Integração de *Data Lakes* Pedológicos através de *Workflows ETLH*

Sabrina Santos Cruz de Oliveira[1], Emerson de Barros Duarte[1], Elton Carneiro Marinho[2], Sérgio Manuel Serra da Cruz[1,2]

[1]Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais – Universidade Federal Rural do Rio de Janeiro
Caixa Postal 74.583 – 26.285-060 – Nova Iguaçu – RJ – Brasil

[2]Programa de Pós-Graduação em Informática – Universidade Federal do Rio de Janeiro
{sabrina, serra}@pet-si.ufrrj.br, emersonbd@ufrrj.br,
elton.marinho@ppgi.ufrj.br

Abstract. *Pedology is the science that studies the soil. Currently, datasets from pedological projects are isolated, disconnected and dispersed in the Web in the most varied formats and classifications. This work presents an approach and experiments based on ETLH workflows capable of being coupled to the OpenSoils platform to load, clean, transform and harmonize large amounts of legacy soil data, annotating with retrospective provenance on operations and data. The resulting database expands the accessibility, sharing and reuse of harmonized pedological data.*

Resumo. *A pedologia é a ciência que estuda o solo. Atualmente, os datasets oriundos de projetos pedológicos se encontram isolados em data lakes sob os mais variados formatos e classificações. O objetivo deste trabalho é apresentar uma abordagem e experimentos baseados em workflows ETLH capazes de serem acoplados à plataforma OpenSoils para carregar, limpar, transformar e harmonizar grandes massas de dados legados, agregando descritores de proveniência retrospectiva sobres operações e dados. O banco resultante amplia a acessibilidade, localização, compartilhamento e reúso de dados pedológicos harmonizados.*

1. Introdução

A cadeia do agronegócio brasileiro começa a sentir os efeitos de *Big Data*. Elos dessa cadeia passam por grandes transformações e demandam integração de dados legados com informações produzidas em tempo real por tratores, drones e sensores [Agrapresse, 2015]. No entanto, esses dados são de difícil integração do ponto de vista semântico e muitas vezes estão dispersos nas redes. Para aprimorar a gestão agrícola é necessário desenvolver estratégias de agregação de diversas fontes de dados.

A cadeia do agronegócio é dinâmica e complexa. Neste trabalho focaremos em um dos principais recursos da cadeia: o solo. Investigaremos os problemas de integração de dados relacionados aos *data lakes* pedológicos herdados. A pedologia é a ciência que estuda as origens, morfologia, classificações e propriedades físico-químico-biológicas dos solos do planeta [Santos, 2018]. Do ponto de vista da computação, é uma ciência tão intensiva em dados quanto a bioinformática [Cruz et al 2018].

Os *data lakes* pedológicos são grandes conjuntos de dados (semiestruturados e fracamente documentados) oriundos de grandes projetos de mapeamentos de solos que ocorreram nos últimos 50 anos. Sob a ótica da Ciência de Dados eles apresentam duas classes de problemas principais:

I) ***Estruturais*** – arquivos de tipos e formatos diversos, dados incompletos, irregulares, *outliers* e falhas de preenchimentos.

II) ***Semânticos*** – os solos classificados de acordo com taxonomias que são periodicamente atualizadas ao longo do tempo. Desde 2017 se utiliza a 5ª edição do Sistema Brasileiro de Classificação de Solos (SiBCS) [Santos, 2018]. Ou seja, as classes taxonômicas que descrevem propriedades pedológicas variam no tempo, mas os solos previamente classificados com regras anteriores não são automaticamente reclassificados. Essa dinâmica, introduz inconsistências semânticas nos *datasets* ao serem analisados à luz da taxonomia corrente.

Essas duas classes de problemas se evidenciam em algumas bases públicas como, por exemplo, no BDSolos [Embrapa, 2006] e no repositório FeBR [Rosa & Anjos, 2020]. Ambas armazenam grandes volumes de dados legados contendo estruturas e atributos distintos, são de difícil integração e por conseguinte são de difícil atualização, tendo baixo reuso por parte da comunidade. Adicionalmente, as bases não são adequadamente anotadas com padrões de metadados ou enriquecidos com proveniência. Essas limitações trazem impactos nas análises de dados ou na produção de relatórios e mapas digitais, pois poderão incorporar desde vieses analíticos até erros de classificação que afetarão decisões/interpretações sobre os projetos pedológicos.

Como contribuição, este trabalho discute o uso de *workflows* de Extração-Transformação-Carga-Harmonização (*Extraction-Transformation-Load-Harmonization - ETLH*) que mapeiam a taxonomia SiBCS no contexto de tratamento de dados legados de projetos pedológicos. Os *workflows* e experimentos foram capazes de tratar e carregar grandes volumes de dados pedológicos herdados, transformando-os em dados harmonizados e anotados com proveniência retrospectiva (que consiste na captura os dados das etapas que foram executadas e informações sobre este ambiente) e carregando-os diretamente na plataforma *OpenSoils* [Cruz et al., 2018, 2019].

O presente artigo está organizado da seguinte forma: após esta introdução, é apresentada uma fundamentação teórica, seguida pelos trabalhos relacionados na Seção 3. Materiais e métodos na Seção 4, o desenvolvimento e os resultados na Seção 5 e por fim, a conclusão, limitações e trabalhos futuros na Seção 6.

2. Referenciais teóricos

2.1. *Dados Pedológicos*

Os solos são recursos naturais não renováveis. Fisicamente o solo é composto por diversos materiais dispostos em camadas; sua análise é conduzida através de aberturas denominadas trincheiras (ou perfis) que são realizadas diretamente no campo onde pedólogos, agrônomos ou geólogos coletam diversos tipos de dados morfológicos e georreferenciados (imagens,

descrições das camadas, profundidades, composição física e química, transições, cores, erosões, texturas, entre outros atributos das camadas) [Solos, 2013].

Os mapeamentos digitais dos solos são obtidos através de observações diretas e experimentos de difícil reprodutibilidade. Essas atividades são realizadas diretamente no campo ou em laboratórios "de campanha" e envolvem muitos profissionais, instrumentos pesados ou sensores sensíveis. Os dados são posteriormente complementados por análises químicas e físicas realizadas em laboratórios especializados. Em geral, os *datasets* pedológicos são volumosos e gerados por equipes distintas e dispersas no tempo e, por vezes, geograficamente distantes [McBratney & Minasny, 2006]. Produz-se (involuntariamente) *data lakes* desconectados que contém várias falhas em séries não harmonizadas. Apesar disso, são de extrema importância para diversas áreas além da fertilidade na agricultura, como por exemplo para sustentabilidade, mudanças climáticas, economia e ecologia.

Uma parte dos dados de solos estão dispersos na Web ou em repositórios públicos e privados [Rosa e Anjos, 2020]. Constata-se que sua localização não é simples e seu acesso não é transparente, o que dificulta sua reutilização tanto por gestores e pesquisadores quanto por agricultores. [Cruz et al., 2018]. Adicionalmente, muitos *datasets* são dispersos em mapas de papel ou mesmo em páginas HTML, arquivos PDF e planilhas armazenadas em servidores sob os mais variados formatos. Não raro, possuem diferentes estruturas semânticas e dados não harmonizados. Isto é, em vários casos não se usam os mesmos padrões de unidades de medidas para os mesmos atributos pedológicos ao longo do tempo, acentuando ainda mais as inconsistências dos dados.

2.2. Processos ETLH em Pedologia

Data Cleaning é um conjunto de técnicas que detectam e removem anomalias nos dados (sintáticas, semânticas e de cobertura) [Rahm & Do, 2000]. A limpeza de dados pedológicos adota diversas abordagens gerais ou especializadas. As gerais são voltadas para tratar tipos de problemas já conhecidos, como por exemplo a detecção de *outliers*, registros duplicados, nulos, falhos, etc. Nesta pesquisa usamos processos *ETLH* para representar as tarefas de *Data Cleaning*, fundir *data lakes*, customizá-los e produzir *datasets* enriquecidos [Simitsis, 2003].

O *ETLH* usou as especificações da taxonomia SiBCS para harmonizar os dados pedológicos ao implementar as tarefas especializadas voltadas para tratar problemas de domínio, ou seja, os que demandam conhecimentos na área de pedologia. Podemos citar como exemplo a correção de coordenadas geográficas, faixas de pH, tabelas de cores, concentrações de elementos, percentuais silte-areia-argila, entre outros.

2.3. Proveniência de dados

O conceito de proveniência de dados refere-se à origem ou à procedência de um determinado objeto [Davidson & Freire, 2008]. Entretanto, os aspectos fundamentais da proveniência não se resumem apenas aos dados, mas também aos processos produtores de dados. No momento, apesar de serem aplicados há tempos na e-ciência, os estudos de proveniência de dados na área da pedologia ainda são incipientes.

No que tange aos dados pedológicos, sua proveniência também possui diversas granularidades (fina e grossa) [Davidson & Freire, 2008]. Neste trabalho adotamos a retrospectiva de baixa granularidade nos *workflows* para anotar a harmonização dos dados.

3. Trabalhos relacionados

O BDSolos[1] é um banco de dados desenvolvido pela EMBRAPA que demanda alto conhecimento de solos para que seja plenamente consultado, além disso apresenta limitações na extração de dados. Por outro lado, o FeBR[2] surgiu como um disco virtual público, mas evoluiu para um conjunto de planilhas de projetos pedológicos. Em linhas gerais, importou parte dos dados do BDSolos e disponibilizou com uma organização que possibilita que qualquer pessoa, mesmo com poucos conhecimentos sobre pedologia, consiga ter acesso aos projetos pedológicos.

A organização no FeBR consiste, basicamente, em três arquivos no formato .txt para cada projeto, com valores separados por tabulação. Um contém dados de identificação do projeto, outro os dados das observações realizadas no projeto e por fim, um com os dados pedológicos das descrições das camadas. O FeBR está direcionado apenas a plataformas desktop.

4. Materiais e Métodos

4.1. *OpenSoils*

OpenSoils (www.opensoils.org) é uma plataforma computacional gratuita, aberta, elástica, distribuída, multiusuário e multicamada (Figura 1). É orientada para armazenar dados curados e harmonizados (primários e secundários) de solos do Brasil e seus metadados de proveniência [Cruz et al., 2018, 2019].

A plataforma oferece facilidades aos usuários que vão desde a coleta e organização até o armazenamento de longo prazo. Ou seja, é uma alternativa aos *data lakes*. Ela pode ser utilizada em diversas etapas dos projetos pedológicos que começam diretamente no campo, passam por laboratórios especializados e vão até os processos computacionais de análise/visualização de dados e aprendizado de máquina visando a extração do conhecimento.

Atualmente, *OpenSoils* conta com uma versão Web e aplicativos móveis que se comunicam através de APIs e troca síncrona de dados. O banco de dados do *OpenSoils* é do tipo relacional, seu esquema lógico é capaz de armazenar uma grande quantidade de dados de propriedades pedológicas que seguem definições da 5a. edição da taxonomia SiBCS.

O *schema* completo possui 46 tabelas, porém, neste artigo, destacamos 22 tabelas que armazenam dados pedológicos harmonizados utilizados em projetos pedológicos. No fragmento exibido, podemos correlacionar dados morfológicos que são produzidos em ambientes distintos. Por exemplo, há integração de dados coletados em campo (tabelas *projeto, observação, relevo, descrição geral, horizontes, morfologia entre outros*) com dados obtidos em laboratório através de experimentos físicos (*curva de retenção de água*) e químicos (*ataque sulfúrico, pasta saturada*).

[1] https://www.bdsolos.cnptia.embrapa.br/consulta_publica.html
[2] https://www.pedometria.org/febr/

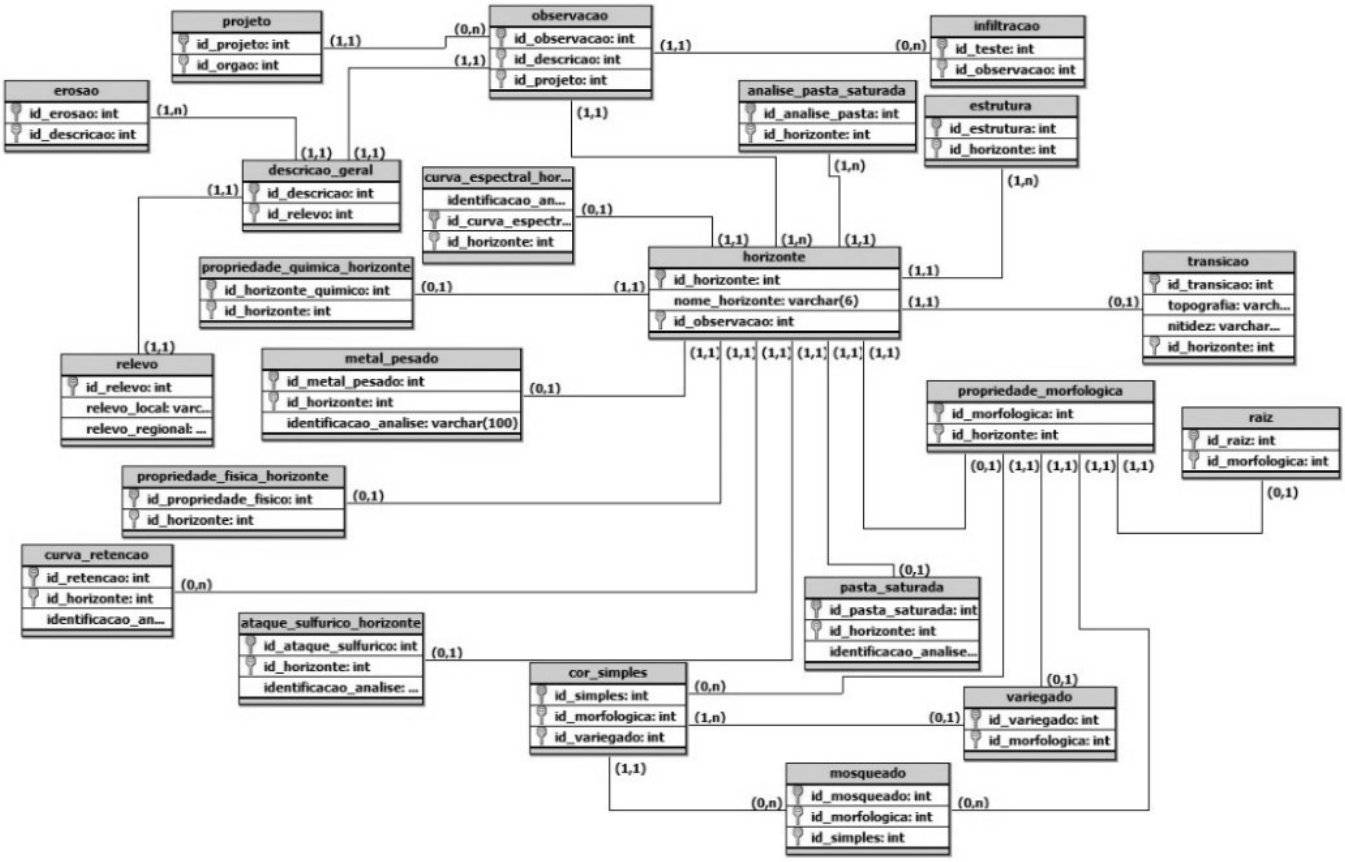

Figura 1. Fragmento com parte do esquema conceitual do *OpenSoils*

4.2. *Workflows* ETLH

Os *workflows ETLH* foram desenvolvidos no Pentaho Data Integration (PDI) [Hitachi, 2004], considerando a SiBCS. Os *workflows* foram incorporados à plataforma *OpenSoils*. Podem ser executados pelos gestores da plataforma para fazer cargas massivas de dados a partir dos *data lakes* pedológicos.

Os *workflows ETLH* são compostos por fluxos de *tasks* e *jobs*. As *tasks* são tarefas de conexão, captura de dados, verificação de nulos, duplicatas, *outliers* ou de validação, harmonização de unidades/dados, ajuste de sistemas de coordenadas etc. As *tasks* são as unidades "mínimas de processamento" dentro do processamento dos *jobs* de um *workflow*.

As primeiras *tasks* dos *jobs* são responsáveis por validações e harmonizações segundo a SiBCS (por exemplo: validações de relevo, descrição geral, observação, erosão, horizonte, propriedade química do horizonte, propriedade física do horizonte, propriedade morfológica, ataque sulfúrico, transição, estrutura, análise pasta saturada, consistência, cerosidade, raiz, nódulos e concreções, superfície de fricção, cor simples, mosqueado e metal pesado). A Figura 2 ilustra uma *task* relacionada com a análise dos dados de erosão e o registro da proveniência retrospectiva das transformações.

Um dos maiores desafios técnicos enfrentados na construção dos *workflows ETLH*, foi a inexistência no PDI de uma funcionalidade que realizasse automaticamente o relacionamento entre os registros e as tabelas da plataforma *OpenSoils*. Desenvolvemos,

nesse sentido, a *task* Execute SQL, um *script* capaz de executar qualquer declaração SQL diretamente no banco de dados da plataforma *OpenSoils*, a solução foi adotada para realizar as atualização das chaves estrangeiras e manter o relacionamento dos dados de cada projeto.

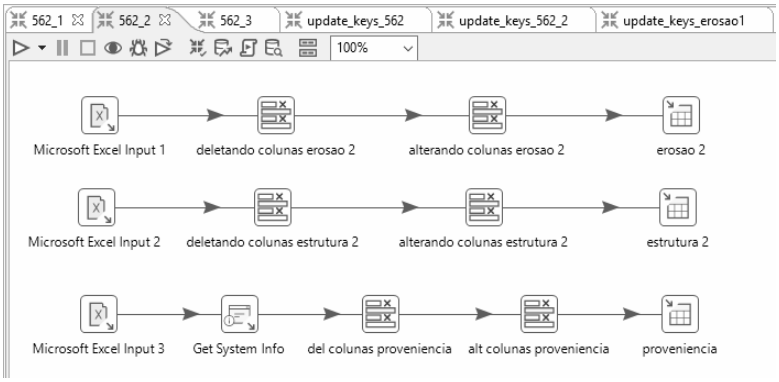

Figura 2. Exemplo de *task* que realiza transformações em atributos pedológicos

A execução das consultas SQL pelos *workflows ETLH* requer a análise da lógica de relacionamento das tabelas envolvidas. Existem quatro *tasks* responsáveis por manter a consistência e atualizar os campos de chaves estrangeiras. Por exemplo, a Figura 3 ilustra um fragmento de uma declaração SQL inserida na *task* Execute SQL, ela atualiza dados harmonizados e cria o relacionamento entre as tabelas *revelo* e *descrição geral*.

```
update descricao_geral g join
       (select (@rng := @rng + 1) as rn, g2.id_descricao
        from (select g2.* from descricao_geral g2 order by g2.id_descricao desc)
             (select @rng := 0) params
        limit 6
       ) g2
       on g.id_descricao = g2.id_descricao join
       (select (@rnr := @rnr + 1) as rn, r.id_relevo
        from (select r.* from relevo r order by r.id_relevo desc) r cross join
             (select @rnr := 0) params
       ) r
       on g2.rn = r.rn
   set g.id_relevo_fk = r.id_relevo;
```

Figura 3. Atualização do atributo relevo *id_relevo_fk* na tabela descrição geral

Adicionalmente, foram concebidas *tasks* específicas (atuantes como *sub-workflows*) que capturam a proveniência retrospectiva sobre as transformações realizadas pelas demais tarefas dos *workflow*s. Dessa maneira, os metadados de proveniência retrospectiva aderentes à especificação PROV[3] da W3C são persistidos na tabela *proveniência*. Dentre eles, destacam-se *timestamp*, processo, usuário e o IP de quem faz operações *ETLH* sobre cada item de dados dos projetos. Essa *task* aumenta a rastreabilidade dos dados e contribui para aumentar a confiabilidade dos dados pedológicos da plataforma *OpenSoils*.

Figura 4. Exemplo de um *simple job* no *OpenSoils*

A seguir, desenvolvemos os *simple jobs* para incorporar o conjunto de *tasks* capazes de processar as sequências de operações *ETLH*. Por exemplo, um *simple job* se inicia com

[3] w3.org/TR/prov-primer/

o processamento da *task* START (obrigatória em qualquer fluxo) e segue a lógica de limpeza e harmonização daquele atributo. A Figura 4 ilustra um *simple job* com todas as *tasks* de um projeto. A *task* 562_1 refere-se à primeira parte de tratamento e carga de dados do projeto 562. O mesmo ocorre em 562_2 e 562_3. As demais *tasks* são responsáveis única e exclusivamente pela atualização de chaves e garantem a semântica e consistência dos dados do projeto 562 no banco de dados relacional.

Os *workflows ETLH* da plataforma reúnem todos os *simple jobs* em um único *job* composto. Dessa forma, a execução é automatizada e se dá de uma só vez fazendo a carga/transformações dos *data lakes*. Os *workflows ETLH* do *OpenSoils* são parametrizáveis e exportam apenas as colunas contendo os dados harmonizados e a proveniência. Utilizam *tasks* Execute SQL para a realização das consultas e em seguida, encadeiam-se com *task* Microsoft Excel Output que transformam os dados em planilhas no formato .xls.

5. Experimentos Computacionais e Resultados

Os *workflows ETLH* refletem os mapeamentos entre os atributos pedológicos dos projetos com o esquema relacional do *OpenSoils*. Para isso, foram executadas, por projeto, oito transformações responsáveis pela limpeza dos dados, transformação, harmonização, carga e criação de relacionamentos a partir da inserção das chaves estrangeiras, possibilitando o relacionamento entre as tabelas de dados pedológicos.

Os experimentos mediados pelos *workflows ETLH* coletaram dados de perfis de solos dispersos em centenas de planilhas de pedológicos registrados no FeBR, além das tabelas do BDSolos. A execução dos experimentos modelados como um *job* composto por *tasks*, processou 193 projetos pedológicos, totalizando mais 9 mil perfis de solos em mais 800 mil registros carregados na plataforma *OpenSoils* em pouco mais de 2 horas de processamento. Cada um deles passou pelo processo de *ETLH*, renomeação de colunas, relacionamento e consistência de chaves e registro de proveniência.

Os dados produzidos pelos experimentos são abertos e oriundos de *data lakes* de projetos pedológicos realizados durante décadas em todo o Brasil. Agora, podem ser integralmente acessados tanto pela plataforma *OpenSoils* Web e sua API quanto pelo app *OpenSoils* Edu [Cruz, 2018], já disponível na *Google Play Store*.

6. Conclusão

Os solos do Brasil representam ao mesmo tempo um patrimônio público insubstituível e um recurso natural não renovável que se encontroa sob permanentes riscos de erosão, contaminação e maus usos. No entanto, apesar de serem um elemento central na cadeia do agronegócio, verificam-se poucas pesquisas sobre a gestão eficiente de longo prazo dos *datasets* pedológicos. Os atuais sistemas apresentam gargalos relacionados a dispersão alta e baixa integração de dados, resultando em pouca transparência e limitadas informações de proveniência de dados. Essas condições reduzem a acessibilidade, interoperabilidade e reuso de dados pedológicos.

Este artigo apresenta um esforço para mitigar as limitações do *data lakes* pedológicos, produzir *datasets* harmonizados e anotados por proveniência sobre dados legados. Desenvolvemos *workflows ETLH* acopláveis à plataforma *OpenSoils* que são capazes de harmonizar dados segundo a SiBCS e anotá-los com proveniência retrospectiva. Como contribuição, estruturamos e disponibilizamos os dados pedológicos em um banco de dados relacional para que pudessem ser acessados pelos usuários a partir dos aplicativos

móveis do *OpenSoils*. Além disso, com breves ajustes ou seguindo-se o modelo de entrada atual, os workflows podem ser utilizados para outros formatos de dados de entrada e de saída. Nossos experimentos foram capazes de processar e harmonizar centenas de milhares de registros de perfis de solos em todo o Brasil.

Como trabalhos futuros pretendemos difundir a plataforma em sociedades e programas nacionais de solos e, aprofundar as investigações sobre o banco do *OpenSoils*, disponibilizá-lo de forma compatível com os princípios FAIR [Wilkinson et al., 2016] e incorporar processos de *Smart Contracts* e FAIRificação de dados pedológicos [Marinho et al., 2020].

Referências

Agrapresse (2015). Big data: une nouvelle révolution agricole en marche. Hebdo, Agra Presse, pp. 1–7.

Cruz, S. M S., et al. (2019). Desenvolvendo Sistemas Agrícolas de Próxima Geração: Um Estudo em Ciência de Solos. In Anais do X Workshop de Computação Aplicada a Gestão do Meio Ambiente e Recursos Naturais (pp. 135-144). SBC.

Cruz, S. M. S. et al. (2018) "Towards an e-infrastructure for Open Science in Soils Security". In: XII BRESCI 2018, 2018, Recife. Proceedings of the XII Brazilian E-Science Workshop. Porto Alegre: SBC.

Davidson, S. B.; Freire, J. (2008) "Provenance and scientific workflows: challenges and opportunities". In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2008. (SIGMOD '08), p. 1345–1350.

Hitachi, V. (2004). "Pentaho Data Integration".

Marinho, E. C. et al. (2020). "Proteção de Dados: Proposta de gerenciamento de dados de solos usando os princípios FAIR e a tecnologia blockchain". In: 10ª. Conferencia de Directores de Tecnología de Información y Comunicación en Instituciones de Educación Superior, TICAL2020 y 4° Encuentro Latinoamericano de e-Ciencia. Equador.

McBratney, A. B.; Minasny, B. Australian Centre for Precision Agriculture, Faculty of Agriculture, Food and Natural Resources, McMillan Building A05, The University of Sydney, Sydney, New South Wales 2006, Australia.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3–13.

Rosa, A. S., Anjos, M. A. (2020). Uma plataforma para facilitar o acesso aos dados do Repositório Brasileiro Livre para Dados Abertos do Solo. SEI-SICITE.

Santos, H. G. et al. (2018). Sistema brasileiro de classificação de solos. 5. ed. rev. e ampl. Brasília, DF: Embrapa.

Solos, Embrapa. (2013). Sistema brasileiro de classificação de solos. Centro Nacional de Pesquisa de Solos: Rio de Janeiro.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.