Uso de *workflows ETLH* para integrar *datasets* pedológicos: estudo para adequação aos princípios FAIR

Sabrina Santos Cruz de Oliveira[1], Emerson de Barros Duarte[1], Élton 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. *Currently, datasets from pedological projects are isolated in silos under the most varied formats. The objective of this work is to present an approach and experiments based on ETLH and FAIRification workflows capable of loading, cleaning, transforming, identifying and harmonizing large masses of legacy data on the OpenSoils platform. Additionally, we discuss the main steps of the FAIRification process.*

Resumo. *Atualmente, os datasets oriundos de projetos pedológicos se encontram isolados em silos sob os mais variados formatos. O objetivo deste trabalho é apresentar uma abordagem e experimentos baseados em workflows ETLH e de FAIRificação capazes de carregar, limpar, transformar, identificar e harmonizar grandes massas de dados legados na plataforma OpenSoils. Adicionalmente, discutimos as principais etapas do processo de FAIRificação.*

1. Introdução

Estudos recentes da Embrapa, SEBRAE e INPE[1] indicam que a agricultura digital apoiada na Ciência de Dados e na acelerada adoção de tecnologias digitais têm o potencial de transformar importantes setores da cadeia do agronegócio brasileiro. O estudo indica que a internet no campo é o principal atrativo para os agricultores que buscam ampliar ou conquistar novos mercados, reduzir custos ou agregar valor à produção. Portanto, é possível conjecturar que, em breve, essa cadeia, demandará novas práticas

[1]https://www.embrapa.br/en/busca-de-noticias/-/noticia/54770717/pesquisa-mostra-o-retrato-da-agricultura-digital-brasileira

correlacionadas com a gestão de grandes volumes de dados para auxiliar nos processos de tomada de decisão.

Segundo Jonquet *et al*. (2018), os dados agronômicos são de difícil integração e interoperabilidade tanto do ponto de vista técnico quanto semântico. Análises mais refinadas dos *datasets* produzidos nesta cadeia podem auxiliar tanto em sistemas mais inteligentes quanto numa produção mais sustentável [Cruz *et al*., 2018]. No caso dos solos, as dificuldades também são grandes. Portanto, disponibilizar sistemas mais transparentes e sustentáveis que indiquem os melhores tipos de solos para usos adequados, sequestro de carbono, racionalização do uso de recursos naturais e redução da aplicação fertilizantes e produtos químicos são de grande relevância.

Os desafios e as oportunidades para desenvolver pesquisas e inovação na cadeia do agronegócio são variadas, mas ainda timidamente exploradas pela comunidade brasileira de computação. Neste trabalho focaremos em um dos principais recursos da cadeia: os dados de solos. Segundo Bessa e Cruz (2021), esse é um dos elementos centrais na ligação entre diversas áreas do conhecimento, tais como computação, segurança de solos e humanidades digitais.

Como objeto de estudo, investigaremos as necessidades do processo de FAIRificação de dados relacionados aos *datasets* de dados pedológicos. Destacamos que a Pedologia é uma ciência complexa que ainda não se apropriou da perspectiva da Ciência Aberta para a gestão de seus *datasets* apoiados pelos princípios FAIR [Wilkinson *et al*., 2016]. Do ponto de vista da computação, determinados aspectos da Pedologia podem ser considerados como uma ciência tão intensiva em dados quanto a Bioinformática pois envolve a correlação de múltiplas variáveis ambientais [Cruz *et al*., 2018].

Segundo Oliveira *et al*. (2021), os *datasets* de dados pedológicos caracterizam-se por serem grandes conjuntos de dados oriundos de grandes projetos de mapeamento de solos que ocorreram nos últimos 60 anos. Por exemplo, o RADAMBRASIL[2], projeto governamental criado em 1970, realizou um inventário de grande parte do País usando mapeamento em escala 1:1.000.000. Em geral, esses dados não são disponibilizados para a sociedade brasileira como dados abertos governamentais, muitas vezes permanecendo inacessíveis. Sob a ótica da Ciência de Dados, a maioria desses *datasets* apresentam duas categorias de problemas principais:

- ***Problemas Estruturais*** – Segundo Rosa & Anjos (2020), há pouco acesso e estímulos a políticas de dados primários abertos na área. Adicionalmente, os arquivos são dispersos em vários servidores, sendo de tipos e formatos diversos, muitas vezes incompletos, irregulares, contendo *outliers* e mesmo falhas de preenchimentos. Adicionalmente, os dados são pouco transparentes e estão desconectados em silos de dados.

- ***Problemas Semânticos*** – Via de regra, os solos do Brasil são classificados de acordo com taxonomias que são periodicamente atualizadas. Desde 2017 é utilizada a 5ª edição do Sistema Brasileiro de Classificação de Solos (SiBCS) [Santos, 2018]. As classes taxonômicas descrevem propriedades e atributos pedológicos cujas classificações podem variar no tempo. No entanto, devido a uma série de condicionantes, os solos

[2] https://www.embrapa.br/en/pronasolos

previamente classificados com regras anteriores não são reclassificados automaticamente à medida que novas taxonomias são liberadas. Essa dinâmica, involuntariamente, agrega inconsistências semânticas nos *datasets* se forem (re)analisados à luz da taxonomia corrente. Adicionalmente, se constata que atualmente inexistem *datasets* públicos e abertos totalmente alinhados aos princípios FAIR ou aderentes aos princípios da Ciência Aberta.

Essas categorias de problemas se evidenciam em duas bases públicas de dados de solos do Brasil: BDSolos[3] e no repositório FeBR[4]. Ambas são de grande valor, armazenam expressivos volumes de dados legados contendo estruturas e atributos distintos. No entanto, são de difícil integração e por conseguinte de difícil atualização, sua integração demanda esforços de processamento, isso muitas vezes é oneroso para pesquisadores e instituições, levando baixo reuso por parte da comunidade.

Como contribuição, este trabalho apresenta *workflows* que desempenham tarefas de Extração-Transformação-Carga-Harmonização (*ETLH*) e de FAIRificação que consideram a taxonomia SiBCS atual no contexto do tratamento e estabelecimento de relacionamentos entre dados legados de projetos pedológicos. Os experimentos foram capazes de tratar grandes volumes de dados pedológicos legados, transformando-os em dados harmonizados e anotados com proveniência retrospectiva e carregando-os diretamente na plataforma *OpenSoils* [Cruz *et al*., 2018, 2019]. Além disso, este trabalho aprofunda discussões sobre o trabalho de Oliveira *et al*. (2021), pontuando as adequações dos dados para que se alinhem aos princípios FAIR, onde discutimos os pontos que ainda necessitam de evolução e melhorias.

Este trabalho está organizado da seguinte forma: as seções 2, 3 e 4 abordarão uma visão geral sobre *datasets* pedológicos, princípios FAIR e proveniência de dados, respectivamente. A seção 5 apresenta os trabalhos relacionados, a seção 6 os materiais e métodos e na seção 7 discutimos os resultados obtidos até o momento. Na seção 8 apresentamos as conclusões e indicações de trabalhos futuros.

2. *Datasets* Pedológicos

Fisicamente o solo é composto por diversos materiais dispostos em camadas; sua análise é conduzida através de aberturas denominadas tradagens, trincheiras (ou perfis) que são realizadas diretamente no campo onde pedólogos, agrônomos ou geólogos coletam diversos tipos de dados ambientais e morfológicos georreferenciados (imagens, descrições das camadas, profundidades, composição física e química, transições, cores, erosões, texturas, entre outros atributos das camadas) [Santos, 2018].

Os dados de solos são obtidos através de observações diretas e em experimentos que algumas vezes podem ser considerados de difícil reprodutibilidade. Essas atividades são realizadas diretamente no campo ou em laboratórios "de campanha" e podem envolver muitos profissionais, instrumentos e sensores. Os dados são posteriormente complementados por diversos tipos de análises laboratoriais realizadas em ambientes especializados. Em geral, os *datasets* pedológicos são volumosos e gerados por equipes distintas com especialidades diversas e dispersas no tempo e, por vezes, geograficamente

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

distantes. Com isso, acabam produzindo, involuntariamente, *datasets* desconectados que contém várias falhas em séries de dados não harmonizadas.

Uma parte dos dados de solos já está dispersa na Web ou em repositórios públicos e privados formando silos de dados desconectados. Constata-se que a recuperação dos *datasets* não é um processo trivial e seu acesso não é transparente e nem mesmo automatizado, o que dificulta sua reutilização tanto por gestores quanto por pesquisadores. Porém, essa situação penaliza especialmente os agricultores que muitas vezes se veem excluídos do acesso a esse patrimônio digital.

Muitos *datasets* não são classificáveis como dados abertos. Por exemplo, existem apenas seis *datasets* abertos no portal de dados abertos do Governo Federal[5]. Mesmo no sítio do recentíssimo programa PRONASOLOS[6], os *datasets* não estão abertos e nem disponíveis ao público. Neste caso, atualmente, são ofertadas apenas interfaces de visualização de dados consolidados sob a forma de mapas.

Muitas vezes tais dados ainda estão dispersos em mapas de papel ou mesmo em arquivos fechados e planilhas armazenadas em servidores sob os mais variados formatos. Não raro, os dados são apresentados de modo agregado na forma de tabelas com estatísticas ou possuem diferentes estruturas semânticas, com poucos metadados descritivos, ausência quase absoluta de rastros de proveniência [Cruz *et al*., 2019, Bessa *et al*., 2021], por fim, os dados não são harmonizados. Isto é, vários *datasets* não usam as mesmas classificações taxonômicas, variáveis ou padrões de unidades de medidas para os mesmos atributos pedológicos ao longo do tempo, acentuando ainda mais as inconsistências dos dados.

Do ponto de vista da Ciência de Dados, esse conjunto de características prejudica e retarda o desenvolvimento de trabalhos de pesquisa e as entregas de soluções para a sociedade civil. Resumidamente, elas reduzem a vida útil dos dados. Portanto, estudar e difundir práticas mais eficiente ligadas à Ciência Aberta e uso de repositórios públicos confiáveis e focados em preservação digital e baseadas em princípios FAIR e "FAIRificação" de dados poderão trazer maior reprodutibilidade das pesquisas e visibilidade aos *datasets*.

3. Princípios FAIR

Os princípios FAIR foram originalmente publicados em 2016 com o objetivo de fornecer orientações para a aplicação em repositório de dados [Wilkinson *et al*., 2016]. FAIR é um acrônimo para *Findable* (Localizável), *Accessible* (Acessível), *Interoperable* (Interoperável) e *Reusable* (Reutilizável). São princípios orientadores de alto nível que podem ser aplicados em diversas áreas do conhecimento, incluindo dados da cadeia do agronegócio. No caso das ciências agrárias, os princípios FAIR visam promover a agregação e uniformização dos dados e dos sistemas, incluindo os dados pedológicos, sob protocolos que consideram aspectos éticos, legais, culturais e barreiras técnicas, reduzindo custos de gestão dos dados e assegurando sua qualidade e transparência.

Os dados pedológicos e seus metadados devem ser fáceis de serem localizados e individualizados por identificadores persistentes, para que sejam consumidos tanto por humanos quanto por máquinas (*Findable*). Após localizados, os dados precisam ser acessíveis, ou recuperáveis através do seu identificador, mesmos que os dados originais já

[5] https://dados.gov.br/dataset?q=solos&tags=Brasil
[6] http://pronasolos.agenciazetta.ufla.br/

não estejam disponíveis, possivelmente incluindo autenticação e autorização e protocolos comuns (*Accessible*). Os dados precisam se integrar a outros dados (*Interoperable*), precisam interoperar com aplicativos ou fluxos de trabalho distintos para suportar análises, armazenamento e processamento mais eficazes. Logo, o suporte de estruturas semânticas (ontologias, taxonomias, etc.) é essencial. Tudo isso para que os *datasets* possam ser mantidos e reutilizados ao longo do tempo. Os metadados e os dados devem ser bem descritos, para que possam ser replicados e/ou (re)combinados em diferentes configurações (*Reusable*)[7]. Cada um dos quatro princípios possui subdivisões, além das descrições e características sobre sua aplicação. Graças a isso, hoje somos capazes de realizar análises e traçar metas para tornar os dados, independentemente de sua origem e projetar processos de FAIRificados de dados.

FAIRificação é o processo de tornar os dados brutos aderentes aos princípios FAIR, tornando-os inteligíveis para humanos ou máquinas [Veiga *et al.*, 2019]. Esse processo ainda é pouco explorado na agricultura digital e no agronegócio com seus dados agrícolas e pedológicos. No Brasil, os primeiros passos começam a ser trilhados na área Agro através do GO-FAIR Brasil[8].

4. Proveniência de dados

O tema proveniência de dados ainda é pouco difundido na cadeia do agronegócio. O conceito, consagrado na área da computação, refere-se à origem ou à procedência de um determinado objeto [Buneman; Khanna; Tan, 2000 e Davidson & Freire, 2008]. Entretanto, os aspectos fundamentais da proveniência não se resumem apenas aos dados, mas também aos processos e aos agentes transformadores. Apesar de serem aplicados há tempos nas áreas de banco de dados, e-ciência, e humanidades digitais, os estudos de proveniência de dados na área da pedologia e mesmo em ciência de dados ainda são incipientes.

A proveniência dos dados aumenta a confiança do consumidor de dados [Allemang; Bobbin, 2016]. A família de documentos PROV[9] da W3C define um modelo e serializações para permitir o intercâmbio interoperável de informações de proveniência em ambientes heterogêneos como a web.

A proveniência pode ser classificada como prospectiva quando captura a especificação de uma tarefa computacional, seu fluxo de trabalho. Descreve as etapas que devem ser seguidas para gerar um determinado tipo de dado. É a captura de uma especificação abstrata de fluxo de trabalho como uma receita para derivação futura de dados [Lim *et al*., 2010]. A proveniência retrospectiva captura os dados, as etapas que foram executadas e informações sobre este ambiente [Da Cruz; Do Nascimento, 2016]. Neste trabalho adotamos a retrospectiva de baixa granularidade nos *workflows ETLH* e de FAIRificação dos dados.

5. Trabalhos relacionados

O Brasil possui poucos bancos de dados de solos, a maioria dos dados encontram-se em repositórios dispersos e isolados. O BDSolos é um banco de dados relacional

[7] https://www.go-fair.org/fair-principles/
[8] https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-brazil-office/
[9] Um conjunto de documentos que definem vários aspectos que são necessários para alcançar a visão de intercâmbio interoperável de informações de proveniência em ambientes heterogêneos como a web.

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 surgiu como um disco virtual público, mas evoluiu para um conjunto de planilhas de projetos pedológicos. Em linhas gerais, adotou alguns princípios da Ciência Aberta e, importou parte dos dados do BDSolos, disponibilizando-o sob a forma de planilhas que possibilitam 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. Até o momento, não se verificou maior alinhamento aos princípios FAIR ou mesmo aporte do leque de tecnologias de Web Semântica. Em termos internacionais, organizações tais como a ISRIC[10] e a OGC[11] já advogam abertamente a necessidade da adoção dos princípios FAIR como um dos mecanismos de padronização de dados para a área de solos.

6. Materiais e métodos

6.1. Opensoils

OpenSoils (www.opensoils.org) é uma plataforma computacional gratuita, aberta, elástica, distribuída, multiusuário e multicamada. 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].

Atualmente, o *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 pedológicos que seguem definições presentes na 5a. edição da taxonomia SiBCS.

O *schema* completo possui 46 tabelas. É possível correlacionar dados produzidos por processos observacionais de campo com os produzidos em laboratórios. Por exemplo, há integração de dados coletados em campo (tabelas *projeto, observação, relevo, descrição geral, horizontes, morfologia, entre outros*), além de dados obtidos em laboratório através de experimentos físicos (*curva de retenção de água*) e químicos (*ataque sulfúrico, pasta saturada*).

6.2. *Workflows ETLH* e de Fairificação de Dados

Para a carga de dados legados no banco de dados relacional do *OpenSoils*, foram desenvolvidos *workflows ETLH* a partir da ferramenta Pentaho Data Integration (PDI)[12]. Os *workflows ETLH*, originalmente descritos por Cruz *et al*. (2021), são compostos por fluxos de *tasks* e *jobs*. As *tasks* são as unidades "mínimas de processamento" dentro do processamento dos *jobs* de um *workflow*. As *tasks* variam desde simples tarefas de conexão e captura de dados, verificação de nulos, duplicatas, outliers, por exemplo validação, harmonização de unidades/dados, ajuste de sistemas de coordenada, até tarefas classificatórias descritas pela taxonomia.

[10] https://www.isric.org/international-soil-standards
[11] https://www.ogc.org/
[12] https://help.hitachivantara.com/Documentation/Pentaho/7.1/0D0/Pentaho_Data_Integration

As primeiras *tasks* dos *jobs* são responsáveis pela conexão e ingestão de dados no *schema* conceitual do banco de dados da plataforma, as demais são responsáveis pelas validações e harmonizações segundo a SiBCS (por exemplo: validações de classes e outros atributos tais como relevo, descrição geral, observação, erosão, horizonte, propriedade química do horizonte, propriedade física do horizonte, propriedade morfológica, etc).

Um dos maiores desafios enfrentados na construção dos *workflows ETLH*, foi a inexistência no PDI de uma funcionalidade que realizasse automaticamente o relacionamento entre os registros dos *datasets* e as tabelas da plataforma *OpenSoils* (característica essencial no processo de FAIRificação de dados). Nesse sentido, utilizamos a *task* Execute SQL, um script capaz de executar qualquer declaração SQL diretamente no banco de dados. Assim, após a carga dos dados, os *scripts* SQL são executados para realizar as atualizações das chaves estrangeiras e manter o relacionamento entre os dados de um mesmo projeto.

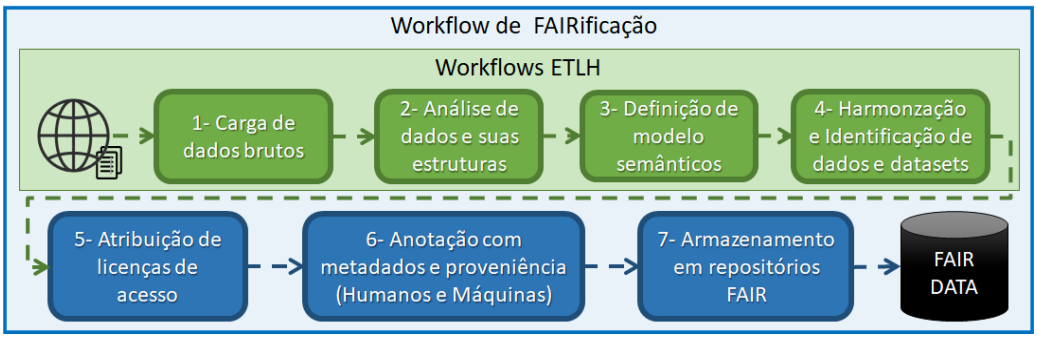

Figura 1. Representação conceitual das etapas do processo de *ETLH* e FAIRificação de dados.

A Figura 1, ilustra, o relacionamento entre os *workflows ETLH* e de FAIRificação de dados da plataforma *OpenSoils*. Em geral, a FAIRificação dos dados representa um processo de alto nível que pode ser materializado sob diversas etapas e tecnologias [Goble *et al*., 2020]. As principais etapas da FAIRificação são: 1) mapeamento e carga de *datasets (*dados brutos) disponíveis na Web; 2) análise dos dados para verificar conteúdo; conceitos representados, estruturas, as relações entre os elementos que constituem os dados; 3) definição/escolha de um modelo semântico para a representação do conjunto de dados (idealmente a partir de vocabulários ou ontologia bem fundamentadas), fornecendo uma estrutura para organizar/estruturar os dados sem ambiguidades; 4) permitir a identificação e harmonização dos dados, promovendo a interoperabilidade e a integração com outros tipos de dados e sistemas; 5) atribuir uma licença/autorização para acessar aos dados; 6) anotar os dados com metadados e proveniência permitindo que tanto seres humanos como máquinas possam localizá-lo; 7) promover o armazenamento (dados e metadados) a longo-prazo em repositórios FAIR e/ou publicar os dados FAIRificados adicionados de licença para que, os metadados possam ser indexados e localizados por mecanismos de pesquisa.

7. Resultados e Discussão

7.1 Experimentos

Os experimentos iniciais, foram mediados pelos *workflows ETLH* coletaram dados de perfis de solos dispersos em centenas de arquivos registrados no FeBR, além das tabelas do BDSolos. A execução dos experimentos modelados como um *job* e composto por *tasks*, processou 193 projetos pedológicos, totalizando mais 9 mil perfis de solos e mais de 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 projetos pedológicos realizados durante décadas em todo o Brasil, podendo ser integralmente acessados tanto pela plataforma *OpenSoils Web* e sua API quanto pelo app *OpenSoils* Edu [Cruz, 2018], já disponível na *Google Play Store* e na *App Store*.

7.2 Discussão sobre a FAIRificação

Até o momento, o banco de dados do *OpenSoils* adequa-se a pelo menos uma ou mais subdivisões de cada um dos princípios FAIR. Nesta seção pontuamos as adequações já existentes e as que ainda precisam ser concluídas em novos desdobramentos da pesquisa.

7.2.1. Localizável

Embora seja um requisito funcional da plataforma, os dados harmonizados ainda não estão plenamente linkados na Web de Dados, como na DBPedia. Apesar de identificáveis, por enquanto, encontram-se em banco de dados relacional. Entretanto, os metadados de descrição são padronizados e um dicionário de dados possibilita a compreensão de cada uma das tabelas e colunas que compõem o banco atualmente. Essa característica poderá ser explorada nas etapas do *workflow* de FAIRificação de dados.

7.2.2. Acessível

Os dados podem ser acessados pela plataforma Web ou pela plataforma móvel, como citado na seção 7.1. Podem ser acessados por outros sistemas abertos através de API. Os metadados possibilitam a construção de diversas formas de busca dos dados e se mantêm acessíveis constantemente, mesmo que, em algum momento, algum dado não se encontre mais disponível. Assim como os dados, os metadados possuem acesso facilitado pelas plataformas.

7.2.3. Interoperável

A partir da harmonização dos dados realizada pelos *workflows ETLH,* conseguimos atingir uma maior qualidade dos dados. Os dados encontram-se prontos para serem compartilhados e utilizados na representação do conhecimento. Os dados fazem referência a outros dados existentes na plataforma, integrando relacionamentos persistentes entre si.

7.2.4. Reutilizável

Desenvolvemos, junto aos *workflows ETLH*, um processo de registro e coleta de proveniência retrospectiva bem definidos apoiada na especificação PROV. Após a carga de novos *datasets* ou alteração de dados já existentes, uma tabela é atualizada com um novo registro que descreve quem, o que e como foi feita aquela operação, além de guardar o novo e o antigo valor. Um dos pontos a ser incorporado ao projeto é a anonimização

dos dados de usuários/projetos de forma a permitir sua reutilização, assegurando a privacidade e sob as normas definidas pela Lei Geral da Proteção de Dados Pessoais[13] (LGPD). No caso de necessidade de exportação para repositórios FAIR deverá proceder-se a técnicas de pseudo minimização, tais como a encriptação e utilização de códigos identificadores de documentos.

8. Conclusão

Podemos inferir que nesta era de constantes avanços da agricultura digital teremos grandes demandas relacionadas a gestão de dados devido a necessidade de reutilizar dados para desenvolver novos produtos ou serviços inovadores para atender os diversos atores da cadeia do agronegócio brasileiro. Os solos do Brasil são um elemento central nessa cadeia, porém, verificam-se poucas pesquisas sobre a gestão eficiente de longo prazo dos dados pedológicos.

Os atuais sistemas e repositórios apresentam gargalos relacionados à alta dispersão e baixa integração de dados, resultando em pouca transparência e informações limitadas de proveniência de dados. Essas condições reduzem a acessibilidade, interoperabilidade e reuso de dados pedológicos.

O estudo dos princípios FAIR em agricultura digital e humanidades digitais são recentes e implementados de forma tímida devido à resistência dos investigadores. Este artigo apresenta um esforço para compreender esse contexto e oferecer novos elementos que possam colaborar com a mitigação das inconformidades relacionadas às limitações dos *datasets* pedológicos e produzir dados de solos FAIRificados e harmonizados. Desenvolvemos *workflows ETLH* acopláveis à plataforma *OpenSoils* que, até o momento, se mostraram capazes em apoiar a FAIRificação segundo a taxonomia da SIBICS, anotando dados com metadados de proveniência retrospectiva e identificadores. Além disso, buscamos analisar as lacunas existentes para adequação dos dados aos princípios FAIR na plataforma *OpenSoils*.

A forma como os projetos de FAIRificação são conduzidos depende do orçamento disponível e do tipo e tamanho da organização. Nossos experimentos processaram centenas de milhares de registros de perfis de todos os tipos de solos em todo o Brasil. Na prática, esses dados já podem ser exportados ou acessados com maior praticidade e agilidade a partir dos aplicativos móveis ou da API do *OpenSoils*. Como trabalhos futuros, além da adequação integral aos princípios FAIR, pretendemos linkar os dados na Web de Dados e incorporar uma estrutura de *FAIR Digital Objects*, possibilitando assim a integração com diferentes bases de dados já disponíveis.

Referências

Bessa, A. C. F.; Cruz, S. M. S. (2021). Investigando a adoção de Princípios FAIR e Proveniência de Dados na Agricultura Digital sob a perspectiva das Humanidades Digitais: um estudo de caso na plataforma OpenSoils. In: HDRIO2021, 2021, Rio de Janeiro. II Congresso Internacional em Humanidades Digitais.

[13] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

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.

Da Cruz, S. M. S; Do Nascimento, J.A.P. (2016). SisGExp: Rethinking Long-Tail Agronomic Experiments. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). [S. l.: s. n.], vol. 9672, p. 214–217. https://doi.org/10.1007/978-3-319-40593-3_24.

Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M. R., ... & Schober, D. (2020). FAIR computational workflows. Data Intelligence, 2(1-2), 108-121.

Oliveira, S. S. C., de Barros Duarte, E., Marinho, E. C., & da Cruz, S. M. S. (2021, September). Integração de Data Lakes Pedológicos através de Workflows ETLH. In Anais da VII Escola Regional de Sistemas de Informação do Rio de Janeiro (pp. 48-55). SBC.

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.

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.

Simitsis, A. (2003, September). Modeling and managing ETL processes. In *VLDB PhD Workshop* (Vol. 76).

Veiga, V. S. de O., Henning, P., Dib, S., Penedo, E., Lima, J. D. C., Silva, L. O. B. da, & Pires, L. F. (2019). Plano de gestão de dados fair: uma proposta para a Fiocruz | Fair data management plan: a proposal for Fiocruz. Liinc Em Revista, 15(2). https://doi.org/10.18617/liinc.v15i2.5030

Wilkinson, M.; Dumontier, M.; Albersberg, I. *et al*. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18