1 Summary

More information, and introductions to the concepts of the **data tree** and corresponding **metadata** can be found at the project source repository:
`https://github.com/geophysics-ubonn/ubg_data_toolbox`

1.1 The DataTools in the Jupyter terminal

The following commands can be run in all standard terminal. We recommend to use a Jupyter Terminal. Note that the Python package *ubg_data_toolbox* must be installed. This can be done by executing:

```
pip install ubg_data_toolbox
```

You can also install the toolbox from within a Jupyter Notebook cell by executing:

```
!pip install ubg_data_toolbox
```

1.2 The Commands

The following commands are used to manage **data trees**:

- Check a single measurement (**m_***) directory:

  `dm_m_check_dir`

- Check a complete data directory (**dm_***) directory:

  `dm_check_dirtree`

- Add data (file(s) or directory) to a data tree, interactively:

  `dm_add`

- Initialise a new *metadata.ini* based on the directory structure:

  `dm_init_metadata`

- List all measurement directories

  `dm_list_measurements`

Usually command line options can be queried by appending ”-h” to the command. Example:

```
$ dm_add -h
usage: dm_add [-h] -t TREE -i INPUT [INPUT ...]

Add one measurement to a given data directory structure

options:
```

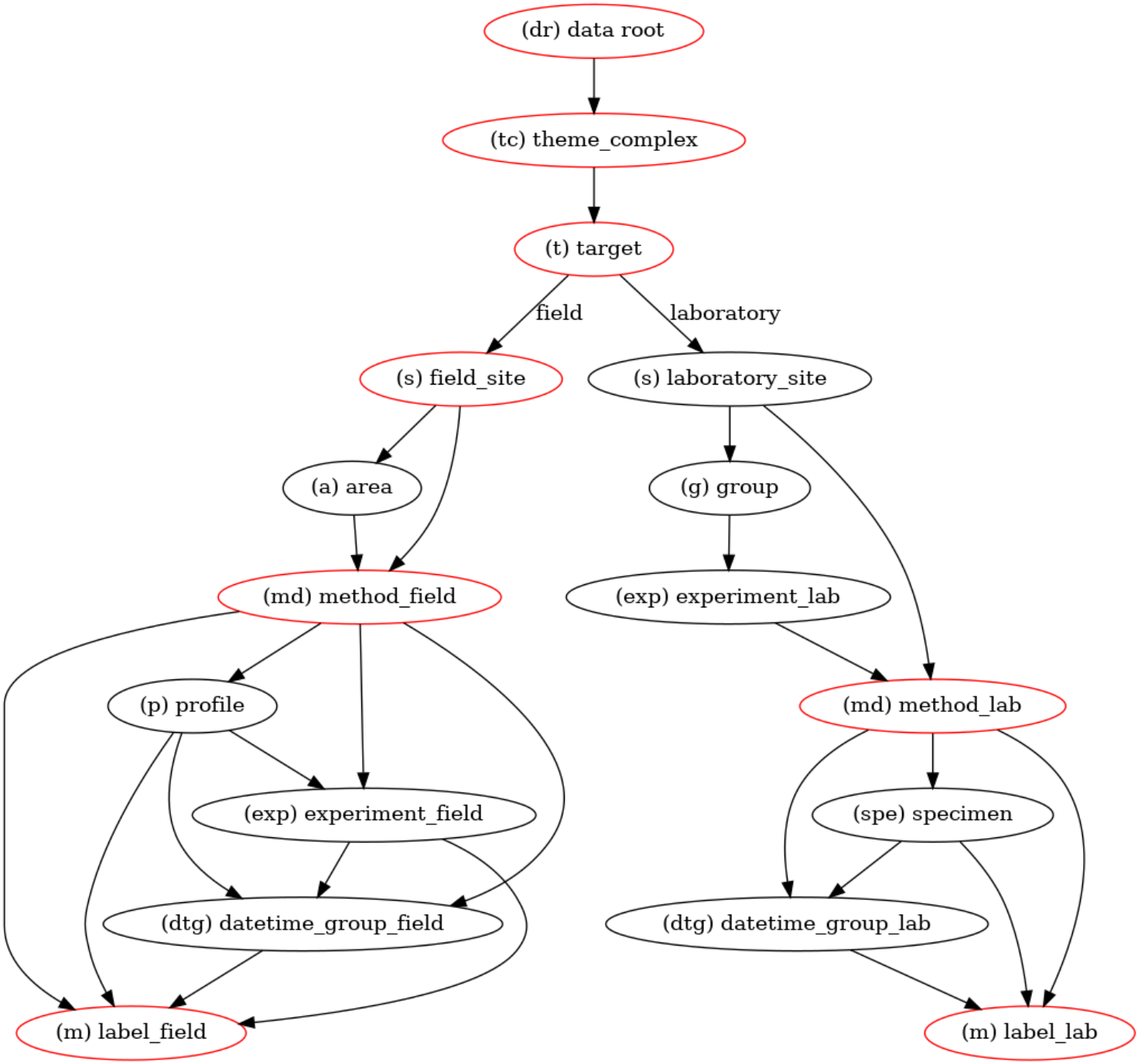
```
-h, --help            show this help message and exit
-t TREE, --tree TREE  Path of data tree (should start with: dr_
-i INPUT [INPUT ...], --input INPUT [INPUT ...]
                      Path to measurement (data/directory/directory tree)
```

1.3 Procedures

- In absence of a data tree, use **dm_add** to add measurements to a newly created directory structure
- For additional measurements, it usually is also convenient to keep on using **dm_add**. However, you can also use the following procedure:
  - Create the **m_*** in the correct place BEFORE creating the *metadata.ini* file.
  - THEN, use **dm_init_metadata** to initiate the *metadata.ini* file from the directory structure. Note that there may be missing, but required, metadata entries that can not be extracted automatically from the directory tree.
- Check the directory tree with **dm_check_dirtree** and fix any reported issues

2 Directory Structure

The directory structure is defined as follows. Each directory must start with a prefix, followed by an underscore, followed by a name/value. For example, the top directory could be called: **dr_datatree**. Note that some levels are optional (indicated by additional arrows in the figure below).



An example a directory tree (with only one measurement) is:

```
dr_datatree/
  tc_hydrogeophysics
      t_field
          s_Spiekeroog
              a_North
                  md_ERT
                      p_p_01_nor
                          m_01_p1_nor
                              metadata.ini
                              RawData
                                  data.dat
```

3 Metadata

Metadata is collected in *metadata.ini* files that reside in the individual measurement (m_)-directories. An example *metadata.ini* file could look like:

```
[general]
label = 20240610_ert_p1_nor
person_responsible = Maximilian Weigand
person_email = mw@domain.com
theme_complex = Hydrogeophysics
datetime_start = 20240610_1200
description = A small test measurement
    Note that some entries are multi-line capable!
survey_type = field
method = ERT
completed = yes

[field]
site = Spiekeroog
area = north
profile = p_01

[geoelectrics]
profile_direction = normal
```

You can add arbitrary *[sections]* and key=value pairs. However, the following set of metadata entries is pre-defined, with some of the **required**:

key	multi-line	required field	required lab	Doublin Core	description
section: [general]					
label	✖	True	True		Label of the individual measurement, This is the identifier for a given measurement at a given profile. Usually we construct the label using three parts: datetime, running number, one or two important keywords. Example: 20240516_01_p1_nor
person_responsible	✖	True	True	creator	The person that is responsible for this data set. This must not necessarily be the person that conducted the measurement.
person_email	✖	True	True		Email address of the person now maintaining this data set.
attending_persons	✖	False	False	contributor	All persons that were involved during the measurement. Optional: Add email addresses in parentheses, e.g. Maximilian Weigand (mweigand@geo.uni-bonn.de)
theme_complex	✖	True	True	subject	Theme complex that the measurement falls under. This is the most general category for a given measurement
project	✖	False	False	part of title	?
datetime_start	✖	True	True	date	Starting datetime of the measurement/measurements. Use date format YYYYmmdd_HHMM_s . YYYY: Year (e.g., 2004), mm: Month, dd: Day of month, HH: hour (1-24), MM: Minute (1-60), SS: Second Leave unknown parts out (e.g., seconds)
datetime_end	✖	False	False	date	Ending datetime of the measurement/measurements
description	✔	True	True	description	Description (should be short, comprehensive, and with links to detailed documentation)
survey_type	✖	True	True		Field or laboratory measurements? Allowed values: field, laboratory
method	✖	True	True	False	Which method(s) were used? (e.g.: ERT, SP, GPS, GPR)
experiment	✖	False	False		Label for the experiment that a measurement is assigned to
description_exp	✔	False	False		Description (should be short, comprehensive, and link to detailed documentation)
restrictions	✔	False	False	license	State any licensing restriction of the data set. Especially, note down any copyright owned by a party that is not the Department of Geophysics, Uni Bonn

completed	✖	True	True		States if the measurement series is finished or still ongoing. Possible values: yes, no
keywords	✖	False	False	subject	Keywords, separated by comma.
related_dois	✔	False	False	references	
missing	✔	False	False		?
problems	✔	False	False		Known restrictions/problems of the dataset (entries should be time stamped, multi-line entries required)
signed_off_by	✔	False	False		?
analysis_links	✔	False	False		?
dt_group	✖	False	False		Datetime group – Used to group measurements, e.g. into days or years
section: [field]					
survey_start	✖	False	False		Starting datetime of survey. Intended for the field data tree. Format: yyyymmdd hh:mm:ss
survey_end	✖	False	False		Ending datetime of survey. Intended for the field data tree. Format: yyyymmdd hh:mm:ss (same as survey_start)
site	✖	True	False		The general area of the measurement, e.g. a town name. This is further clarified in the metadata entries ”area”, ”profile”, ”coordinates”
area	✖	True	False		A more localized specification of the measurement area, e.g., an identifier of a certain field or street
profile	✖	True	False		The profile that was measured on. One common naming scheme consistent of the character ”p”,a running number, and a signifying key word. Example: p_01_nor.Use ”complete_area” for unspecific locations, i.e., whole-day gps measurements at one location

coordinates	✔	False	False		Coordinates of representative location(s) (i.e., starting point of measurement profile). One coordinate per line The use of WGS84 coordinates is preferred (EPSG 4326). Please state the use of other coordinate systems in the metadata entry "coordinates_desc". Coordinates should be included in decimal notation, with a least 6 decimal digits (ca. 5-12cm precision). See https://wiki.openstreetmap.org/wiki/Pre Format: One coordinate per line. Either two or three columns, separated by the character ";". The first two columns always are: latitude and longitude. An optional third column, "description" can hold identifiers, such as "start", "end", etc. A header column, starting with "#", is optional. Example: #lat;lon;description 50.706019097;7.210912815;start
coordinates_desc	✔	False	False		Description of coordinates. State used representation (e.g., WGS84 or UTM) here. Do not forget the UTM zone
section: [geoelectrics]					
spacing	✖	False	False		Electrode spacing
profile_direction	✖	True	False		Profile direction. Allowed values: normal, reciprocal
electrode_positions	✔	True	False		Electrode positions (x,y,z).The are the final electrode positions used for generating FE meshes. For 2D profiles, provide only (x,z) data.As the unit use meter [m].Note that somewhere in the metadata explanations on how initial coordinates (i.e. gps data) was transformed to yield these coordinates
section: [laboratory]					
site	✖	False	True		Laboratory measurement site
group	✖	False	False		High-level group of experiments
experiment_start	✖	False	False		Starting datetime of experiment. Intended for the laboratory data tree. Format: yyyymmdd hh:mm:ss
experiment_end	✖	False	False		Ending datetime of experiment. Intended for the laboratory data tree. Format: yyyymmdd hh:mm:ss (same as experiment_start)
specimen	✖	False	False		Sample material, e.g. sandstone; used mainly for laboratory measurement metadata.
permeability	✖	False	False		Permeability of sample material

porosity	✖	False	False		Porosity of sample material
section: [device]					
device	✖	False	False		Used measurement instrument.
device_serial	✖	False	False		Serial number of instrument, required if several devices of one type exist (e.g. the DT80)
programming	✖	False	False		Optional file path to a script/file containing the programming (script) used for the measurements(s)