Repositório Brasileiro de Dados de Ferro do Solo

ALESSANDRO SAMUEL-ROSA
Universidade Federal de Santa Maria

2017-07-25

Contents


```
knitr::opts_chunk$set(dev.args = list(bg = 'transparent'))
```

Prefácio

Chapter 1

Divulgação e Sensibilização

Chapter 2

Organização dos Dados

Os dados são organizados em tabelas bidimensionais usando planilhas eletrônicas padronizadas construídas a partir de experiências de iniciativas tanto locais como globais. São quatro as tabelas utilizadas, como nomes: *dataset*, *observacao*, *camada* e *metadado*. O propósito e conteúdo de cada uma dessas tabelas são descritos nas sessões que seguem.

2.1 Tabela *dataset*

A maximização da disseminação e uso de um conjunto de dados, o apropriado reconhecimento de seu autor e instituição responsável, bem como a identificação que alterações e ajustes feitos após a publicação da sua primeira versão, dependem da provisão de dados mínimos sobre aquele conjunto de dados. No Fe-BR esses dados são inseridos na tabela denominada *dataset*.

Por conter dados essenciais sobre um conjunto de dados, a tabela *dataset* é a primeira a ser revisada e processada. Quaisquer informações faltantes são solicitadas ao autor/responsável pelo conjunto de dados. Da mesma forma, sempre que informações incoerentes são identificadas, o autor/responsável pelo conjunto de dados é consultado iterativamente até que quaisquer dúvidas sejam completamente sanadas.

A tabela *dataset* possui cerca de vinte itens a ser preenchidos estruturados em uma sequência de linhas. Nessa tabela os dados são inseridos na segunda coluna, imediatamente ao lado da sua respectiva identificação. Assim, a primeira linha é dedicada à identificação propriamente dita do conjunto de dados no repositório, a segunda ao título do conjunto de dados, e assim por diante, até as últimas linhas, dedicadas a termos usados descrição características chave e indexação do conjunto de dados.

item	data
dataset_id	...
dataset_titulo	...
...	...
palavras_chave	...

Os itens da tabela *dataset* são descritos em detalhe a seguir.

2.1.1 `dataset_id`

Código identificador único do conjunto de dados. Esse código é usado como identificador dos diferentes conjuntos de dados no Fe-BR. Em geral, o código de identificação de cada conjunto de dados é definido

pela equipe do Fe-BR de acordo com a ordem das contribuições. Por exemplo, o código de identificação da primeira contribuição é `dataset_id = ctb0001`. No caso dos conjuntos de dados obtidos diretamente do SISB, o código de identificação utilizado no Fe-BR é o mesmo usado naquele sistema, garantindo assim a comunicação direta entre o SISB e Fe-BR.

2.1.2 `dataset_titulo`

Título do conjunto de dados, geralmente relacionado ao projeto, tese, dissertação, etc que gerou o conjunto de dados. Preferência é dada ao uso de um título específico para o conjunto de dados, o que identifica melhor o seu conteúdo do que o título do trabalho ou projeto que gerou. Isso é importante para aumentar as chances de o conjunto de dados ser identificado em ferramentas de busca. Quanto ao formato, usa-se a mesma língua da fonte – apesar de títulos em língua portuguesa serem preferidos – 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'.*

2.1.3 `dataset_descricao`

Descrição do conjunto de dados, cotendo dados básicos para maximizar o reuso futuro do conjunto de dados sem que seja necessário entrar em contato com o seu autor/responsável. A descrição inclui aspectos como:

1. Os motivos para a realização do estudo que levou à produção 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 – um determinado conjunto de dados pode ser usado juntamente com outros conjuntos de dados cujas observações do solo tenham sido feitas de modo dissimilar. Quando for o caso, a descrição do conjunto de dados deve incluir ainda:

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 (RJ). 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 submetidos 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.

2.1.4 dataset_versao

Versão do conjunto de dados. Dado necessário nos casos em que o conjunto de dados foi alterado/modificado por uma razão técnica e/ou científica. Por exemplo, um conjunto de dados contendo inúmeros perfis do solo cujas localizações originais não foram georreferenciadas – versão 1 – é modificado de forma a possuir coordenadas espaciais aproximadas dos locais de observação de todos os perfis do solo – versão 2. Sempre que uma nova versão de um conjunto de dados é preparada, uma descrição das alterações/modificações deve ser apresentada junto da descrição do conjunto de dados em `dataset_descricao`.

2.1.5 dataset_licenca

Licença de uso do conjunto de dados. Dado que define com o conjunto de dados pode ser usado a partir da sua publicação no Fe-BR. A legislação brasileira ainda não especifica, exatamente, qual deve ser a licença de uso dos conjuntos de dados produzidos via financiamento público. Contudo, existe consenso de que a licença de uso mais apropriada para esse tipo de conjuntos de dados é a licença Creative Commons CC BY 4.0. A licença CC BY 4.0 permite que um conjunto de dados seja distribuido, remixado, adaptado e usado para criar outros produtos, mesmo que para fins comerciais, desde que seja atribuido o devido crédito ao autor/responsável pelo conjunto de dados original. Assim, trata-se da licença mais flexível dentre as licenças Creative Commons disponíveis, maximizando a disseminação e uso dos conjuntos de dados.

Para o caso de conjuntos de dados produzidos via financiamento privado, onde as partes envolvidas têm interesse em compartilhar os dados desde que não para fins comerciais, usa-se a licença Creative Commons CC BY-NC. A licença CC BY-NC possui exatamente os mesmos termos da licança CC BY, exceto pelo fato de que a distribuição, remixação, adaptação e derivação de outros produtos não sejam usados para fins comerciais.

2.1.6 publicacao_data

Data de publicação do conjunto de dados. A data de publicação do conjunto de dados não é necessariamente a mesma data de publicação do trabalho que produziu ou utilizou o conjunto de dados pela primeira vez, mas sim a data em que o conjunto de dados foi tornado efetivamente público e disponível. Quando ajustes e/ou correções são feitas, então a data de publicação é a data em que o conjunto de dados foi tornado disponível no Fe-BR.

A data de publicação é apresentada usando o formato numérico padrão dd-mm-aaaa (dia-mês-ano). Preferência é dada à inserção dos dados de ambos dia, mês e ano. Contudo, dentre os três dados, o mais importante é aquele relativo ao ano. Quando, por exemplo, tanto o dia como o mês de observação são desconhecidos, usa-se `xx-xx`. Caso a data de publicação seja completamente desconhecida, usa-se a data de publicação do trabalho que gerou ou utilizou o conjunto de dados pela primeira vez.

2.1.7 organizacao

Identificação e endereços físico e eletrônico da organização responsável pela geração do conjunto de dados. Quando o conjunto de dados é provido por organização outra que não aquela responsável pela sua geração, identifica-se a organização que atualmente detém responsabilidade técnica sobre o conjunto de dados, ou seja, a organização provedora do conjunto de dados. O mesmo aplica-se aos conjuntos de dados produzidos por duas ou mais organizações. No caso de conjuntos de dados produzidos ou sob responsabilidade de organizações como a Embrapa, que possuem várias unidades distribuídas pelo território brasileiro, identifica-se a unidade da organização responsável pelo conjunto de dados, nunca apenas a organização. Em todo caso, é fundamental que ambos endereço físico e eletrônico sejam válidos e atuais.

Os dados necessários para a indentificação da organização responsável pelo conjunto de dados são os seguintes:

- organizacao_nome
- organizacao_url
- organizacao_pais_id
- organizacao_municipio_id
- organizacao_codigo_postal
- organizacao_rua_nome
- organizacao_rua_numero

2.1.8 autor

autor_nome; autor_email

Identificação do autor – ou autora ou autores –, ou seja, a pessoa responsável pela geração do conjunto de dados. Quando o acesso ao autor é impossível ou limitado, identifica-se a pessoa que atualmente detém responsabilidade técnica sobre o conjunto de dados. Contudo, preferência é dada à identificação do autor do conjunto de dados. No caso de trabalhos acadêmicos, como monografias, dissertações e teses, identifica-se o autor e o orientador principal do trabalho acadêmico. Nesse caso, e sempre que houver dois ou mais autores/responsáveis, os respectivos dados de identificação são separados usando ponto e vírgula. É fundamental que o endereço de e-mail do autor/responsável seja válido e atual.

2.1.9 contribuidor

contribuidor_nome, contribuidor_email, contribuidor_organizacao

2.1.10 dataset_referencia_i

Referência permanente válida a documentos ou artigos científicos onde a versão atual do conjunto de dados foi usado pela primeira vez. Tantas referências quantas forem julgadas pertinentes podem ser inseridas, numerando-as sequencialmente em ordem de importância, usando para isso o índice i. Preferência é dada ao uso de um Digital Object Identifier (DOI), mas uma URL também é aceita. No caso de o responsável pelo conjunto de dados não ser o seu autor, é imprescindível que a referência principal, i = 1, seja a um trabalho do autor do conjunto de dados. Já a referência secundária, i = 2, pode ser a um trabalho do responsável pelo conjunto de dados. No caso de trabalhos acadêmicos, como monografias, disserações e teses, informa-se uma referência ao trabalho acadêmico e ao artigo onde o conjunto de dados foi utilizado pela primeira vez. No caso de o conjunto de dados ter passado por alterações/modificações, uma referência ao trabalho onde a nova versão do conjunto de dados tenha sido usada pela primeira vez também é inserida. Para inserir novas referências, basta inserir, em sequência, novas linhas na tabela.

2.1.11 area_conhecimento

Área de especialidade da Agronomia – Ciência do Solo à qual o conjunto de dados está relacionado. Segundo o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), são seis as áreas de especialidade:

- Gênese, Morfologia e Classificação dos Solos
- Física do Solo
- Química do Solo
- Microbiologia e Bioquímica do Solo*
- Fertilidade do Solo e Adubação
- Manejo e Conservação do Solo

2.1.12 palavras_chave

Lista de termos que descrevem aspectos importantes do conjunto de dados, preferencialmente diferentes daqueles constantes no título e compostos por até três palavras, separados por ponto e vírgula. Uma lista de termos bem elaborada aumenta o potencial de descoberta do conjunto de dados por mecanismos de busca.

2.1.13 categoria_vcge

2.2 Tabela observacao

Nesta aba são inseridas as informações espaciais mais fundamentais sobre as observações do solo.

2.2.1 observation_id

Identificador exclusivo da observação usado no conjunto de dados de origem. Como se trata de um código, não devem ser usados espaços ou caracteres especiais. Espaços devem ser substituídos por um *underscore*, `_`, ou traço, `-`. Caracteres especiais devem ser substituídos pelo caractere correspondente simplificado, ou seja, `á` torna-se `a`, `ç` torna-se `c`, e assim por diante. Exemplo: `observation_id = Perfil-01`.

2.2.2 observation_date

Data da observação no formato dd-mm-aaaa. A data de observação é um dos atributos que, juntamente com as coordenadas espaciais, definem uma observação em um conjunto de dados. Isso significa que a observação repetida, ao longo do tempo, de um mesmo local no espaço, caracteriza a constituição de uma nova observação, portanto, a definição de uma nova `observation_id`. O valor padrão, para os casos em que a data de observação é desconhecida, é `observation_date = xx-xx-xxxx`. Quando somente o dia e/ou o mês de observação são desconhecidos, usa-se o primeiro valor da sequência, ou seja, `dd = 01` e `mm = 01`.

2.2.3 coord_sistema

Sistema de coordenadas de referência (SCR) utilizado para o georreferenciamento das observações do solo. A especificação do SCR é fundamental para possibilitar o uso apropriado de dados espaciais, especialmente para fins de correlação/cruzamento com outros dados espaciais. Como a especificação do SCR pode ser feita de diversas maneiras, são aceitas quaiquer descrições mais populares, não padronizadas, como WGS 84 / UTM zone 23S. Contudo, para fins de organização dos dados e posterior automatização de processos computacionais, o Fe-BR adota os códigos padronizados e aceitos internacionalmente do European Petroleum Survey Group (EPSG). Assim, a partir da descrição fornecida com o conjunto de dados, identifica-se o código EPSG correspondente, que é usado em substituição daquela.

2.2.4 coord_x; coord_y

Coordenada X e Y, ou seja, o mesmo que Longitude e Latitude, desde que em coordenadas geográficas.

2.2.5 coord_accuracy; coord_source

Erro das coordenadas geográficas em metros.

coord_source	Definição
GPS	Aparelho GPS
MAPA	Mapa analógico ou digital
WEB	Serviço web como o Google Maps

2.2.6 country_id; state_id; city_name

Identificação do município, estado – ou unidade federativa – e país onde a observação foi realizada. Como o Fe-BR trata apenas de conjuntos de dados produzidos no Brasil, usa-se o código ISO 3166-1 alpha-2 do Brasil, ou seja, BR. No caso do estado – ou unidade federativa –, usa-se a abreviação da respectiva unidade federativa (UF). O nome do município é escrito por extenso, conforme encontrado na fonte.

2.2.7 sample_type; sample_number; sample_area

Tipo de amostragem. Opções: SIMPLES ou COMPOSTA.Número de amostras. Um (1) quando o valor da variável anterior é SIMPLES ou mais para COMPOSTA. Expressar o valor com números inteiros. Área amostral (m^2). Por exemplo, a área de uma trincheira costuma ser de 1 m^2.

2.2.8 taxon_<sistema>_<ano>

Identificação do sistema de classificação – taxonomia – utilizado para classificar a observação do solo, incluindo nome e ano, bem a classificação – táxon – atribuída à observação, incluíndo nome e sigla. Tanto o nome como o ano de publicação do sistema de classificação utilizado auxilia na verificação de possíveis inconsistências na classificação da observação e no processo de harmonização dos dados. O nome da classificação atribuída à observação do solo deve ser escrito por extenso, mantendo quaisquer espaços e caracteres especiais presentes, prefere

Código	Sistema taxonômico
fao-unesco	Legend of the World Soil Map
fao-wrb	World Reference Base for Soil Resources
sibcs	Sistema Brasileiro de Classificação do Solo
st	Soil Taxonomy
usc	Universal Soil Classification

Para os conjuntos de dados que incluem observações do solo classificadas usando dois ou mais sistemas de classificação, ou diferentes versões do mesmo sistema de classificação, usa-se o índice i presente no nome de cada uma das quatro colunas relativas à classificação do solo para identificar os respectivos sistemas de classificação. Assim, colunas adicionais são inseridas para cada sistema de classificação, onde i = 1 indica o primeiro sistema de classificação, i = 2 o segundo sistema de classificação, e assim por diante. Caso apenas um sistema de classificação tenha sido usado, usa-se apenas i = 1.

2.3 Tabela *camada*

Nesta aba são inseridos os dados das camadas amostradas e onde o conteúdo de ferro tenha sido determinado.

2.3.1 observation_id

Identificador exclusivo da observação usado no conjunto de dados de origem. Refere-se à tabela *observation*.

2.3.2 layer_number

Número da camada ou horizonte na observação, atribuído consecutivamente de cima para baixo. Há alguma confusão com essa variável, por vezes sendo atribuído valor numérico sequencial para todo o conjunto de

dados. É preciso ficar claro que a numeração refere-se à ordem das camadas dentro do perfil à que pertencem, ou seja, a primeira camada amostrada de um perfil receberá, sempre, o número inteiro 1. O valor padrão, para o caso do perfil observado possuir apenas uma camada, é `layer_number = 1`.

2.3.3 `layer_name`

Nome da camada ou horizonte. Exemplo: A, B, C etc. Conjuntos de dados que são produto de trabalhos edafológicos raramente terão informações sobre o nome da camada amostrada. Caso essa informação esteja presente, deve-se verificar com o responsável pelo conjunto de dados se a informação está correta ou foi criada apenas para preencher a coluna da planilha eletrônica. O valor padrão, para o caso do perfil observado ser oriundo de estudo edafológico, é `layer_name = NA`.

2.3.4 `sample_code`

Código laboratorial da amostra. Usado para identificar as repetições de laboratório. Alguns grupos de pesquisa destinam suas amostras para análise em laboratórios especializados, os quais identificam as amostras com códigos específicos, que chamamos de `sample_code`. Em outros casos, as análises são feitas pelo laboratório do próprio grupo de pesquisa, o qual não necessariamente atribui um código de identificação único às amostras. Mas esses laboratórios podem realizar as análises com duas ou mais repetições, sendo então o número da repetição de uma determinada amostra identificado como `sample_code`. O valor padrão, para o caso em que não há código de identificação único e/ou a amostra não foi analisada com repetições, é `sample_code = 1`.

2.3.5 `upper_depth`; `lower_depth`

Profundidade do limite superior e inferior, respectivamente, da camada ou horizonte observados.

2.3.6 `fe_<extração>_<determinação>`

O valor padrão, para o caso em que o conteúdo de ferro não está disponível, é `fe_<extração>_<determinação> = NA`.

Chapter 3

Dificuldades e Desafios

Licença de uso. Item com o qual tem havido mais confusão até agora, sobretudo porque a maioria das pessoas não está familiarizada com a atribuição de licença de uso aos dados que produzem. Na maioria dos casos será necessário entrar em contato com o responsável pelo conjunto de dados para explicar o que cada licença significa.