriculture)
  + ArcGIS data products and online groups
  + Database credentials
  + Box.com folders

More resources and offboarding checklists from [Harvard Research Data Management](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\DataManagement\Resources\StaffTurnover) can be found in our [DataManagement shared drive](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\DataManagement\Resources\StaffTurnover).

# 7. Documentation

Documentation is the process of recording all aspects of project design, sampling, lab analysis, data cleaning, data analyses, data quality control and assurances procedures, and development of decision-support tools. Seem familiar? All of these aspects are also part of our data life cycle. We document our data throughout the life cycle to:

* standardize our procedures
* enable reproducibility
* establish credibility
* ensure others (including our future selves) use and interpret data correctly
* provide searchability

*All documentation (including this document) should be updated (and versioned) as procedures change and lessons are learned.*

Samples collected by WSDA for the SoSA should follow the procedures and standards described in the below documentation.

External data must have at least the documentation outlined in [Section 7.4](#sec-external-data) to be integrated into the SoSA dataset.

## 7.1 Project-level

Project-level documentation includes all descriptive information about the SoSA dataset, as well as planning decisions and process documentation. Documentation includes quality assurance project plans (QAPP), standard operating procedures (SOP), and other high-level documents (i.e., request for proposals, applications, meeting agendas/notes, etc.).

### Quality assurance project plan (QAPP)

The QAPP is the highest level of project documentation and covers everything from the project description, personnel roles and responsibilities, project timelines, data and measurement quality objectives, study design, and overviews of field, laboratory, and quality control.

Ours can be found in <Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\QAPP>, though it needs to be updated.

### Standard operating procedures (SOP)

SOPs provide detailed instructions for field, lab, or data processing procedures and decision making processes.

All sampling related SOPs can be found in <Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\SOPs>.

#### SoSA sampling

The purpose of this SOP is to specify the procedures for a typical site visit in which soil sampling is conducted to measure physical, chemical, and biological soil health indicators. Procedures include equipment preparation prior to sampling, best practices for filling out field forms, the selection of sampling locations, sampling protocols, sample handling and storage, and submitting samples to the lab. This SOP serves to ensure data quality by creating audit trails and enabling verification that data are present, complete, and accurate. Additionally, this SOP will be used to maintain consistent sample collection procedures throughout the state for WSDA employees and partners.

#### Quality control / quality assurance (QA/QC)

This SOP outlines the process for screening sample metadata and lab results for completeness, consistency, and quality. Procedures involve subject matter expertise, investigation, communication with sampling teams and labs, algorithmic quality control, and tagging sample results with quality codes (as shown in the below table). Data are then integrated into the statewide database according to a SOP not yet authored.

#### SOPs we don’t have (yet)

* assigning sample, producer, field IDs
* data cleaning
* data integration into database
* data storage
* external data integration

## 7.2 Dataset-level

Dataset-level documentation applies to lab results, sample locations, grower information, and management data. We use readmes and changelogs to document what each dataset contains, how they are related, potential issues to be aware of, and any alterations made to the data.

### Readme

readme files are plain text documents that contain information about the files in a folder, explanation of versioning, and instructions/metadata for data packages. These files are saved as .txt, not as MS Word documents that take longer to open and can only be opened on computers with Microsoft installed.

#### Describe contents of folder

For example, the \_completeDataset folder contains a [readme.txt](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\_completeDataset\readme.txt) that describes each of the files’ structure, contents, and other pertinent information, such as the data source (i.e. PRISM mean annual precipitation and temperature are 30-year normals from 1991-2020 at 800 m resolution).

#### Explain versions

Another example is the [readme.txt](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\2023_Sampling\LabData\Raw\readme.txt) in the 2023\_Sampling > LabData > Raw folder, which explains why there are two different versions of the lab results and where to find additional information.

#### Provide instructions

An example of a [readme.txt](https://app.box.com/s/jl714sbos0jus6mr9zesqsfpv718j783) that provides instructions on how to use the folder contents can be found in the [ArcGIS soil sample points box.com folder](https://app.box.com/s/4d6xg8q5dr1cjjoxver0g3h35d3s3bao) that is shared with outside partners.

### Changelog

Changelogs are also simple and concise plain text documents saved in a folder alongside data files to document any changes to the dataset.

At the bare minimum, the changelog.txt should contain:

* date of modification
* initials of who made the changes
* description of the changes

## 7.3 Variable-level

Variable-level documentation includes data dictionaries and code books, which are often talked about interchangeably. However, we’ll refer to the *data dictionary* as a tabular collection of names, definitions, and attributes about the variables in a dataset created (ideally) in the planning phase of the project before data are collected. In contrast, *codebooks* provide descriptive, variable-level information and univariate summary statistics to allow users to understand the contents of a dataset without opening it. The codebook is created or updated after data are collected, cleaned, and validated.

### Data dictionary

In a data dictionary, each row is a different variable and each column is a different attribute of that variable. With a data dictionary, any user should be able to properly interpret each variable in our data.

Our [dataDictionary.xlsx](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils_completeDataset\dataDictionary.xlsx) in the [\_completeDataset folder](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils_completeDataset) contains two tabs (labResults and sampleLocations) that describe the attributes of each variable.

### Codebook

Codebooks provide more information (i.e., existing values/ranges and summary statistics) than the data dictionary and can be used to understand a very high-level summary of the processed data. There are many R packages that generate codebooks; however, we have not implemented this type of documentation for our project yet.

Crystal Lewis gave the lightning talk [*A Comparison of Packages to Generate Codebooks*](https://github.com/Cghlewis/codebook-pkg-comparison). I’d like to generate codebooks for our datasets once they live in a database.

## 7.4 External data

External data refers to any data not directly collected by WSDA or trained partners (e.g., WSU or conservation districts) that follow our SOPs. These can include other studies pre-dating WaSHI, special soil health surveys, or publicly available datasets.

The Data Scientist and Senior Soil Scientist will decide whether to integrate an external dataset case by case by considering the below questions:

* How does the study design fit into SoSA goals?
* Who collected the soil samples?
* What field procedures were used and how were they documented?
* Who analyzed the soil samples? With which methods and QA/QC procedures?
* Which pieces of metadata and management data accompany the lab results?
  + Farm, producer and field info[[3]](#footnote-149)
  + Sampling date
  + Sampling depth
  + Latitude and longitude
  + Production system (current crop, crop rotation, etc.)
  + Tillage, livestock grazing, irrigation, soil fertility and amendments, conservation practices, etc.
* Is there a data dictionary or codebook to describe the variables and measurements, units, missing values, etc.?

Generally, the external data should **1)** be well documented, **2)** be collected and analyzed by well-trained scientists and labs; and **3)** have adequate accompanying metadata and management data to facilitate interpretation of the results.

Some publicly available datasets to consider are listed in <Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\DataSources>.

# 8. Data flows

## 8.1 Project structure

## 8.2 Original data

## 8.3 Working data

## 8.4 Clean data

# 9. Data sharing

Our data sharing policies promote FAIR principles such that our data are “*as open as possible, as closed as necessary*” (European Commission 2016). Data should be *open* enough to facilitate efficient re-use and avoid duplicating data collection efforts, enhance scholarly rigor, promote engagement across the research and public communities, and realize many other benefits (Whyte and Pryor 2011). However, data must also be as *closed* as necessary to protect grower privacy and to honor prior agreements with growers or other researchers.

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 9.2](#sec-maintain-confidentiality).

## 9.1 WaTech data categorization

Under the Washington state policy 141.10 (Securing Information Technology Assets), state agencies must classify data into categories based on the sensitivity of the data. WaTech provides [guidance](https://watech.wa.gov/Categorizing-Data-State-Agency) on the four categories of data.

#### Category 4: “Confidential information requiring special handling”

**WaTech**: Data requires strict handling requirements applied by statues (e.g. HIPAA) or regulations (e.g. rules on employee files).

**SoSA**: We don’t manage any data falling under this category.

#### Category 3: “Confidential information”

**WaTech**: Data includes “personal information” as defined in RCW 42.56.590 (security breaches) and RCW 19.255.010 (personal information disclosure). An individual’s first name or first initial and last name *in combination* with at least one of the following elements: social security number, driver’s license or Washington identification card number, or any account numbers that permit access to their financial account.

**SoSA**: While we do not collect any of the above elements in combination with grower names, we still protect individual and farm names, and latitude and longitude coordinates as confidential information.

#### Category 2: “Sensitive information”

**WaTech**: Data are intended for official use only and withheld unless specifically requested.

**SoSA**: Lab results and management surveys fall under this category. Access to this data requires a [data share agreement](#data-share-agreement).

#### Category 1: “Public information”

**WaTech**: Data is not covered in any of the above categories or is already released to the public.

**SoSA**: De-identified and aggregated data such as the number of soil samples and from which counties and crops they were collected fall under this category. For example, the [SoSA dashboard](https://nras.maps.arcgis.com/apps/dashboards/d4b2d135d47a4ff7a89285fecf67cef2) is publicly available and the map zoom is disabled at the 1:1,600,000 scale (counties level).

## 9.2 Maintain confidentiality

Under no circumstances should the following Category 3 data be released to external collaborators or made publicly available:

* farm name
* grower first and last name
* field names that contain street names or other identifying information
* latitude and longitude coordinates or other geospatial identifiers
* any information that identifies the individual farm or grower

Category 2 data should be anonymized and aggregated to honor our [data privacy statement](#data-privacy-statement) by either removing or replacing Category 3 confidential information with dummy data. The {randomNames} R package can be used to replace real names with fake names. Latitude and longitude should be rounded to a precision that does not identify the farm or fields sampled. See the example R script below, copied from the private [soils-internal GitHub repo](https://github.com/WA-Department-of-Agriculture/soils-internal/blob/main/R/anonymizeData.R).

R code to anonymize data

```{r}  
library(dplyr)  
library(readxl)  
library(janitor)  
library(randomNames)  
  
load("./data/\_completeDataset/2020-2023\_labResults\_wide.Rdata")  
load("./data/\_completeDataset/2020-2023\_labResults\_long.Rdata")  
crops <- read\_excel("./data/reference/crops.xlsx")  
points <- read.csv("./data/\_completeDataset/2020-2023\_sampleLocations.csv") |>  
 select(sampleId, longitude, latitude) |>  
 mutate(across(where(is.numeric), \(x) round(x, 0)))  
  
# Join lab data with gis data  
data <- left\_join(allResults\_wide, points)  
  
# Join data with crop dictionary to get crop groups  
data <- left\_join(data, crops, by = c("crop" = "Crop Type")) |>  
 select(-crop) |>  
 relocate("crop" = "Crop Group", .after = county)  
  
# Anonymize  
anonData <- data |>  
 # Farm name, producer name, sample ID  
 mutate(  
 farmName = paste0("Farm ", sprintf("%03d", cur\_group\_id())),  
 producerName = randomNames(which.names = "first",  
 sample.with.replacement = FALSE),  
 producerId = paste0(  
 paste0(sample(LETTERS, 3, replace = TRUE), collapse = ""),  
 "0",  
 paste0(sample(1:9, 1, replace = TRUE), collapse = "")  
 ),  
 fieldName = paste0("Field ", fieldId),  
 sampleId = paste0(  
 substr(year, 3, 4), "-", producerId, "-", fieldId  
 ),  
 .by = c(county, producerName, producerId)  
 ) |>  
 # county  
 mutate(  
 county = paste0("County ", cur\_group\_id()),  
 .by = c(county)  
 )  
  
# Grab 100 random samples  
anon\_subset <- anonData |>  
 slice\_sample(n = 100)  
  
write.csv(anon\_subset, file = "./data/\_completeDataset/exampleData.csv",  
 na = "",  
 row.names = FALSE)  
```

## 9.3 Data share agreement

SoSA CoPIs created a [Data Sharing and Scope of Work Agreement](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\DataSharing\StateoftheSoils_DataShareAgreement.pdf) that details the type of data to be shared, the scope of work in which the data may be used, and terms for using SoSA data.

Once the agreement has been signed by both CoPIs and the “Partnering Scientists”, the agreement should be saved in its own folder within [Y:/NRAS/Soil\_Health\_Initiative/StateOfTheSoils/DataSharing](Y:\NRAS\Soil_Health_Initiative\StateOfTheSoils\DataSharing). If this agreement is also part of a grant proposal, it should also be saved in its corresponding grant folder [Y:/NRAS/Soil\_Health\_Initiative/ContractsGrants/Grants](Y:\NRAS\Soil_Health_Initiative\ContractsGrants\Grants).

The request correspondence, code to subset the requested data, and the final dataset sent should all be saved in this folder. Internally documenting this information allows us to track publications and attributions resulting from this data sharing. This documentation also helps us track the broader impact of how our collaborators’ use the SoSA dataset.

## 9.4 Public access

Currently the only SoSA data publicly available are the counts of samples across the project, counties, and crop types displayed in the [ArcGIS Online Dashboard](https://nras.maps.arcgis.com/apps/dashboards/d4b2d135d47a4ff7a89285fecf67cef2).

A small, anonymized subset is included as example data in the [{washi}](https://github.com/WA-Department-of-Agriculture/washi) and [{soils}](https://github.com/WA-Department-of-Agriculture/soils) R packages for demonstration purposes.

**TODO**: discussion with team on the below ideas

In the future when the data are more mature and hosted in a proper database, we may publish an anonymized subset in a public repository such as:

* GitHub via an R package or Shiny app
* [Zenodo](https://zenodo.org/): integrates with GitHub and is citable with a DOI
* [Data.WA.gov](https://data.wa.gov/): open data portal for the State of Washington

More inspiration and ideas for enhancing data discoverability and sharing across the agricultural and soil health communities include:

* [USDA LTAR data dashboards](https://ltar.ars.usda.gov/data/data-dashboards)
* [CAF LTAR metadata tool](https://meta.cafltar.org/)

[Chapter 16](https://datamgmtinedresearch.com/share) of *Data Management in Large-Scale Education Research* discusses more considerations for data sharing and choosing public repositories (Lewis 2023).

## 9.5 Acknowledgments

All research and data 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 were 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.

# 10. Code Style Guide

## 10.1

Template snippet

Section break snippet

.R and .qmd file names

function names

argument names

variable names

https://indrajeetpatil.github.io/second-hardest-cs-thing/

# References

Bryan, Jennifer. 2018. “Excuse Me, Do You Have a Moment to Talk about Version Control?” *The American Statistician* 72 (1): 20–27. <https://doi.org/10.1080/00031305.2017.1399928>.

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

European Commission. 2016. “H2020 Programme Guidelines on FAIR Data Management in Horizon 2020.” <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>.

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>.

Whyte, Angus, and Graham Pryor. 2011. “Open Science in Practice: Researcher Perspectives and Participation.” *International Journal of Digital Curation* 6 (March): 199–213. <https://doi.org/10.2218/ijdc.v6i1.182>.

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. Code files include anything that gets pushed to GitHub where it becomes a URL. [Google recommends kebab-case](https://developers.google.com/search/docs/crawling-indexing/url-structure) because hyphens are better than underscores for URLs. [↑](#footnote-ref-89)
2. See [Chapter 10](#sec-code-style-guide) for the code style guide. [↑](#footnote-ref-91)
3. Enough farm, producer and field info to distinguish unique farmers and fields and assign IDs to. This info doesn’t need to be personally identifiable information. [↑](#footnote-ref-149)