SIDB: The Soil Incubation Database version 1.0

Carlos A. Sierra and others

Abstract

Dynamics of carbon release over time are generally assessed using soil incubation studies, where a certain amount of soil organic matter is left to decompose under controlled environmental conditions, and the temporal dynamics of released CO_2 are monitored over time.
The Soil Incubation Database SIDB, is a compendium of time series of CO_2 evolution from multiple published studies. It compiles metadata and time series data in human and machine readable formats that can be imported into different programing platforms for analysis. Metadata for each entry is stored in `yaml` files that contain: 1) information about the study and the creator of the entry, 2) physical an ecological information about the study site, 3) description about the incubation setup and treatments, 4) description about the specific variables in the `data` file that corresponds to the study. The `data` file is stored in comma delimited format `csv`, and contains information about the time of measurement and the amount of CO_2 release. Information about units of the data are stored in the corresponding metadata file. To facilitate data use, we provide an `R` package that can load all or particular entries from the database and makes them available as an R `list` for analysis. A website developed with `Jekyll` provides general information about the project. All source code for the database, the R package, and the website is publicly available on `GitHub`. We invite contributions from the soil science and ecological research communities to contribute data, code, and perform analysis from this database.

Introduction

- Introduce concepts of SOM decomposition dynamics and the use of incubation studies to address this topic.
- Motivate the need for synthesis of studies
- Goals of database and manuscript

The SIDB open source software project

The Soil Incubation Database version 1.0 is conceived as an open source software project, which not only provides open access to data, but also provides software for reading and analysis of data as well as documentation for future use and development. Therefore, the SIDB project is managed under a version control system centrally stored in GitHub. The structure of the repository is as follows

```
SIBD project
|   Readme.md
|-- data
    |-- entry1
        |-- metada.yaml
        |-- data.csv
    |-- ...
|-- docs
    |-- _config.yml
    |-- index.html
    |-- _layouts
    |-- _includes
    |-- assets
    |-- css
|-- Rpkg
    |-- DESCRIPTION
    |-- NAMESPACE
    |-- R
    |-- man
```

The three main folders, `data`, `docs`, and `Rpkg` provide access to the database, the website, and the R package, respectively. These are described in detail in the sections below.

The Database

The soil incubation data is stored in the `data` folder. This folder contains all entries in the database. Within each subfolder there are two files containing both data and metadata for each incubation study. The metadata file has the extension `.yaml`, and the data is stored in a comma separated file `.csv`. The name of each subfolder follows the convention: `AuthornameYEARJournalAbbrv`.

The metadata file

The metadata file is simply a text file that includes all relevant information about the incubation study. The yaml format is both human and machine readable, so it is very easy to write all relevant information about a study in these files. You can inspect the available entries for examples on how to write yaml metadata files.

Each file must contain the following basic information:

```
citationKey: Smith2017SBB
doi: 10.1016/2017.03.025.2
entryAuthor: John Stewart
entryCreationDate: 2014-02-19
```

In the yaml format, each separate field has a name followed by : and a value. The citationKey field is simply an ID to cite or identify the name of each entry. It has the same convention as the name of the folder for each particular entry. The doi field is the digital object identifier for the publication where the data was originally obtained. This doi can be used to automatically retrieve information from other databases such as crossref. The entryAuthor field is the name of the person who created the entry. entryCreationDate is the date at which the entry was created and uploaded in the database. This date follows the convention YYYY-MM-DD.

In yaml it is possible to create hierarchies with different type of information. This is very useful to store diverse fields related to the research sites where the soils were sampled for the incubation. For each metadata file, it is required to include a siteInfo field as follows

```
siteInfo:
    studySite: central Piedmont region of North Carolina, USA
    ecosystemType: forest
    climate: temperate
    soilType: Alfisol
    texture:
         percentClay: 40
         percentSilt: 20
         percentSand: 40
    coordinates:
         latitude: 36.6166667
         longitude: -79.150
```

Notice that there is an indentation pattern in this example. Indentation is used by yaml to create subfields within a field. So, studySite is a subfield of siteInfo, and latitude is a subfield of coordinates. You can create subfields for any field in case you need to add additional information such as different study sites with different climates. More subfields can be added to the siteInfo subfield as necessary.

Another important field that must be added is the incubationInfo

```
incubationInfo:
    desc: "Soils were incubated at three different temperatures: 4, 22,
           and 40 degrees Celsius as well as ambient and elevated CO2"
```

```
  treatments:
    - temperature
    - moisture
    - CO2
  incubationTime:
    time: 48
    units: days
```

In this example, the `incubationInfo` field has a subfield with a description `desc` on how the incubations were carry out. This is very important information to document details about the incubation experiment. It also includes the subfield `treatments` that lists the different variables that were modified by the treatments. Notice that each treatment is preceded by a `-`, which indicates that each element corresponds to an array.

The last field that must be added is the `variables`. This field contains, in order of appearance, the variables in the `.csv` file that is included in the same folder with the incubation data.

```
variables:
    V1:
     name: time
     units: days
    V2:
     name: T1
     temperature: 4
     moisture: 30
     CO2: 400
     units: "mg CO$_2$ g^{-1} soil day^{-1}"
     desc: "CO2 production rate measured at 4 degrees celsius, 30 % vwc,
        and 400 ppm"
    V3:
     name: T2
     temperature: 22
     moisture: 10
     CO2: 800
     units: "mg CO$_2$ g^{-1} soil day^{-1}"
     desc: "CO2 production rate measured at 22 degrees celsius, 10 % vwc,
        and 800 ppm"
```

The number of variables `V` in this field must correspond to the number of variables in the `.csv` file.

Data

The `data.csv` file for each entry in the database contains the time series of incubation data in a comma separated values format. The first column of the data file must contain the times at which CO2 measurements were made. Subsequent columns must contain the respiration measurements. The format of the data is irrelevant (eg. units, order of treatments) as long as the relevant information to identify each respiration column is described in the `variables` field of the metadata file.

The website

Documentation of the project, which includes the database and the R package, is presented in the project's website (https://soilbgc-datashare.github.io/sidb/). The source code of the website is stored in the `docs` folder. We use Jekyll to build the website from a set of files where content and layout are clearly separated. All content is either written in `Markdown` or `yaml` formats. Basic information about the configuration of the site is stored in the `_config.yml` file. Basic layouts and `html` formatting styles are included in the `_layout` and `_includes` folders, while pictures and `css` formatting styles are stored in the `css` and `assets` folders, respectively.

To serve the website locally, users can run the following code inside the `docs` folders

```
jekyll serve
```

provided a recent installation of Jekyll is available. The site will be served at a local host and can be viewed in any web browser.

The website is publicly served by `GitHub Pages`. Everything new changes are pushed to the SIDB repository, the website is build and served automatically by GitHub.

The R package

Data in the SIDB are stored in a format that can be read in any programming language. We provide an R package to allow users to read the database in R and facilitate analysis. To install the package, open R and run

```
install.packages("devtools")
devtools::install_github('SoilBGC-Datashare/sidb/Rpkg/')
```

Currently, two functions are provided: `loadEntries.R` and `readEntry.R`. As their name suggest, `loadEntries.R` collects all metadata and data from all entries and produces an `R list` with the entire database. The function `readEntry.R` reads individual entries from the database and also produces an `R list`.

Summary statistics of the entries in version 1.0

The number of entries in version 1.0 of the database is 36. Most entries have multiple time series of CO_2 flux release from incubation experiments. The current total number of time series is 561, and the total number of data-points is 10951.

The spatial distribution of the study sites for all entries is summarized in Figure 1. Most entries are concentrated in temperate ecosystems of the northern hemisphere, and tropical ecosystems are under-represented in this version.

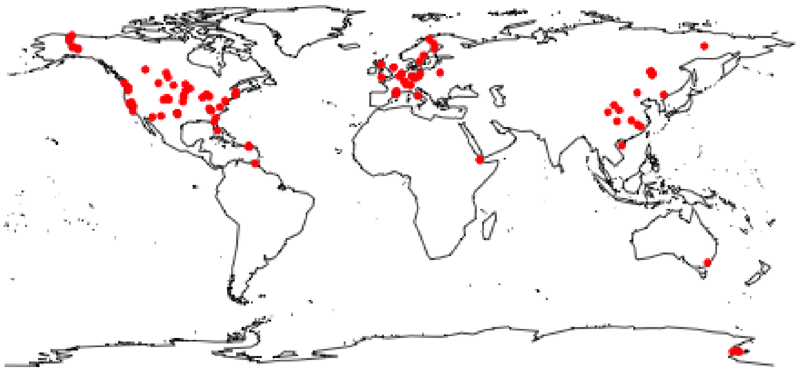

Figure 1: Spatial distribution of study sites for all entries in version 1.0 of the database

The database contains studies with a wide range of incubation times, from 1 to 924 days. A histogram of the incubation time for all entries is presented in Figure 2.

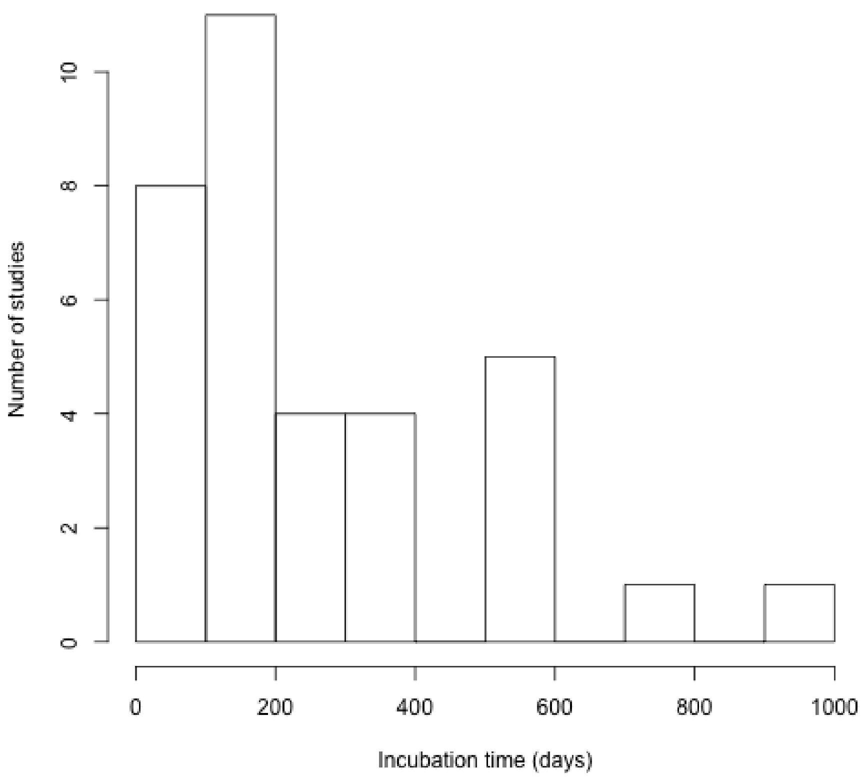

Figure 2: Frequency distribution of incubation times for all entries in version 1.0 of the database

Conclusions