Procedimentos Adotados no Repositório Brasileiro de Dados de Ferro do Solo

ALESSANDRO SAMUEL-ROSA
Programa de Pós-Graduação em Ciência do Solo
Universidade Federal de Santa Maria

2017-08-28

Conteúdo

Apresentação

Nós acreditamos que o uso sustentável do solo para aplicações ambientais no Brasil exige o conhecimento da distribuição horizontal e vertical do conteúdo de ferro no solo. Por quê? Porque, em alguns locais, os óxidos de ferro podem perfazer até 80% da massa do solo!

Os óxidos de ferro são bem conhecidos por sua forte interação com a matéria orgânica do solo, influenciando assim a quantidade de carbono que o solo consegue armazenar de forma estável. Os óxidos de ferro também são bem conhecidos pela sua forte afinidade com íons fosfato, determinando assim a disponibilidade de fósforo para as plantas. A intrincada relação entre fósforo e matéria orgânica reforça a necessidade do conhecimento da variação tridimensional do teor de ferro do solo em todo o território brasileiro. Tal conhecimento poderia ajudar, por exemplo, na melhoria dos sistemas de classificação do solo, no desenvolvimento de sistemas avançados de recomendação que assegurem o uso mais eficiente de fertilizantes fosfatados, e na construção de políticas públicas de uso e ocupação do solo que respeitem as reais capacidades desse precioso e não-renovável bem natural.

Foi pensando na necessidade de impulsionar o avanço do conhecimento para o uso sustentável do solo brasileiro que decidimos desenvolver um projeto inovador. O objetivo: a construção colaborativa de um repositório de dados de ferro do solo com cobertura nacional que fosse público, gratuito e, sobretudo, fácil de usar e manter. Lançado em dezembro de 2016, o Repositório Brasileiro de Dados de Ferro do Solo (Fe-BR) já conta com a impressionante soma de 232 conjuntos de dados oriundos de todas as vinte e sete unidades federativas do Brasil. Muitos dos conjuntos de dados já foram processados e estão aqui, à sua inteira disposição, esperando para serem usados. Contribua você também!

Capítulo 1

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, com nomes:

1. *dataset*, para os dados sobre o conjunto de dados,
2. *observacao*, para os dados das observações do solo,
3. *camada*, para os dados das camadas e horizontes, e
4. *metadado*, para os dados sobre os dados.

Uma descrição detalhada do propósito e conteúdo específico de cada uma dessas tabelas é apresentado nas sessões que seguem.

1.1 Tabela *dataset*

A maximização da disseminação e uso de um conjunto de dados, o apropriado reconhecimento de seus autores ou autoras 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 sempre a primeira a ser revisada e processada. Quaisquer dados faltantes são solicitados aos autores/responsáveis pelo conjunto de dados. Da mesma forma, sempre que dados incoerentes são identificados, os autores/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 de um conjunto de dados de ferro do solo.

A tabela *dataset* possui cerca de vinte itens estruturados em uma sequência de linhas. 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 Fe-BR, 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.

item	dado
dataset_id	...
dataset_titulo	...
dataset_descricao	...
...	...
palavras_chave	...

item	dado
categoria_vcge	...

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

1.1.1 `dataset_id`

Código identificador único do conjunto de dados. Esse código é usado para identificar cada uma dos conjuntos de dados incluídos no Fe-BR. Em geral, o código de identificação de um novo conjunto de dados inserido no Fe-BR é definido pela sua posição (ordem das contribuições) em relação aos demais conjunto de dados já inseridos no Fe-BR. Por exemplo, o código de identificação do primeiro conjunto de dados contribuído (`ctb`) para o Fe-BR é `dataset_id = ctb0001`.

No caso dos conjuntos de dados obtidos diretamente do Sistema de Informação de Solos Brasileiros (SISB), o código de identificação utilizado no Fe-BR é o mesmo usado naquele sistema. Isso possibilita estabelecer comunicação direta do SISB com o Fe-BR, o que permite atualizar os dados do primeiro caso ajustes nos dados sejam feitos durante o seu processamento no Fe-BR.

1.1.2 `dataset_titulo`

Título do conjunto de dados, geralmente relacionado ao título do projeto, tese, dissertação, etc que o gerou. Preferência é dada ao uso de título específico para o conjunto de dados, o que identifica melhor o seu conteúdo do que o 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, 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'.*

1.1.3 `dataset_descricao`

Descrição do conjunto de dados, contendo dados básicos para maximizar o reuso futuro do conjunto de dados sem que seja necessário entrar em contato com os seus autores/responsáveis. A descrição do conjunto de dados inclui 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 deve incluir também:

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 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.

1.1.4 `dataset_versao`

Versão do conjunto de dados. Dado necessário para identificar os casos em que o conjunto de dados foi 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 – `dataset_versao = 1` –, é modificado consideravelmente de maneira que todos os perfis do solo possuam coordenadas espaciais aproximadas dos locais de observação – `dataset_versao = 2`.

Sempre que uma nova versão de um conjunto de dados é preparada, uma descrição sumária das modificações deve ser incluída na descrição do conjunto de dados em `dataset_descricao`. Os responsáveis pelas modificações devem ser identificados conforme descrito abaixo.

1.1.5 `dataset_licenca`

Licença de uso do conjunto de dados. Dado que define como 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 gerados por projetos mantidos via financiamento público. Contudo, existe algum consenso de que uma licença de uso 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 aos autores/responsáveis pelo conjunto de dados original. Assim, trata-se de uma das licença mais flexível dentre as licenças Creative Commons disponíveis, maximizando a disseminação e uso dos conjuntos de dados. Assim, a licença CC-BY 4.0 é a licença padrão do Fe-BR. Maiores informações sobre as licenças de uso recomendadas para conjuntos de dados públicos podem ser encontradas no fórum da Infraestrutura Nacional de Dados Abertos (INDA).

Para os conjuntos de dados gerados por projetos mantidos via financiamento privado, onde as partes envolvidas têm interesse em compartilhar os dados desde que não para fins comerciais, recomenda-se usar a licença Creative Commons CC-BY-NC 4.0. A licença CC-BY-NC 4.0 possui praticamente os mesmos termos da licença CC-BY 4.0, exceto pelo fato de que a distribuição, remixação, adaptação e derivação de outros produtos não podem ser usados para fins comerciais.

1.1.6 `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 o gerou ou utilizou pela primeira vez, mas sim a data em que o conjunto de dados foi tornado efetivamente público e disponível para acesso. Quando modificações são

feitas no conjunto de dados durante ou após a sua inserção no Fe-BR, resultando assim em uma nova versão, então a data de publicação é a data em que essa nova do conjunto de dados foi disponibilizada no Fe-BR.

A data de publicação é apresentada na sequência dia, mês e ano, usando formato numérico dd-mm-aaaa ou dd/mm/aaaa, ou seja, dois dígitos para o dia e mês, e quatro dígitos para o 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 algum dos dados não é conhecido, usa-se o símbolo `xx` em substituição. Caso a data de publicação do conjunto de dados seja completamente desconhecida, usa-se a data de publicação do trabalho que o gerou ou utilizou pela primeira vez.

1.1.7 `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, pode-se optar por identificar a organização que atualmente detém responsabilidade técnica sobre o conjunto de dados, ou seja, a organização provedora do conjunto de dados. Isso é especialmente importante caso modificações consideráveis tenham sido feitas nos dados. O mesmo pode ser aplicado aos conjuntos de dados gerados por duas ou mais organizações. No caso de conjuntos de dados gerados ou sob responsabilidade de organizações como a Embrapa, que possuem várias unidades, 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`, o nome da organização,
- `organizacao_url`, o endereço da organização na Internet,
- `organizacao_pais_id`, o país da organização,
- `organizacao_municipio_id`, a cidade da organização,
- `organizacao_codigo_postal`, o código postal da organização,
- `organizacao_rua_nome`, a rua da organização,
- `organizacao_rua_numero`, o número da organização em sua rua.

1.1.8 `autor_<...>`

Identificação dos autores ou responsáveis pela geração do conjunto de dados. Quando o acesso aos autores é impossível ou limitado, identifica-se as pessoas que atualmente detém responsabilidade técnica sobre o conjunto de dados. No caso de trabalhos acadêmicos, como monografias, dissertações e teses, identifica-se o autor e o orientador principal do trabalho acadêmico.

Os dados necessários para a indentificação dos autores do conjunto de dados são os seguintes:

- `autor_nome`, o nome completo dos autores, e
- `autor_email`, o endereço de e-mail atual dos autores.

Sempre que um conjunto de dados tiver dois ou mais autores, insere-se os respectivos dados de identificação separados por ponto e vírgula. É importante que o endereço de e-mail dos autores do conjunto de dados seja válido e atual. Isso é fundamental para permitir o contato com os autores sempre que surgirem dúvidas sobre o conjunto de dados. O conhecimento do nome completo dos autores também é fundamental para permitir os devidos créditos lhes sejam atribuídos sempre que o conjunto de dados for distribuido, remixado, adaptado ou usado para criar outros produtos.

1.1.9 `contribuidor_<...>`

Identificação dos autores ou responsáveis por modificações no conjunto de dados. Modificações constituem contribuições à melhoria do conjunto de dados, sendo assim apresentadas na sua descrição e geralmente resultando em uma nova versão do conjunto de dados. Como a inserção de um conjunto de dados no Fe-BR sempre requer algum tipo de modificação, os integrantes da equipe do Fe-BR sempre figuram como autores de contribuições à sua melhoria.

Os dados necessários para a indentificação dos responsáveis pelas modificações feitas num conjunto de dados são os seguintes:

- `contribuidor_nome`, o nome dos autores das modificações,
- `contribuidor_email`, o endereço de e-mail atual dos autores das modificações, e
- `contribuidor_organizacao`, a organização à qual os autores das modificações estão afiliados.

A apropriada identificação dos autores das modificações no conjunto de dados é fundamental para permitir o contato com os mesmos sempre que surgirem dúvidas sobre as novas versões do conjunto de dados. Além disso, permite atribuir o devido crédito quando se deseja referenciar uma versão específica do conjunto de dados. Contudo, recomenda-se sempre dar crédito aos autores originais do conjunto de dados, independentemente da sua versão.

1.1.10 `dataset_referencia_i`

Referência permanente e válida a documentos ou artigos científicos onde a versão original e/ou a versão atual do conjunto de dados foi usado, preferencialmente, pela primeira vez. Preferência é dada ao uso de um Digital Object Identifier (DOI), mas uma URL também é aceita.

Tantas referências quantas forem julgadas pertinentes podem ser inseridas, numerando-as sequencialmente em ordem de importância, usando para isso o índice `i`. No caso de os responsáveis atuais pelo conjunto de dados não serem os seus autores originais, é imprescindível que a referência principal, `i = 1`, seja a um trabalho dos autores do conjunto de dados. Já a referência secundária, `i = 2`, pode ser a um trabalho dos responsáveis atuais pelo conjunto de dados. No caso de trabalhos acadêmicos, como monografias, disserações e teses, pode-se inserir 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 modificações consideráveis, uma referência ao trabalho onde a nova versão tenha sido usada pela primeira vez também deve ser inserida. Para inserir novas referências, basta inserir, na sequência, novas linhas na tabela `dataset`, usando o `i` para indicar sua ordem.

1.1.11 `area_conhecimento`

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

- 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

1.1.12 palavras_chave

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

1.1.13 categoria_vcge

Categoria do Vocabulário Controlado do Governo Eletrônico (VCGE). O VCGE é um esquema padronizado de assuntos e categorias usado para facilitar a apresentação e identificação dos serviços disponibilizados em uma estrutura de diretórios *online*. O objetivo do VCGE é ajudar os cidadãos a encontrar informações em catálogos de dados públicos sem a necessidade de conhecer a organização responsável pelo assunto ou categoria.

1.2 Tabela *observacao*

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

1.2.1 observacao_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 são 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: `observacao_id = Perfil-01`.

1.2.2 observacao_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 `observacao_id`. O valor padrão, para os casos em que a data de observação é desconhecida, é `observacao_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`.

1.2.3 coord_sistema

Sistema de referência de coordenadas (SRC) utilizado para o georreferenciamento das observações do solo. A especificação do SRC é fundamental para possibilitar o uso apropriado de dados espaciais, especialmente para fins de correlação/cruzamento com outros dados espaciais. Como a especificação do SRC 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.

1.2.4 coord_x; coord_y

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

1.2.5 `coord_precisao`; `coord_fonte`

Precisão com que as coordenadas espaciais foram determinadas (em metros).

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

1.2.6 `pais_id`; `estado_id`; `municipio_id`

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.

1.2.7 `amostra_<...>`

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.

1.2.8 `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, insere-se tantas colunas quantas forem necessárias.

1.3 Tabela *camada*

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

1.3.1 observacao_id

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

1.3.2 camada_<...>

camada_numero

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 `camada_number = 1`.

camada_nome

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 `camada_nome = NA`.

1.3.3 amostra_codigo

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 `amostra_codigo = 1`.

1.3.4 profund_<...>

Profundidade do limite superior, `profund_sup`, e inferior, `profund_inf`, da camada ou horizonte observados.

1.3.5 fe_<extração>_<determinação>

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

1.4 Tabela *metadado*

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.