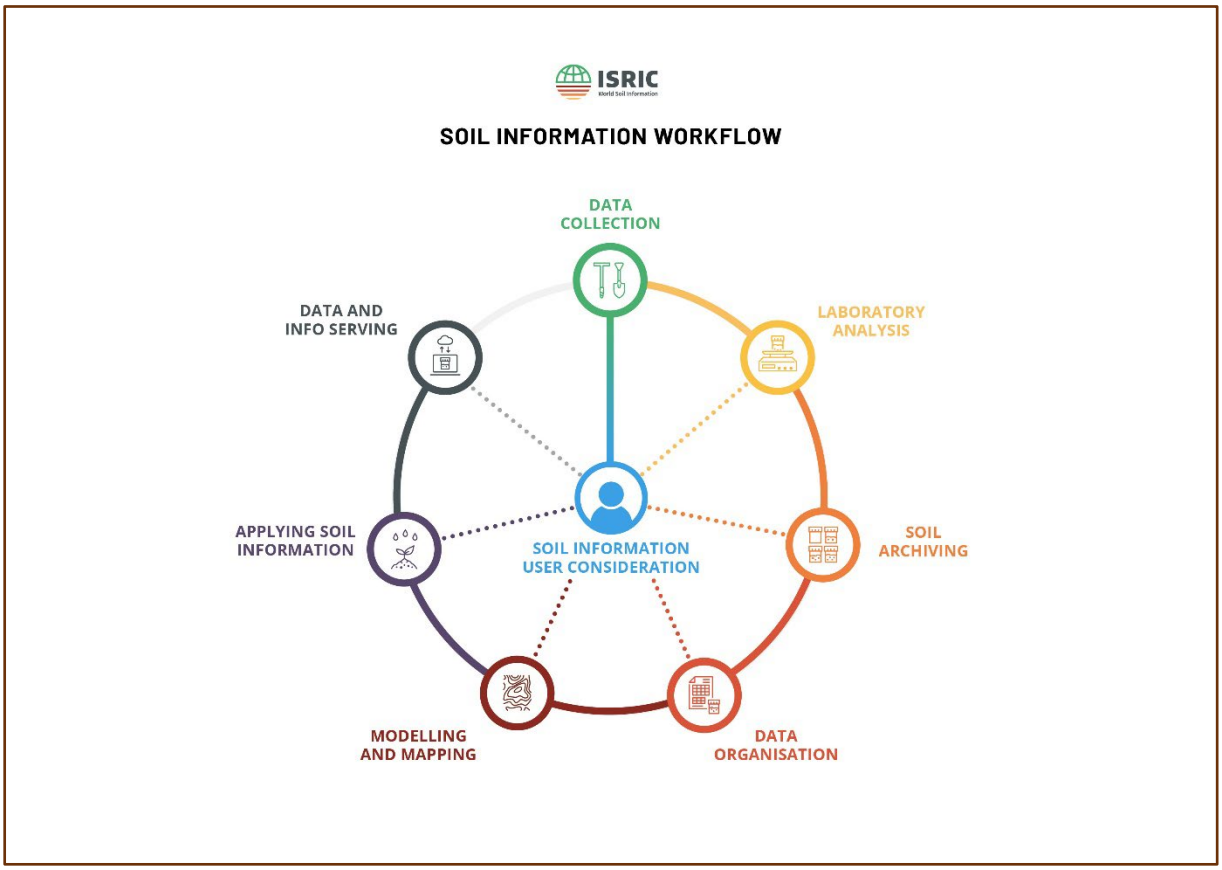
Basic principles for compiling a profile dataset for consideration in WoSIS

Niels H. Batjes

August 2022



Director, ISRIC - World Soil Information
Droevendaalsesteeg 3 (building 101)
6708 PB Wageningen
The Netherlands
E-mail: soils@isric.org

This document was prepared for the ISRIC Community of Practice, with reference to the WoSIS Procedures Manual 2020.

Additional information on ISRIC can be accessed through http://www.isric.org.

Citation

Batjes, N.H., 2022. Basic principles for compiling a profile dataset for consideration in WoSIS. CoP report, ISRIC – World Soil Information, Wageningen.

Contents

Summary

This document was prepared in the framework of ISRIC's Community of Practice for soil information providers. It outlines basic principles for compiling a profile dataset for consideration in WoSIS (World Soil Information Service). Templates for data entry are provided, but data providers can also submit datasets in their original format provided the basic observation and measurement (O&M) principles, as defined in this document, are met. More information about the various steps required for submitting profile data to WoSIS and their further processing are given in the '2020 WoSIS Procedures Manual' together with the workflows for handling, quality-assessing and standardising soil profile data.

1. Introduction

The World Soil Information Service (WoSIS), hosted by ISRIC – World Soil Information, aims to serve the user with a selection of standardised and ultimately harmonised soil profile data. The quality-assessed and standardised data may be used to underpin digital soil mapping and a range of global assessments. WoSIS is an important building block of ISRIC's evolving, searchable data infrastructure.

The aims of the World Soil Information Service are to:

- Safeguard world soil data 'as is' (especially for soil legacy data)
- Share soil profile (point) data upon their standardization and harmonisation
- Provide quality-assessed, standardised soil data for digital soil mapping and a range of environmental applications.

To be considered in WoSIS a soil dataset should include data for commonly required soil properties (Batjes 2016; FAO 2006; Schoeneberger *et al.* 2012; van Engelen and Dijkshoorn 2013), but no minimum dataset is prescribed. However, sufficient information (metadata) should be provided to assess the source and quality of the data as well as the license. A dictionary table describing the meaning of all (often abbreviated) column headings used in the dataset tables should be provided with the metadata themselves. Similarly, the use of dictionary tables (or thesauri) is recommended for describing all coded data entries as well as any abbreviated descriptive soil property value (e.g., 'W' means 'well drained' when defined according to the FAO (2006) Guidelines).

Currently, ISRIC is developing ETL (Extract, Transform, Load) procedures that extract, transform, and load soil data from multiple sources to WoSIS, as a unified data repository. Ultimately, the ETL procedures themselves will be made available as a web-based service to facilitate uniform data submission and cleansing.

2. Observations and measurements

Soil data should be consistently given as the result of observations and measurements (O&M). In this document, an observation (O) is the outcome of an 'act of measuring or otherwise determining the value of a property', while a measurement (M) is the outcome of a 'set of operations having the object of determining the value of a quantity' (OGC 2013).

Soil records are considered complete and thus processable into WoSIS when:

1) The lineage[1] of the soil record is well described, and
2) The soil data are consistently expressed as the result of observations and measurements (O&M). Those values, either numeric, categorical or descriptive, are expressed according to the associated domain as dictated by the references used for defining units of expression or pick lists.

[1] https://en.wikipedia.org/wiki/Data_lineage.

Typically, each soil profile is characterized by a number of consecutive layers or horizons, with defined upper and lower boundaries[2]. Alternatively, for soil fertility assessments only the upper 10 to 30 cm will be sampled.

3. Preparing your soil profile data

Often, soil profile data need to be prepared to facilitate their standardization in WoSIS. As indicated, sharing soil data for consideration in WoSIS does not require the use of a specific data entry template with a priori standards, nor is there a minimum dataset size. However, the general principles of describing the data should be followed so that they can be readily understood by the WoSIS database managers.

Below, we present a suggested template for compiling and submitting datasets (in case no digital dataset are available yet). The corresponding templates and a worked example, both in 'xlsx' and 'ods' format, can be downloaded from the ISRIC website[3].

The template consists of one spreadsheet with seven different sheets:

- read_me_first
- dataset
- profile
- layer
- column_definition
- standard_attribute
- standard_domain

By convention, the sheet and column names should not contain diacritical marks, symbols, spaces, upper-case characters and not start with a number.

Sheet **read_me_first** briefly explains the purpose and lay-out of the template.

Sheet **dataset** serves to describe the main characteristics of your dataset. It has the following rows:

- title: Dataset title, project or thesis title.
- version: Dataset version.
- publication_date: Publication date (yyyy-mm-dd).
- abstract : Concise description of the dataset.
- license: Access and use constraints of the dataset; please provide an equivalent to Creative Commons license (https://creativecommons.org/choose/).
- organization_name: Organisation name
- organization_url: Link to organisation or project web page.

[2] If the bedrock or an impenetrable layer is observed, this should be specified in the dataset to make the observation explicit.
[3] https://git.wur.nl/batje001/WoSIS/-/blob/master/Templates/.

- organization country : Organisation country.
- organization_city: Organisation city.
- organization_postal_code: Organisation postal code.
- Organization_delivery_point: Organisation address.
- author_first_name: Author first name (1st author).
- author1_last_name: Author last name (1st author).
- author1_email: Author email (1st author).
- authorX_first_name: Author first name (Xth author, repeat as needed).
- authorX_last_name: Author last name (Xth author).
- authorX_email: Author email (Xth).
- laboratory_name: Laboratory name.
- laboratory_country: Laboratory country.
- laboratory_city: Laboratory city.
- laboratory_postal_code: Laboratory postal code.
- laboratory_delivery_point: Laboratory address.
- Any further items may be entered after these rows when necessary.

Sheet **profile** serves to describe the main characteristics of each profile. It starts with the following columns:

- profile_code: Unique identifier of the profile as used in the source dataset.
- observation_date: Date of the observation in format (yyyy-mm-dd).
- coordinate_system: Coordinate system used. Please indicate the correspondent EPSG code (e.g., WGS 84, EPSG: 4326).
- x_coord: X coordinate, if in geographic coordinates (degrees), the same as Longitude.
- y_coord: Y coordinate, if in geographic coordinates (degrees), the same as Latitude.
- classification_system_name: The soil classification system used to classify the profile.
- classification_system_year : The publication year of the soil classification system used.
- profile_classification_name: Classification according to system and year previously defined.
- site_attribute 1: Add name for first attribute in given column (these are so-called site properties, e.g., drainage conditions).
- site_attribute 2: Add name for second attribute in given column.
- site_attribute x : Add name for next attribute.
- (...)

In case a profile has been classified according to several systems, for example (CPCS 1967),(FAO 1988) and (IUSS WG-SIS 2015), new columns for this can be added to sheet profile, for example as 'classification_system_name 2', 'classification_system_year 2' and 'profile_classification_name 2'.

Importantly, each row may only contain data for a given soil profile.

Sheet **layer** serves to describe properties for each (taxonomic) horizon respectively fixed-depth layer, for a given profile. It starts with the following columns:

- profile_code: Unique identifier of the profile. Provides the reference to data in sheet profile.

- layer_name: Horizon designation, for example A, B or C.
- sample_code: Laboratory sample code.
- upper_depth: Depth of upper layer or horizon (cm).
- lower_depth: Depth of lower layer or horizon (cm).
- layer_attribute 1: insert name for first attribute here (e.g., ph cacl2).
- layer_attribute 2: Insert name for second attribute.
- layer_attribute x: Insert name for next attributes.
- (...)

Any other layer description attributes can follow after these columns. For example, ph kcl, organic carbon, clay, silt, sand or bulk density. Note that each column can only contain values measured using one single analytical method, expressed using one uniform unit of measurement (to be specified in sheet column definition).

Importantly, each row in the **layer** sheet can only contain data for defined combinations of profile and layer (e.g., profile code and layer name).

Sheet **column definition** serves to describe all the columns (attributes) that have been specified earlier in the **profile** and **layer** sheets. It starts with the following columns:

- sheet name: Either 'profile' or 'layer'.
- column name: The exact name of the column added after the default (mandatory) ones.
- description: Description of the attribute.
- unit : Units used (e.g., cm), if not used, 'unitless' should be indicated (e.g., for soil pH).
- data type: Data type, use one out of ('Text', 'Integer', 'Real', 'Boolean', 'Date').
- analytical method: Analytical method used in the laboratory. If none is given enter 'Not specified'.
- desc_attribute_standard_ id : Name for standard attribute as defined for WoSIS, see sheet standard attribute.
- domain_name: Domain name from the **standard domain** sheet, when applicable (for categorical attributes only). Note: For regional data sets it may be necessary to create look up tables that permit to relate the 'region specific' domains to the standard domains adopted for WoSIS. For example, in case of a data set from China, the Chinese terms (characters) should be 'correlated' to their equivalent in English.
- New rows can be added to sheet **column definition**, as necessary for the given dataset

Table **standard_attribute** contains the standard attributes definitions as used in WoSIS. The exact names as listed in column *desc_attribute_standard_id* are needed to compile the look up tables mentioned earlier. Providing this crucial information, will make it possible for ISRIC to easily incorporate any newly submitted data into WoSIS. Otherwise, we will have to consult the data provider for additional information.

Table **standard_domain** contains the variables and categories (domains) as used in WoSIS; these are derived from the 4th edition of the 'FAO - Guidelines for Soil Description' (FAO 2006).

4. Submitting your profile data

Once compiled according to the above general principles, your data is ready to be sent to the WoSIS team for further processing.

Importantly, the citation of the data set and associated license should be provided to permit proper citation[4] and further ingestion/standardisation into WoSIS itself.

Any new data contributions will be processed as soon as possible, and this generally in sequence of their receipt at ISRIC.

In the near future, data providers will be able to make their soil (profile) data suited for ingestion into WoSIS using a set of web-based ETL procedures.

5. References

Batjes NH 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *GEODERMA* 269, 61-68.

CPCS 1967. *Classification des sols (Edition 1967)*, Grignon (FR), 87 p. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers16-03/12186.pdf

FAO 1988. *FAO-Unesco Soil Map of the World, Revised Legend, with corrections and updates (Reprint of World Soil Resources Report 60, FAO, Rome, 1988)*. World Soil Resources Report 60, FAO, Rome; reprinted with updates as Technical Paper 20 by ISRIC, Wageningen,1997, 140 p. http://www.isric.org/sites/default/files/ISRIC_TechPap20.pdf

FAO 2006. *Guidelines for soil description (Fourth ed.)*, FAO, Rome, 97 p. http://www.fao.org/docrep/019/a0541e/a0541e.pdf

IUSS WG-SIS 2015. IUSS Working Group on Soil Information Standards International Union of Soil Sciences (IUSS). http://www.soilinformationstandards.org/

OGC 2013. *OGC Abstract Specification Geographic information — Observations and measurements (ver. 2.0)*, Open Geospatial Consortium. https://portal.opengeospatial.org/files/?artifact_id=41579

Schoeneberger PJ, Wysocki DA, E.C. Benham and Soil Survey Staff 2012. *Field book for describing and sampling soils (ver. 3.0)*. National Soil Survey Center Natural Resources Conservation Service, U.S. Department of Agriculture, Lincoln (NE)http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_052523.pdf

van Engelen VWP and Dijkshoorn JA 2013. *Global and National Soils and Terrain Digital Databases (SOTER) - Procedures manual (Ver. 2.0)*. ISRIC Report 2013/04, IUSS, ISRIC and FAO, Wageningen, 198 p. http://www.isric.org/sites/default/files/isric_report_2013_04.pdf

[4] See https://www.isric.org/explore/wosis/wosis-contributing-institutions-and-experts