Bringing Together Brazilian Soil Scientists to Share Soil Data

Alessandro Samuel-Rosa[1]; Ricardo S. D. Dalmolin[2]; Paulo Ivonir Gubiani[2]; Wenceslau Teixeira[3]; Stanley R. de M. Oliveira[4]; João Herbert M. Viana[5]; Carlos G. Tornquist[6]; Lúcia Anjos[7]; José João L. L. de Souza[8]; Eloi Ribeiro[9]; Marta Ottoni[10]; Paula S. C. de Medeiros[11]; José Miguel Reichert[2]; Diego S. Siqueira[12]; José Marques Júnior[12]; José A. M. Demattê[13]; André C. Dotto[13]; Leonardo Collier[14]; Gustavo M. Vasques[3]; Gustavo Valladares[15]; Fabrício A. Pedron[2]; João C. Pedroso Neto[16]; José M. F. Alba[17]; Ronaldo P. de Oliveira[3]; João Henrique Caviglione[18]; Pablo Miguel[19]; Humberto G. dos Santos[3]; Carlos A. Flores[17]; Igo Lepsch[13]; Diego José Gris[2]; Nícolas Augusto Rosin[2]; Jean M. Moura-Bueno[2]

[1] Estudante, Universidade Federal de Santa Maria, Av. Roraima 1000, Cidade Universitária, Bairro Camobi, Santa Maria - RS, CEP 97105-900, Brasil, e-mail: alessandrosamuelrosa@gmail.com; [2] Professor/Professor/Professor/Estudante/Estudante/Estudante, Universidade Federal de Santa Maria; [3] Pesquisador, Embrapa Solos; [4] Pesquisador, Embrapa Informática Agropecuária; [5] Pesquisador, Embrapa Milho e Sorgo; [6] Professor, Universidade Federal do Rio Grande do Sul; [7] Professora, Universidade Federal Rural do Rio de Janeiro; [8] Professor, Universidade Federal do Rio Grande do Norte; [9] Pesquisador, ISRIC - World Soil Information; [10] Pesquisadora, Serviço Geológico do Brasil; [11] Pesquisadora, Instituto Brasileiro de Geografia e Estatística; [12] Estudante/Professor, Universidade Estadual Paulista; [13] Professor/Estudante, Universidade de São Paulo; [14] Professor, Universidade Federal de Goiás; [15] Professor, Universidade Federal do Piauí; [16] Pesquisador, Empresa de Pesquisa Agropecuária de Minas Gerais; [17] Pesquisador, Embrapa Clima Temperado; [18] Pesquisador, Instituto Agronômico do Paraná; [19] Professor, Universidade Federal de Pelotas.

INTRODUCTION

A great deal of soil data has already been produced as part of soil survey and research projects. Most of the published datasets are published as a single paper, and the primary data is unavailable to other researchers. If not published soon, then the data probably will remain unpublished forever. As data underutilization is a waste of resources and refrains the advancement of knowledge, many isolated soil data rescue and sharing efforts have emerged (ARROUAYS et al., 2017). But a consistent solution to the problem of permanently safeguarding and promoting the reusability of all kinds of soil data has yet to be developed.

Lately, soil scientists have increased their concerns with data discoverability and reusability, the first two of the Three Laws of Open Data (EAVES, 2009). Discoverability is the ability of a dataset to be found by someone else. Reusability is the ability that a dataset has to be used again by its producer and/or someone else. Both discoverability and reusability are critical to ensuring the reproducibility of the research, a basic principle of the scientific method.

Brazilian soil scientists have recently created a soil data repository using community-built standards and following open data policies in an attempt to address the issues mentioned above. The Free Brazilian Repository for Open Soil Data – **febr** –, accessible through www.ufsm.br/febr, is a centralized repository targeted at storing open soil data and serving it in a standardized and harmonized format, for various applications. This paper describes the features of **febr** and the opportunities that it creates for soil science.

FEATURES

Data model

Unlike existing soil databases, **febr** was designed to allow individualized management of datasets. First, because such a design highlights datasets authors, helping them to be properly acknowledged and cited by others. Second, because it gives the flexibility to accommodate many types of data of any soil variable. This is accomplished by storing datasets using a directory structure, each dataset being in its own directory.

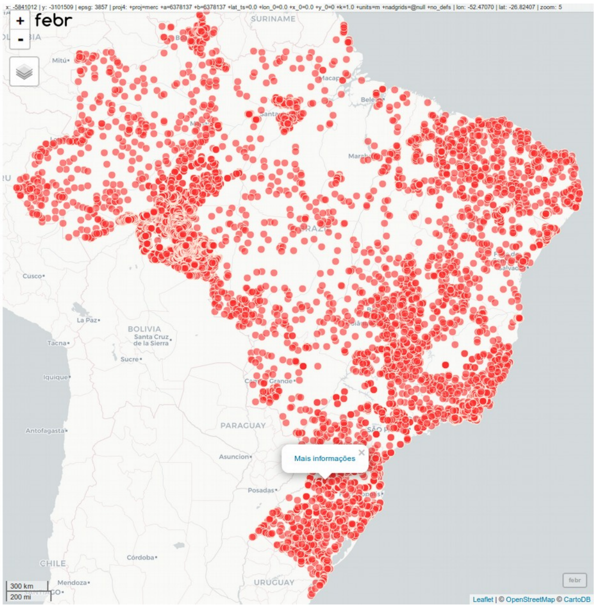
The data model used in **febr** to organize each dataset takes into consideration that soil observations generally have four operational dimensions. The first two are the *x* and *y* horizontal spatial coordinates, referring to some predetermined standard coordinate reference system (CRS). The third is the temporal coordinate, *t*, the moment – according to some predetermined standard calendar and time zoning system – when the soil was observed. The fourth and last operational dimension is the soil observation depth, *z*, as measured using some predetermined standard scale, e.g. metres. At a point in (geographic) space and time, [*x*, *y*, *t*], or in space, time and depth, [*x*, *y*, *t*, *z*], a soil observation is accompanied by an attribute space. The latter is a multi-dimensional space defined by a set of attributes of the environment (land use, slope, parent material) or soil layer (pH, cec, carbon content), respectively.

Two tables per dataset are used to cope with the multi-dimensional character of soil data: 'observacao', for space-time data, and 'camada', for space-time-depth data. The relation between them is established using a identification key included in both tables – the column 'observacao_id'. A third table, 'metadado', is used to store the data about the methods used to produce the soil data, i.e. the metadata. A fourth table, 'dataset', stores data about the dataset as a whole (general description, author name and contact, version). Last, all **febr** maintainers tasks (changes, improvements, todo list) are recorded in table 'tarefa'.

The directory structure of **febr** is implemented in Google Drive, while spreadsheets (Google Sheets) are used to store data tables. Spreadsheets are familiar to any soil scientist, eliminating the hurdle of having to learn a new software and/or data structures. For this reason, it is very easy to enter, manipulate, and visualize soil data in **febr**. It facilitates the participation of soil students and experts, as well as non-specialists in soil research, in soil data recovery and quality assessment exercises.

Visualization

There are two ways to search for data in **febr**. The first is via the global visualization page, www.ufsm.br/febr/view, which shows the distribution of soil observations across the Brazilian territory – provided they have spatial coordinates (**Figure 1**). Various additional geographical data layers can be accessed in the global visualization page to assist navigation and geolocation (terrain, street and road network, vegetation). These data comes from different providers such as Esri, OpenStreetMap, OpenWeatherMap, and Stamen, through the JavaScript library Leaflet accessed through the R-packages leaflet and mapview. By clicking on a point, a popup window appears, showing a link to the page of the dataset in the **febr** catalog.

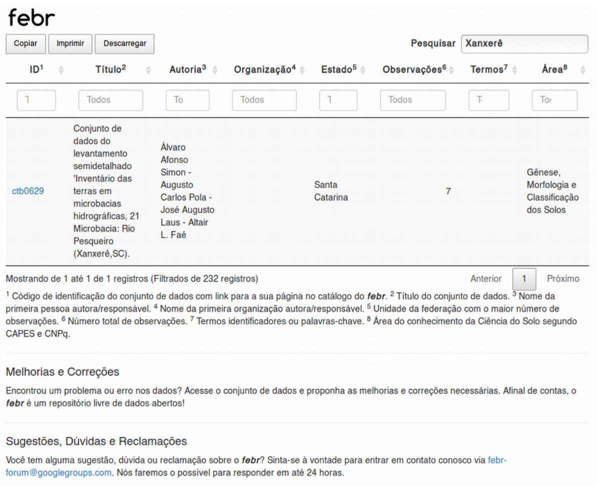


x: -5841012 | y: -3101509 | epsg: 3857 | proj4: +proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +y_0=0 +k=1.0 +units=m +nadgrids=@null +no_defs | lon: -52.47070 | lat: -26.82407 | zoom: 5

Figure 1. Snapshot of the visualization page of **febr**. The popup window on a point in Xanxerê (SC) has a link to its page in the **febr** catalog.

Search

The second search tool available in **febr** is the dedicated search page, www.ufsm.br/febr/search (**Figure 2**). It was implemented using the JavaScript library DataTables – a plugin for the JavaScript library jQuery – through the R-package DT. In addition to finding a dataset, the search page helps discovering other data sets and learning how they are related to each other. For this purpose, seven search criteria are used: dataset title, dataset first author name, dataset first author organization, federative unit (state) with the largest number of observations, total number of observations, indexing terms and knowledge area of Soil Science according to CAPES and CNPq. Clicking on the identification code of a dataset opens its page in the **febr** catalog.



febr

Copiar | Imprimir | Descarregar

Pesquisar: Xanxerê

ID[1]	Título[2]	Autoria[3]	Organização[4]	Estado[5]	Observações[6]	Termos[7]	Área[8]
1	Todos	To	Todos	1	Todos	T	To
ctb0629	Conjunto de dados do levantamento semidetalhado 'Inventário das terras em microbacias hidrográficas, 21 Microbacia: Rio Pesqueiro (Xanxerê,SC).	Álvaro Afonso Simon - Augusto Carlos Pola - José Augusto Laus - Altair L. Faé		Santa Catarina	7		Gênese, Morfologia e Classificação dos Solos

Mostrando de 1 até 1 de 1 registros (Filtrados de 232 registros)

Anterior | 1 | Próximo

[1] Código de identificação do conjunto de dados com link para a sua página no catálogo do ***febr***. [2] Título do conjunto de dados. [3] Nome da primeira pessoa autora/responsável. [4] Nome da primeira organização autora/responsável. [5] Unidade da federação com o maior número de observações. [6] Número total de observações. [7] Termos identificadores ou palavras-chave. [8] Área do conhecimento da Ciência do Solo segundo CAPES e CNPq.

Melhorias e Correções

Encontrou um problema ou erro nos dados? Acesse o conjunto de dados e proponha as melhorias e correções necessárias. Afinal de contas, o ***febr*** é um repositório livre de dados abertos!

Sugestões, Dúvidas e Reclamações

Você tem alguma sugestão, dúvida ou reclamação sobre o ***febr***? Sinta-se à vontade para entrar em contato conosco via febr-forum@googlegroups.com. Nós faremos o possível para responder em até 24 horas.

Figure 2. Snapshot of the search page of **febr**. The search term "Xanxerê" returned a single result (dataset_id = 'ctb0629').

Catalog

Each dataset has an individual page in the **febr** catalog, www.ufsm.br/febr/catalog/. Implemented using the R-package bookdown, the catalog details the key features of each dataset. This includes a general description, especially its origin and how it was produced, and the people and institutions responsible for its production. Secondary information and the spatial distribution of observations – provided they have spatial coordinates – are also available. Unlike the global visualization page, the local catalog spatial visualization tool can be used to evaluate the quality of geospatial data in more detail. The catalog also includes a search tool via the jQuery JavaScript library. Finally, the catalog gives access to dataset directories and respective spreadsheets in Google Drive, where they can be edited (upon request) or downloaded from.

Download

Datasets in **febr** can be downloaded in two different ways. The first is to access and download the spreadsheets directly from Google Drive, where the output file format can be selected in the *File* menu, with the options XLSX, ODS, PDF, HTML, CSV, and TSV. The second way is to use the R-package **febr** (**Figure 3**). It has four key functions, dataset, observation, layer, and metadata, each of them designed to download the tables 'dataset', 'observacao', 'camada', and 'metadado', respectively. Various arguments can be passed to these functions to select the datasets and variables that should be downloaded. The connection with Google Sheets is established using the R-package googlesheets.

```
# Install packages
if (!require(devtools)) {
  install.packages(pkgs = "devtools")
}
devtools::install_github(
  repo = "febr-team/febr-package")

# Download observations with all variables
obs <- febr::observation(
  dataset = "ctb0629", variable = "all")

# Download layers with all variables
lrs <- febr::layer(
  dataset = "ctb0629", variable = "all")

# Merge data frames
ctb0629 <- merge(
  x = obs, y = lrs,
  by = c("dataset_id", "observacao_id"))
```

Figure 3. Installation and usage of the R-package **febr**. The code chunk shows how to download and merge data from dataset 'ctb0629'.

Routines for data standardization and harmonization have been implemented in the R-package **febr**. The former includes dealing with measurement units and number of decimal places, irregular transitions between sampling layers, symbols used to indicate detection limits (+, <, >) or missing data (N/A, NA, *blank*), coordinate reference systems, and so on. The later consists of rules to translate soil attributes determined using disparate analytical methods into a common attribute space. For example, to harmonize the iron contents in the sulfuric acid extract determined by atomic absorption spectrometry and by inductively coupled plasma spectrometry. The benefit of these routines is that upon download the data is virtually ready for analyses and modelling.

Discussion group

The **febr** is a project with challenging goals. This requires much discussion and the collaboration of countless people. A public discussion group was created to facilitate these. One can send a message to febr-forum@googlegroups.com to join the discussion group and then influence the definition of standards and data management choices, collaborate in research and development activities, propose new features and improvements, help solving questions and provide technical support, and point out data inconsistencies. Also, by posting to febr-forum@googlegroups.com one can learn how to publish data in **febr.**

Documentation

A comprehensive documentation, prepared using the R-package bookdown, is available at www.ufsm.br/febr/book/. Constantly updated, the documentation is fundamental to facilitate the adoption of agreed standards by those interested in publishing data in **febr** as well as by the **febr** maintainers. This guarantees, for example, that **febr** maintainers can be replaced without harming the progress of the activities.

OPPORTUNITIES

The **febr** currently has 14,477 soil observations from 232 datasets covering the Brazilian territory (**Figures 1 and 4**). **Figure 4** in special shows that many states have a poor coverage in **febr**, mainly from the North and Centre-West regions of Brazil, where soil data sharing should be further encouraged. The potential benefits of having this enormous volume of soil data freely available, standardized and harmonized, for short and long term scientific reuse, are countless. For example, improving the Brazilian Soil Classification System and international ones (Universal, WRB, Soil Taxonomy), creating intelligent fertilizer recommendation engines, developing specialized databases, calibrating pedotransfer functions, supporting the upcoming Brazilian National Soil Survey Program (PronaSolos) and international soil mapping initiatives (GlobalSoilMap, Global Soil Partnership), and many others. In the long term, the project results should promote a cultural change towards a more open and collaborative soil science in Brazil. By sharing data through a centralized, collaborative and community-based soil data storing and sharing facility, soil scientists from different fields have the opportunity to increase collaboration and the much needed soil knowledge.

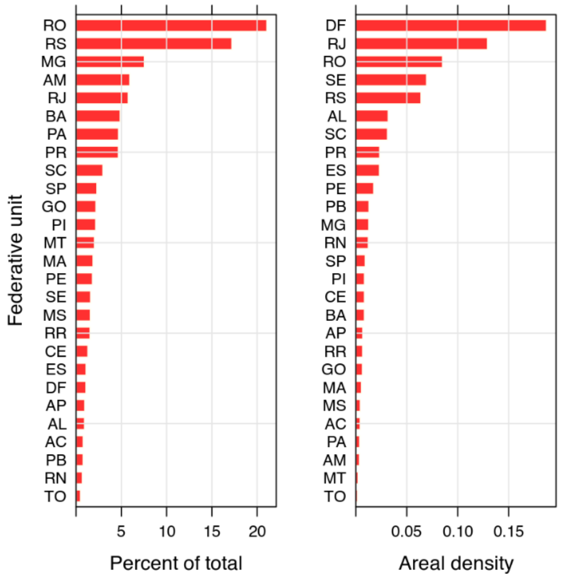


Figure 4. Relative distribution and areal density (per 1000 km^2) of observations among federative units (states) – 42% of observations are from the South and Southeast regions.

ACKNOWLEDGEMENTS: The authors thank all soil scientists who have collaborated in the development of **febr** by sharing data and ideas.

REFERENCES

Arrouays D et al. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ. 2017;14:1-19. doi:10.1016/j.grj.2017.06.001

Eaves D. Three Laws of Open Data. Nov/2009. Accessed in: 29 Nov. 2009. Available at: https://eaves.ca/2009/11/29/three-laws-of-open-data-international-edition/.