Standard Data Formats

Part of Data Management Mibirem Project

Authors: *Mibirem data management group (in alphabetical order)*
B. Anadolu (DND), L. Antonielli (AIT), Sona Aseyednezhad (UU), S. Becarelli (UniPi), J. Beyert (Sensatec), F. Brogioli (DND), J. Brouchon (Altar), A. Cebron (CNRS), D. Donnerer (RTDS), J. Elsmoortel (UGent), D. de Graaff (TAUW), S. di Gregorio (UniPi), A. Kandyla (UU), T. Praamstra (TAUW), T. Reichenauer (AIT), S. Thijs (UHasselt), P. Vandamme (UGent), S. Vandersanden (UHasselt), A. Vauloup (CNRS), A. Zech (UU)

Date: 12. March 2024

1. Field data

Project Partner Involvement:

- linked to WP1 (part. T1.3) & WP5
- CNRS, Sensatec, DND, TAUW, UU, AIT, UniPi, UHasselt

Data types belonging to this group:

- Site descriptions (maps, geometry, locations, etc)
- on site measurements (e.g. PH)
- analytical measurements: environmental parameters = chemical and hydrogeological properties from lab based on field samples

Field site description:

- picture/map of site with sample location (preferably as pdf)
- accompanying file with coordinates (excel/csv) of wells → specify coordinate system (e.g. in meta-data)
- details on sampling locations (e.g. well specifics) in accompanying report
- optional (but highly recommended): photos/pictures of field site, sample locations etc

On site measurements/analytics

- File types:
  - excel files for on-site measurements and analytics (template created)
  - accompanied by readme-file
  - if available reports as pdf
- different files for soil samples and groundwater samples
- in-situ measurements and analytics in the same file (for the same samples), but in different tabs

- excel file naming: code of site (first part) as specified in the handbook + “_S” for soil and “_W” for water
- standard language: english → requires translation of Agrolab reports
- additional data from site (e.g. historical data/reports) → provided by partners,at start in report, if needed, make contact
- excel-file structure: unified header and column structure as well as units
    - in raw format:
        - no analysis performed in excel-sheet, no figures
        - data in matrix format (not different tables in one tab)
    - matrix structure with
        - parameters in columns
        - samples in rows
    - additional specifications
        - second row = units
        - column names: standard parameter names (following agrolab names of english report available from Dutch AgroLab branch, provided by TAUW)
        - column describing samples information:
            - sample name (encrypted, as specified in handbook, e.g. F_PLO_01_W)
            - separate columns for well, depth, date (if relevant aquifer belonging to)
    - different tabs for different type of information:
        - separate tabs for in-situ measurements and analytics
            - avoids confusion when parameters are measured in-situ and repeated by the lab
            - allows additional tabs if (in-situ) measurements are repeated (new data with new sample data)
        - if applicable: additional tabs for
            - metabolite analysis (if available)
            - isotope analysis (if available)

2. DNA related data

Project Partner Involvement:

- linked to WP2 (part. T2.1) & WP3
- lead: UHAS , CNRS, UniPi, AIT
- involved/informed: UGent

Summary of data types belonging to this group:

1. Molecular biomass (quantity of DNA extracted from samples)
2. bacterial 16S rDNA sequences data (metabarcoding : 16S rDNA sequences)
3. bacterial diversity data (effective and taxonomic affiliation)
4. gene quantification (abundance of 16S rRNA genes and other functional genes (degradation of contaminants) obtained through qPCR)
5. genomic data (shotgun metagenome sequences)

Specifics on data types belonging to this group:

1. Molecular biomass (quantity of DNA extracted from samples)
    - *excel file*

      with mass (dry weight) or volume of sample on which DNA extraction was performed, the volume of DNA eluted, purification step (yes/not), the concentration of DNA in ng/µL, quality control data (Absorbance, Ratio)
    - *tiff, jpg files : agarose gel*
    - *metadata file : word or text file with protocole of DNA extraction (Kit used), description apparatus we used for DNA quantification (qBit, Nanodrop, Quantifluor…), purification/precipitation steps (kit)*

2. bacterial 16S rDNA sequences data (metabarcoding: 16S rDNA sequences, form isolates)
    - *ab1 file*
    - *Fasta file (clean sequences : primer removed…)*
    - *Fastq files (2 files/sample)*
    - *metadata file : word or text file with info on DNA extraction Kit, PCR primer for amplification and for 16S sequencing, sequencing technology, which company did the sequencing…*

3. bacterial diversity data (effective and taxonomic affiliation)
    - *excel file : taxonomic affiliation of isolates (using EZBiocloud + date of analysis*
    - *file *.qza (qiime2: 1 file for taxonomy, 1 file for ASV table)*
    - *file*.csv metadata file*
    - *metadata file : word or text file with info on how reads were treated (QIIME, DADA2, Mothur….)*
    - *csv or excel table (otu_table or ASV_Table: relative abundance or reads number for each OTU or ASV for each samples)*
    - *csv or excel table (sample_data: differents variables for each samples)*
    - *csv or excel table (tax_table: Taxonomic classification table with Kingdom, Phylum, Class, Order, Family, Genus, Species affiliation of OTU or ASV for each samples)*
    - *csv or excel table (with calculated alpha diversity indices for each samples)*

4. gene quantification (abundance of 16S rRNA genes and other functional genes (degradation of contaminants) obtained through qPCR)
    - *pdf file for qPCR apparatus report (info on the date of the qPCR, the amplification protocol, the samples, the standards…)*
    - *excel file (same format or same file than for Molecular biomass) : with all info from file in 1. plus the Cq value, the gene concentration in number of gene copies per µL, and the calculation of the number of gene copies per ng of DNA and per g of soil;*
    - *metadata file : word or text file with info of which qPCR mix used, which standard used, which PCR primer used…*

5. genomic data (shotgun metagenome sequences)
    - *Fastq files*

3. Laboratory Experiments

Project Partner Involvement:

- linked to WP2 & WP3
- lead: AIT, UHasselt, UniPi, CNRS, UGent
- involved: Sensa, Altar, DND

Data formats specification for Microcosm-experiments (degradation tests)

- Degradation dynamics of original compounds: xlxs/csv, tif, pdf (from analysis lab)
- generation of intermediate products and end-products:xlxs/csv, tif, pdf
- Time-dependent CO2 accumulation: xlxs/csv, tif
- Changes of soil pH: xlxs/csv, tif
- Oxygen consumption: xlxs/csv, tif
- DNA extraction: xlxs/csv
- 16S rRNA gene amplicon sequencing: fastq, xlxs/csv, tif, PDF
- Quantitative PCR: xlxs/csv, tif, PDF

Data formats specification for Enrichment cultures

- Growth curves (protein measurement): xlxs/csv, tif, PDF
- Degradation dynamics of original compounds: xlxs/csv, tif, pdf
- generation of intermediate products and end-products:xlxs/csv, tif, pdf
- DNA and RNA extraction: xlxs/csv
- 16S rRNA amplicon sequencing: fastq, xlxs/csv, tif, PDF
- Metagenomics and metatranscriptomics: fastq, xlxs/csv, tif, PDF
- Optical density: xlxs/csv
- Cell numbers (e.g. dilution series): xlxs/csv
- Pictures of sample plates and microscopy: .jpeg
- pH progression xlxs/csv
- list of cultures: xlxs/csv

Data formats specification for Stable Isotope Probing (SIP)

- Degradation dynamics of 13C- and 12C-labeling compounds: xlxs/csv, tif, PDF
- DNA extraction: xlxs/csv
- SIP fractionation: xlxs/csv, tif
- 16S rRNA gene amplicon sequencing: fastq, xlxs/csv, tif, PDF
- Metagenomics: fastq, xlxs/csv, tif, PDF

Data formats specification for soil bactraps

- Time-dependent CO2 accumulation: xlxs/csv
- DNA extraction: xlxs/csv
- Quantitative PCR: : xlxs/csv
- Final protocol: PDF

Data formats specification for Isolation of contaminant degraders from enriched sample [UGent]

- List of selected samples: .csv
- List of isolates & data: .Excel
  - Isolate number
  - Isolate description: colony morphology & color
  - Sample: sample name, date of plating
  - Isolation conditions: date of isolation, isolation conditions (medium, temp, AE/AN)
  - Subcultivation: date of plating on special medium (axenic culture generation 1), general medium used, date of plating on general medium (axenic culture generation 2), date of plating on general medium (axenic culture generation 3)
  - MALDI: date pellet prep, days grown before harvest, spotset, target plate number
  - Preservation: date storage in glycerol, medium used for storage
- MALDI-TOF MS:
  - Input file: .xlsm
  - Output file:
    - Full MS spectra .txt
    - Report with top 10 hit list per isolate: .pdf
- Dereplication
  - SPeDE
    - .csv file (overview per isolate of spectrum quality, chosen as reference yes/no, reference number, reference group, isolate name, Bruker tophit en score)
    - .csv file used to add isolate and measurement data into BioNumerics
    - .txt file per spotset used by SPeDE for peak-based data analyses
    - .txt file per spotset used by SPeDE for Pearson peak-based data analysis
    - .txt files per spotset used by SPeDE for Pearson peak-basd data analysis and comparison with MSP present in our database
    - .txt files per spotset with peak data regridded required for SPeDE analysis
    - .csv pairwise matrix with the number of Unique Spectral Features for all isolates analysed with SPeDE
    - .csv overview of references and % of isolates matched to each reference spectrum
    - .csv file to link sample codes (UGent workflow) to isolate name
  - BioNumerics
    - .xlsx file of metadata fields - including manual clustering based on the UPGMA dendrogram that BioNumerics calculates and identification of isolates per cluster based on species consistency (top 10 hits with databases) and identification of other cluster members
- Selected isolates for further analysis .csv

Data formats specification for preservation of microbial consortia

- Description of aliquots .csv or .xlsm

  - Name, preservant added, freezing method, timing for analysis (time-point), aliquot number per time-point, date FCM analysis, date pellet storage, TCC numbers in triplicate, LDC numbers in triplicate, remarks
- Flow cytometry
  - .fsc files per tube run
  - .csv file eport of experiment statistics
  - .csv file cleanup of statistics via Rstudio
  - .csv sample list per experiment
- DNA data
  - see T1.4, point 2

Data formats specification for deposit of bioremediation bacteria and microbiomes

- Overview file .csv