Always quote citation when using data!

Readme file (WoSIS Snapshot, December 2023)

https://doi.org/10.17027/isric-wdc-rgkv-d111

Data set citation:

Calisto, L., de Sousa, L.M., Batjes, N.H., 2023. Standardised soil profile data for the world (WoSIS snapshot – December 2023), https://doi.org/10.17027/isric-wdcsoils-20231130 .

Data set description:

The 'WoSIS 2023 snapshot' comprises data for 228k profiles from 217k geo-referenced sites that originate from 174 countries. The profiles represent over 900k soil layers (or horizons) and over 6 million records. The actual number of measurements for each property varies (greatly) between profiles and with depth, this generally depending on the objectives of the initial soil sampling programmes.

Supplement to:
Batjes N.H., Calisto, L. and de Sousa L.M., 2023. Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023). Earth System Science Data (https://doi.org/10.5194/essd-2024-14).

Data files:

The snapshot includes the following files, as described below. The type of data (e.g., character, integer, number, double, date) is listed under Type. The data type needs to be considered carefully when importing the TSV files into R or Excel, for examples see textbox below.

- *wosis_202312_observations.tsv*: This file, in 'tab-separated values' format[1], lists the four to six letter codes for each observation, whether the observation is for a site/profile or layer (horizon), the unit of measurement and the number of profiles respectively layers represented in the snapshot. It also provides the inferred accuracy for the laboratory measurements.

Column name	**Type**	**Description**
Code	character	Code for the observation
property	character	Description of soil property
procedure	character	Description of analytical procedure
unit	character	Standard unit of measurement (see Table A1)
profiles	integer	Number of profiles that have at least one measurement for the observation
layers	integer	Number of layers that have measurements for the observation
accuracy	double	Inferred accuracy of the laboratory measurements (%)

- *Wosis_202312_sites.*tsv: This file characterises the site location where profiles were sampled. The following field names are used:

Column name	**Type**	**Description**
Site_id	integer	Primary key
longitude	double	Longitude in degrees (WGS84)
latitude	double	Latitude in degrees (WGS84)
positional_uncertainty	character	Positional uncertainty of the profile's site location, expressed in classes
country_name	character	Name of country where site is located
region	character	Region in which site is located
continent	character	Continent in which site is located

- *wosis_202312_profiles.tsv*: Presents the unique profile ID (i.e. primary key), site_id, source of the data, country ISO code and name, positional uncertainty, longitude and latitude (WGS 1984), maximum depth of soil described and sampled, as well as information on the soil classification system and edition. Depending on the soil classification system used, the number of fields will vary. For example, for the World Soil Reference Base (WRB) system the options are publication year (i.e. version), reference_soil_group_code, reference_soil_group_name, and the name(s) of the prefix (primary) qualifier(s) respectively suffix (supplementary) qualifier(s). The terms principal qualifier and

[1] All columns are tab-delimited (tsv), with double quotation marks as text delimiters. File origin to be considered during data import/loading is '65001: Unicode (UTF-8)'.

supplementary qualifier are used since 2015 (IUSS Working Group WRB 2015, 2022); earlier WRB versions used prefix and suffix for this (e.g. IUSS Working Group WRB 2006). Alternatively, for USDA Soil Taxonomy, the version (year), order, suborder, great group, and subgroup can be accommodated (Soil Survey Staff 2014). The following field names are used:

Column name	**Type**	**Description**
Profile_id	integer	Primary key
Profile_code	character	Code for the profile
Dataset_code	character	Identifier for the source data set
Site_id	integer	Identifier for site where profile is located
positional_uncertainty	character	Positional uncertainty of the profile's site location (expressed in classes)
country_name	character	Name of country where site is located
longitude	double	Longitude in degrees (WGS84)
latitude	double	Latitude in degrees (WGS84)
wrb_reference_soil_group_code	character	Code for WRB group (in given version of WRB)
wrb_reference_soil_group	character	Full name for reference soil group
wrb_prefix_qualifiers	character	Name for prefix (i.e. for WRB1988)
wrb_suffix_qualifiers	character	Name for prefix (i.e. for WRB1988)
wrb_principal_qualifiers	character	Name for principal qualifiers (i.e. for WRB 2015 and WRB 2022
wrb_supplementary qualifiers	character	Name for supplementary qualifiers (i.e. for WRB 2015 and WRB 2022)
wrb_publication_year	integer	Version of World Reference Base for Soil Resources
fao_major_group_code	character	Code for major group (in given version of the Legend)
fao_major_group	character	Name of major group
fao_soil_unit_code	character	Code for soil unit
fao_soil_unit	character	Name of soil unit
fao_publication_year	integer	Version of FAO Legend (e.g. 1974 or 1988)
usda_order_name	character	Name of USDA Soil Taxonomy order
usda_suborder	character	Name of USDA Soil Taxonomy suborder
usda_great_group	character	Name of USDA Soil Taxonomy greatgroup
usda_subgroup	character	Name of USDA Soil Taxonomy subgroup
usda_publication_year	integer	Version of USDA Soil Taxonomy

- W*osis_202312_layers.tsv.* This file characterises the layers (or horizons) per profile:

Column name	**Type**	**Description**
Profile_id	integer	Primary key
Layer_id	integer	Sequential number for the layer (or horizon)
profile_code	character	Code for the profile
site_id	integer	Identifier for site where profile is located
layer_name	character	Name of pedogenetic horizon (“as is”)
upper_depth	integer	Upper depth of layer (cm)
lower_depth	integer	Lower depth of layer (cm)
layer_number	integer	Sequential number for the layer (or horizon)
organic_surface	logical	Flag for the presence of an organic layer above the mineral soil (f= false, t= true)
dataset_id	character	Abbreviation for source data set (e.g. WD-ISCN)
date	character	Date the profile was described/sampled (yyyy-mm-dd)
licence	character	Licence for observation as indicated by the data provider (e.g. CC BY)

- *Wosis_202312_xxxx.tsv.* For each observation (e.g. “xxxx” = “BDFIOD”), as defined under “code” in file *wosis_202312_observation.tsv,* the following are listed:

Column name	**Type**	**Description**
Profile_id	integer	Primary key
Layer_id	integer	Sequential number for the layer (or horizon)
profile_code	character	Code for the profile
layer_name	character	Name of pedogenetic horizon (“as is”)
upper_depth	integer	Upper depth of layer (cm)
lower_depth	integer	Lower depth of layer (cm)
organic_surface	logical	Flag for the presence of an organic layer above the mineral soil (f= false, t= true)
value	character [a]	Exported from array listing all measurement values for observation “*xxxx*” for the given layer. In some cases, more than one observation is reported for a given horizon (layer) in the source, for example four values for TOTC: [1:5.4, 2:8.2, 3:6.3, 4:7.7] (see value_avg below)
method_options	character [a]	Exported from array listing the method options for each analytical procedure as distilled from the source data. The content of this array varies with the soil observation under consideration as described in the method option table for each analytical procedure. For example, in the case of

		electrical conductivity (ELCO), the method options include sample pretreatment (e.g. sieved over 2 mm size, solution (e.g. water), ratio (e.g. 1:5), and ratio base (e.g. weight /volume). For details see Batjes and van Oostrum (2023) and related dashboard.
value_avg	double	Average, for above (it is recommended to use this value for "routine" modelling)
dataset_id	character	Abbreviation for source data set (e.g. WD-ISCN)
country_name	character	Name of country where site is located
longitude	double	Longitude in degrees (WGS84)
latitude	double	Latitude in degrees (WGS84)
positional_uncertainty	character	Positional uncertainty of the profile's site location, expressed in four classes (see Table 2)
region	character	Region in which site is located
continent	character	Continent in which site is located
date	character	Date the profile was described/sampled (yyyy-mm-dd)
licence	character	Licence for observation as indicated by the data provider (e.g. CC BY)

[a] **Important**: Two text fields in *method_options*, namely 'value' and 'method_option', contain complex text strings. These were concatenated from PostgreSQL arrays in WoSIS itself when the snapshot was created. See an example for 'soil pH' (taken from this tutorial):

```
{"instrument = [ electrode ] ratio = [ 1:1 ] sample pretreatment = [ sieved over 2 mm sieve ] solution = [ water [H2O] ]"}
```

How to read tsv files into R

First uncompress the ZIP file and specify the name of the uncompressed folder.

```
setwd("/YourFolder/WoSIS_2023_December/")    ## For example: setwd('D:/WoSIS_2023_December/')
```

Then use read_tsv to read the tsv files, specifying the data types for each column (c = character, i = integer, n = number, d = double, l = logical, f = factor, D = date, T = date time, t = time).

```
observations = readr::read_tsv('wosis_202312_observations.tsv', col_types='cccciid')
   observations          ## show columns and first 10 rows
sites = readr::read_tsv('wosis_202312_sites.tsv', col_types='iddcccc')
   sites                 ## show columns and first 10 rows
profiles = readr::read_tsv('wosis_202312_profiles.tsv', col_types='icciccddccccccicccciccccci')
   profiles              ## show columns and first 10 rows
layers = readr::read_tsv('wosis_202312_layers.tsv', col_types='iiciciiilcc')
   layers                ## show columns and first 10 rows
## Do this for each observation 'XXXX', e.g. file 'Wosis_202312_orgc.tsv':
orgc = readr::read_tsv('wosis_202312_orgc.tsv', col_types='iicciilccdccddccccc')
   orgc                  ## show columns and first 10 rows
```

Note: TSV files can also be imported into R without specifying the column names and types, see this short tutorial (R program). Another tutorial shows how tsv files can be read into an Excel spreadsheet.

- *wosis_202312.gpkg.* The data files are also available in GeoPackage format (which stores the files within an SQLite database).

Coding conventions:

Table A1. Coding conventions for observations (i.e. a combination of property, procedure and unit of measurement), and units (Codes are listed in alphabetical order; code also refers to the column name).

Code	Property	Procedure [a]	Unit
BDFI33	Bulk density fine earth [c]	Bulk density of a soil sample that has been desorbed to 33 kPa (1/3 bar)	kg/dm³
BDFIAD	Bulk density fine earth	Bulk density of a soil sample that has been air dried	kg/dm³
BDFIFM	Bulk density fine earth	Bulk density of a soil sample at field-soil water content at time of sampling	kg/dm³
BDFIOD	Bulk density fine earth	Bulk density of a soil sample that has been dried in an oven at 110 °C	kg/dm³
BDWSAD	Bulk density whole soil [e]	Bulk density of a soil sample that has been air dried	kg/dm³
BDWSOD	Bulk density whole soil	Bulk density of a soil sample that has been dried in an oven at 110 °C	kg/dm³
CECPH7	Cation exchange capacity	CEC estimated by buffering the soil at "pH7" (e.g., NH_4Oac)	cmol(c)/kg
CECPH8	Cation exchange capacity	CEC estimated by buffering the soil at "pH7" (e.g., NH_4Oac)	cmol(c)/kg
CFGR	Coarse fragments	Gravimetric content of soil material larger than 2 mm [c]	g/100g
CFVO	Coarse fragments	Volumetric content of soil material larger than 2 mm [c]	cm³/100cm³
CLAY	Clay [d]	Determination of total gravimetric content of clay-size fraction (for class-size limits and analytical methods see 'method_options')	g/100g
ECEC	Cation exchange capacity	Effective CEC conventionally approximated by summation of exchangeable bases (Ca^{2+}, Mg^{2+}, K^{+}, and Na^{+}) plus 1 M KCl exchangeable acidity (Al^{3+} and H^{+}) in acidic soils	cmol(c)/kg

ELCO20	Electrical conductivity	Electrical conductivity assessed on a 1:2 soil water extract. Used for saline soils.	dS/m
ELCO25	Electrical conductivity	Electrical conductivity assessed on a 1:2.5 soil water extract. Used for saline soils.	dS/m
ELCO50	Electrical conductivity	Electrical conductivity assessed on a 1:5 soil water extract. Used for saline soils.	dS/m
ELCOSP	Electrical conductivity	Electrical conductivity assessed on water saturated soil paste. Used for saline soils.	dS/m
NITKJD	Total nitrogen (N)	Kjeldahl wet-oxidation digestion procedure	g/kg
ORGC	Organic carbon (C)	Amount of organic carbon determined according to method specified under 'method_options'	g/kg
ORGM	Organic matter	Determination of organic compounds that accompany soil particles through a 2-mm sieve using loss-on-ignition (LOI) at about 400 degrees Celsius.	g/kg
PHAQ	pH	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ions (H^+) in water[a].	unitless
PHCA	pH	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ions (H^+), in the specified $CaCl_2$ solution.	unitless
PHKC	pH	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ions (H^+), in the specified KCl solution.	unitless
PHNF	pH	A measure of the acidity or alkalinity in soils, defined as the negative logarithm (base 10) of the activity of hydronium ions (H^+), in the specified NaF solution.	unitless
PHETB1	Phosphorus (P)	Phosphorus determined according to the Bray-I method, a combination of HCl and NH_4-F to remove easily acid soluble P forms, largely Al- and Fe-phosphates (mainly applicable for acid soils)	mg/kg
PHETM3	Phosphorus (P)	Determined according to Mehlich-3 method, a weak acid soil extraction procedure that is considered suitable for removing P and other elements in acid and neutral soil. The extract is composed of 0.2 M glacial acetic acid, 0.25 M ammonium nitrate, 0.015 M ammonium fluoride, 0.013 M nitric acid, and 0.001 M ethylene diamine tetraacetic acid (EDTA).	mg/kg
PHETOL	Phosphorus (P)	Phosphorus determined according to the Olsen method (0.5 M sodium bicarbonate ($NaHCO_3$) solution at a pH of 8.5); used extract P from calcareous, alkaline, and neutral soils.	mg/kg
PHPRTN	Phosphorus (P)	Phosphorus retention measured according to the New Zealand method (Blakemore, 1981).	g/100g

PHPTOT	Phosphorus (P)	Phosphorus determined with a "harsh" digest procedure to liberate and measure all forms of element.	mg/kg
PHPWSL	Phosphorus (P)	Phosphorus soluble in soluble in water	mg/kg
SAND	Sand	Determination of total gravimetric content of sand-size fraction (for class-size limits and analytical methods see 'method_options').	g/100g
SILT	Silt [f]	Determination of total gravimetric content of silt-size fraction (for class-size limits and analytical methods see 'method_options').	g/100g
TCEQ	Calcium carbonate equivalent (TCEQ)	Determination of the gravimetric loss of carbonates as carbon dioxide in the presence of excess hydrochloric acid. The quantity of carbonate (CO_3) in the soil is expressed as $CaCO_3$ and as a weight percentage of the less than 2 mm size fraction.	g/kg
TOTC	Total carbon (C)	Total C is quantified by two basic methods: wet or dry combustion (see 'method options'). In total C determinations, all forms of C in a soil are converted to CO_2 followed by a quantification of the evolved CO_2. Total C can be used to estimate the organic C content of a soil. The difference between total and inorganic C is an estimate of the organic C.	g/kg
WG0006	Water retention gravimetric	Water retention assessed at tension 6 kPa (see 'method options')	g/100g
WG0010	Water retention gravimetric	Water retention assessed at tension 10 kPa (see 'method_options').	g/100g
WG0033	Water retention gravimetric	Water retention assessed at tension 33 kPa (see 'method_options').	g/100g
WG0100	Water retention gravimetric	Water retention assessed at tension 100 kPa (see 'method_options').	g/100g
WG0200	Water retention gravimetric	Water retention assessed at tension 200 kPa (see 'method_options').	g/100g
WG0500	Water retention gravimetric	Water retention assessed at tension 500 kPa (see 'method_options').	g/100g
WG1500	Water retention gravimetric	Water retention assessed at tension 1500 kPa (see 'method_options').	g/100g
WV0010	Water retention volumetric	Water retention assessed at tension 10 kPa (see 'method_options').	$cm^3/100cm^3$
WV0033	Water retention volumetric	Water retention assessed at tension 33 kPa (see 'method_options').	$cm^3/100cm^3$

WV0100	Water retention volumetric	Water retention assessed at tension 100 kPa (see 'method_options').	$cm^3/100cm^3$
WV0500	Water retention volumetric	Water retention assessed at tension 500 kPa (see 'method_options').	$cm^3/100cm^3$
WV1500	Water retention volumetric	Water retention assessed at tension 1500 kPa (see 'method_options').	$cm^3/100cm^3$

[a] Method options for each analytical procedure, as described in Batjes and van Oostrum (2023), are provided in file *Wosis_202312_xxxx.ts.*

[c] Generally, the fine earth fraction is defined as being < 2 mm. Alternatively, an upper limit of 1 mm was used in the former Soviet Union and its satellite states (Katchynsky scheme). The actual size limits are specified under "method options" (see dashboard).

[d] Provided only when the sum of clay, silt and sand fraction is ≥ 90 and ≤ 100 percent (Note that users should normalise the totals to 100 percent before using them for mapping or modelling purposes; further, more stringent limits (e.g. ≥ 98 and ≤ 102) may be considered).

[e] No data are being served for this property because the associated licences are flagged as 'restricted' by the data providers.

[f] The lower and upper limits for the 'silt' size fraction can vary markedly between countries, hence these limits have been specified explicitly in WoSIS under "method options" (see Appendix B). Development and application of conversion procedures to one common "silt" fraction (e.g. 0.002-0.05 mm) is beyond the remit of the WoSIS project itself. The necessary pedotransfer functions should be developed (and tested) prior to generating particle size class related soil property maps for a given geography. Research in this direction is being undertaken by the SoilGrids team, based on the "best available" comparative datasets for calibration.

Table A2. Coding conventions and brief descriptions for soil classification, horizon designations and number of occurrences in the WoSIS 2023-snapshot.

Code	Description
CSTX	Classification of the soil profile according to specified edition (year) of USDA Soil Taxonomy, at least at soil order level
CWRB	Classification of the soil profile according to specified edition (year) of the World Reference Base for Soil Resources (WRB), at least at reference soil group level
CFAO	Classification of the soil profile according to specified edition (year) of the FAO-Unesco Legend, at least at major group level
HODS[a]	Horizon designations as provided in the source databases

[a] Where available, the "cleaned" (original) layer/horizon designation is provided for general information; these codes have not been standardised as they vary widely between different classification systems (Bridges 1993; Gerasimova *et al.* 2013). When no horizon designations are provided in the source data bases, we have flagged all layers with an upper depth given as being negative (e.g. -10 to 0 cm that is using pre-1993 conventions) in the source databases as likely being a shallow "organic surface" layer above a mineral soil layer.

Further information:

Our FAQ page (https://www.isric.org/explore/wosis/faq-wosis) provides answers to commonly asked questions.

Specific questions can be posted to a public mailing list, isric-world-soil-information@googlegroups.com; for this, you must first register to the mailing list at https://groups.google.com/forum/#!forum/isric-world-soil-information.

Contact:

Luis Calisto: Database management expert, ISRIC – World Soil Information (luis.calisto@isric.org)

Niels Batjes: WoSIS project lead, ISRIC – World Soil Information (niels.batjes@isric.org)

Data Provider:

ISRIC – World Soil Information (World Data centre for Soils) https://www.isric.org/

References cited in ReadMe file:

Batjes NH and van Oostrum AJM 2023. *WoSIS Procedures for standardizing soil analytical method descriptions*, ISRIC - World Soil Information, Wageningen, 46 p. https://doi.org/10.17027/isric-1dq0-1m83

Bridges EM 1993. Soil horizon designations: past use and future prospects. *CATENA* 20, 363-373. https://doi.org/10.1016/S0341-8162(05)80002-5

Gerasimova MI, Lebedeva II and Khitrov NB 2013. Soil horizon designation: State of the art, problems, and proposals. *Eurasian Soil Science* 46, 599-609. https://dx.doi.org/10.1134/S1064229313050037

IUSS Working Group WRB 2006. *World Reference Base for Soil Resources (2nd ed.)*. World Soil Resources Report 103, FAO, Rome, 145 p. http://www.fao.org/ag/agl/agll/wrb/doc/wrb2006final.pdf

IUSS Working Group WRB 2015. *World Reference Base for soil resources 2014 - International soil classification system for naming soils and creating legends for soil maps (update 2015)*. World Soil Resources Reports 106, Global Soil Partnership, International Union of Soil Sciences, and Food and Agriculture Organization of the United Nations, Rome, 182 p. http://www.fao.org/3/i3794en/I3794en.pdf

IUSS Working Group WRB 2022. *World Reference Base for soil resources 2022 - International soil classification system for naming soils and creating legends for soil maps*, International Union of Soil Sciences, Vienna (Austria), 284 p. https://www.isric.org/sites/default/files/WRB_fourth_edition_2022-12-18.pdf

Soil Survey Staff 2014. *Keys to Soil Taxonomy, 12th ed.* USDA-Natural Resources Conservation Service, Washington, DC.https://www.nrcs.usda.gov/wps/PA_NRCSConsumption/download?cid=stelprdb1252094&ext=pdf