Repositório Brasileiro Livre para Dados Abertos do Solo

Manual de Uso

Alessandro Samuel-Rosa

27-12-2017

# Conjunto de dados

A maximização da disseminação e uso de um conjunto de dados, o apropriado reconhecimento das pessoas e organizações autoras e/ou responsáveis, bem como a identificação de alterações e ajustes feitos após a publicação da sua primeira versão, dependem da provisão de dados mínimos sobre aquele conjunto de dados. No **febr**, esses dados sobre o conjunto de dados são inseridos na tabela denominada *dataset*, palavra que na língua inglesa significa ‘conjunto de dados’.

Por conter dados essenciais sobre o conjunto de dados, a tabela *dataset* é sempre a primeira a ser revisada e processada. Quaisquer dados faltantes são solicitados aos responsáveis pelo conjunto de dados. Da mesma forma, sempre que dados incoerentes são identificados, os responsáveis pelo conjunto de dados são consultados iterativamente até que quaisquer dúvidas sejam completamente sanadas. O devido preenchimento da tabela *dataset* é essencial para o posterior processamento dos dados e metadados do conjunto de dados.

A tabela *dataset* possui cerca de vinte campos a serem preenchidos, os quais são estruturados em uma sequência de linhas (Tabela 1). Especificamente, os dados sobre o conjunto de dados são inseridos na segunda coluna, imediatamente ao lado da sua respectiva identificação e definição contidas na primeira coluna. A primeira linha é dedicada à identificação propriamente dita do conjunto de dados no **febr**, a segunda ao título do conjunto de dados, e assim por diante, até as últimas linhas, dedicadas aos termos usados para a descrição de características chave e indexação do conjunto de dados. Cada um dos campos da tabela *dataset* são descritos em detalhe nas próximas seções.

Tabela 1 Representação da estrutura da tabela *dataset*.

|  |  |
| --- | --- |
| **campo** | **valor** |
| dataset\_id | <???> |
| dataset\_titulo | <???> |
| dataset\_descricao | <???> |
| … | <???> |
| palavras\_chave | <???> |
| categoria\_vcge | <???> |

## Identificação

Um conjunto de dados no **febr** é identificado por dois campos principais (Tabela 2). O primeiro deles, dataset\_id, é o código identificador único do conjunto de dados, o que significa que não há qualquer dataset\_id repetido no **febr**.

Tabela 2 Campos utilizados para a identificação do conjunto de dados.

|  |  |
| --- | --- |
| **campo** | **valor** |
| dataset\_id | código de identificação do conjunto de dados |
| dataset\_titulo | título do conjunto de dados |

Em geral, o código de identificação de um novo conjunto de dados inserido no **febr** é definido pela sua posição em relação aos demais conjuntos de dados (**c**on**t**ri**b**uições, ctb) presentes no **febr**. Por exemplo, o código de identificação do primeiro conjunto de dados inserido no **febr** é dataset\_id = ctb0001, enquanto que do trigésimo terceiro conjunto de dados é dataset\_id = ctb0033.

No caso dos conjuntos de dados obtidos diretamente do Sistema de Informação de Solos Brasileiros ([SISB](https://www.sisolos.cnptia.embrapa.br/)), o código de identificação utilizado no **febr** é o mesmo usado naquele sistema, independente da sua ordem de entrada no **febr**. Por exemplo, dataset\_id = ctb0770, onde 770 é o código de identificação usado no SISB. Essa estratégia possibilita que, no futuro, e em havendo interesse por parte dos mantenedores do SISB, seja estabelecida comunicação direta do SISB com o **febr**. Assim, seria possível atualizar os dados no SISB à medida que ajustes e correções nos dados forem sendo feitos no **febr**.

O segundo identificador de um conjunto de dados é o seu título – dataset\_titulo –, geralmente relacionado ao título do projeto, tese, dissertação, etc que o gerou. Em geral, é mais interessante usar um título específico para o conjunto de dados, o que identifica melhor o seu conteúdo, ao invés do título do trabalho ou projeto que o gerou. Isso é importante para aumentar as chances de o conjunto de dados ser identificado em ferramentas de busca e assim potencializar o seu reuso.

Quanto à formatação do título, é usada a mesma língua da fonte e caracteres em caixa baixa, com exceção da primeira letra do título e dos nomes próprios. Por exemplo, *Conjunto de dados da tese ‘Matéria orgânica e características físicas, químicas, mineralógicas e espectrais de Latossolos de diferentes ambientes’*.

## Descrição

A descrição do conjunto de dados, contendo dados básicos para maximizar o seu reuso futuro, sem que seja necessário entrar em contato com os autores ou responsáveis, é incluída no campo dataset\_descricao. A descrição do conjunto de dados aborda aspectos como:

1. Os motivos para a realização do estudo que levou à geração dos dados,
2. Um resumo dos dados incluídos no conjunto de dados e dos métodos analíticos usados, e
3. Uma descrição do delineamento amostral e a forma de coleta das amostras.

A descrição do delineamento amostral, ou seja, a estratégia utilizada para seleção dos locais de observação e amostragem do solo, fornece os elementos necessários para determinar se – e como – os diferentes conjuntos de dados podem – ou devem – ser utilizados em uma determinada aplicação.

Quando for o caso, a descrição do conjunto de dados também inclui:

1. Os motivos para a presença de camadas/horizontes com dados faltantes para uma ou mais variáveis do solo, e
2. Uma descrição das alterações/modificações realizadas quando da publicação de uma nova versão.

Uma descrição exemplar de um conjunto de dados é a que segue:

“Conjunto de dados produzido como parte da Dissertação submetida como requisito parcial para obtenção do grau de Mestre no Curso de Pós-Graduação em Agronomia da Universidade Federal Rural do Rio de Janeiro. Inclui dados do conteúdo de ferro total determinado via extração com solução de ácido sulfúrico para 20 perfis do solo observados no município de Pinheiral, Rio de Janeiro. A seleção dos locais de observação foi feita com base no conhecimento pedogenético da área de estudo, a partir de informações prévias sobre o meio físico e solo, tendo sido selecionados pontos representativos que contemplassem a variação dos fatores de formação do solo. Em cada local foram abertas trincheiras para descrição dos perfis e coleta de amostras dos horizontes para caracterização do solo segundo procedimentos descritos no Manual de Descrição e Coleta de Solo no Campo (Santos et al., 2013). Devido à existência de limitações orçamentárias, apenas amostras de horizontes selecionados foram submetidas à determinação do conteúdo de ferro total. Em alguns casos, apenas os horizontes B diagnósticos foram analisados, em outros o primeiro horizonte A e um ou mais horizontes subsuperficiais.”

Arrouays, D. et al. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. *GeoResJ*, 14:1–19, 2017. doi:[10.1016/j.grj.2017.06.001](https://doi.org/10.1016/j.grj.2017.06.001).

Batjes, N. H., Ribeiro, E., Oostrum, A. van, Leenaars, J., Hengl, T., Jesus, J. M. de. WoSIS: providing standardised soil profile data for the world. *Earth System Science Data*, 9(1):1–14, 2017. doi:[10.5194/essd-9-1-2017](https://doi.org/10.5194/essd-9-1-2017).

Chagas, C. S., Junior, W. C., Bhering, S. B., Tanaka, A. K., Baca, J. F. M. Organization and structure of the Brazilian soil information system (SigSolos – version 1.0). *Revista Brasileira de Ciência do Solo*, 28(5):865–876, 2004. doi:[10.1590/S0100-06832004000500009](https://doi.org/10.1590/S0100-06832004000500009).

Cooper, M., Mendes, L. M. S., Silva, W. L. C., Sparovek, G. A national soil profile database for Brazil available to international scientists. *Soil Science Society of America Journal*, 69(3):649–652, 2005–5DC. doi:[10.2136/sssaj2004.0140](https://doi.org/10.2136/sssaj2004.0140).

CrowdFlower. *Data Scientist Report 2017*. San Francisco: CrowdFlower, 2017. <https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html>.

Es, H. van. A New Definition of Soil. *CSA News*, 62(10):20, 2017. doi:[10.2134/csa2017.62.1016](https://doi.org/10.2134/csa2017.62.1016).

Hanson, B., Sugden, A., Alberts, B. Making data maximally available. *Science*, 331(6018):649–649, 2011. doi:[10.1126/science.1203354](https://doi.org/10.1126/science.1203354).

Hengl, T. et al. SoilGrids1km–global soil information based on automated mapping. *PLoS ONE*, 9(8):e105992, 2014. doi:[10.1371/journal.pone.0105992](https://doi.org/10.1371/journal.pone.0105992).

Ottoni, M. V., Lopes-Assad, M. L. R. C., Pachepsky, Y., Rotunno Filho, O. C. A hydrophysical database to develop pedotransfer functions for Brazilian soils: challenges and perspectives. In: TEIXEIRA, W. G.; CEDDIA, M. B.; OTTONI, M. V.; DONNAGEMA, G. K. (Eds.).. *Application of Soil Physics in Environmental Analyses*. Basel: Springer International Publishing, 2014. p. 467–494. doi:[10.1007/978-3-319-06013-2\_20](https://doi.org/10.1007/978-3-319-06013-2_20).

Sestoft, P. Organizing research data. *Acta Veterinaria Scandinavica*, 53(Suppl 1):S2, 2011. doi:[10.1186/1751-0147-53-S1-S2](https://doi.org/10.1186/1751-0147-53-S1-S2).

Simões, M. G., Oliveira, S. R. M., Ferraz, R. P. D., Santos, H. G., Manzatto, C. V. Democratization of information on Brazilian soils: Geoportal and soil database accessible via web. *Cadernos de Ciência e Tecnologia*, 32(1/2):55–69, 2015. <http://seer.sct.embrapa.br/index.php/cct/article/view/23295>.

Vines, T. H. et al. The availability of research data declines rapidly with article age. *Current Biology*, 24(1):94–97, 2014. doi:[10.1016/j.cub.2013.11.014](https://doi.org/10.1016/j.cub.2013.11.014).

Wickham, H. Tidy Data. *Journal of Statistical Software*, 59(10):1–23, 2014. doi:[10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10).