UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

UMA INVESTIGAÇÃO SOBRE TRIPLIFICAÇÃO DE DADOS E SUA ADERÊNCIA AOS PRINCÍPIOS FAIR NO CONTEXTO TRANSDICIPLINAR DAS HUMANIDADES DIGITAIS E AGRICULTURA DIGITAL

SABRINA SANTOS CRUZ DE OLIVEIRA

RIO DE JANEIRO
2023

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

UMA INVESTIGAÇÃO SOBRE TRIPLIFICAÇÃO DE DADOS E SUA ADERÊNCIA AOS PRINCÍPIOS FAIR NO CONTEXTO TRANSDICIPLINAR DAS HUMANIDADES DIGITAIS E AGRICULTURA DIGITAL

SABRINA SANTOS CRUZ DE OLIVEIRA

Sob a orientação do Professor
SERGIO MANUEL SERRA DA CRUZ

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Humanidades Digitais no Programa de Pós-graduação Interdisciplinar em Humanidades Digitais, Área de Concentração em Humanidades Digitais.

Rio de Janeiro
2023

S48t Santos Cruz de Oliveira, Sabrina, 1995-

UMA INVESTIGAÇÃO SOBRE TRIPLIFICAÇÃO DE DADOS E SUA ADERÊNCIA AOS PRINCÍPIOS FAIR NO CONTEXTO TRANSDICIPLINAR DAS HUMANIDADES DIGITAIS E AGRICULTURA DIGITAL ——/ Sabrina Santos Cruz de Oliveira. - Rio de Janeiro, 2023.
102 f.: il.

Orientador: Sergio Serra Manuel da Cruz.
Dissertação (Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais, 2023.

1. Humanidades Digitais. 2. Web Semântica. 3. Dados Interligados. 4. Princípios FAIR. 5. Agricultura Digital. 6. Repositório de dados.
I. Serra Manuel da Cruz, Sergio, 1965-, orient. II Universidade Federal Rural do Rio de Janeiro. Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais III. Título.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

SABRINA SANTOS CRUZ DE OLIVEIRA

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Humanidades Digitais, no Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais.

DISSERTAÇÃO APROVADA EM: 29 de setembro de 2023

_____
Sérgio Manuel Serra da Cruz, Dr. – UFRRJ – Orientador

_____
Marcelo Panaro De Moraes Zamith, Dr. – UFRRJ – Avaliador Interno

_____
Mônica Ferreira da Silva, Drª. – UFRJ – Avaliadora Externa

_____
Rafael Elias De Lima Escalfoni, Dr. – CEFET – Avaliador Externo

ATA DE DEFESA DE TESE Nº 377/2023 - PPGIHD (11.39.00.16)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 15/03/2024 11:37)
MARCELO PANARO DE MORAES ZAMITH
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: ###810#1

(Assinado digitalmente em 15/03/2024 12:40)
SERGIO MANUEL SERRA DA CRUZ
PROFESSOR DO MAGISTERIO SUPERIOR
DCOMP (11.39.97)
Matrícula: ###24#6

(Assinado digitalmente em 08/07/2024 16:12) *MÔNICA FERREIRA DA SILVA*
ASSINANTE EXTERNO
CPF: ###.###.407-##

(Assinado digitalmente em 15/03/2024 16:21)
RAFAEL ELIAS DE LIMA ESCALFONI
ASSINANTE EXTERNO
CPF: ###.###.637-##

(Assinado digitalmente em 15/03/2024 11:40) *SABRINA SANTOS CRUZ DE OLIVEIRA*
DISCENTE
Matrícula: 2020######4

Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: ***377****, ano:* ***2023****, tipo:* ***ATA DE DEFESA DE TESE****, data de emissão:* ***15/03/2024*** *e o código de verificação:* ***e30e447d3f***

AGRADECIMENTOS

A presente dissertação de mestrado não teria sido possível sem o suporte e incentivo de algumas pessoas importantes.

Ao meu orientador, Professor Doutor Sérgio Manuel Serra da Cruz, por toda a paciência, empenho e sentido prático com que sempre me orientou neste trabalho, além de todo o apoio e confiança.

Também quero agradecer à Universidade Federal Rural do Rio de Janeiro e a todos os professores do meu curso pela elevada qualidade do ensino oferecido.

Quero agradecer também à minha família e amigos pelo apoio incondicional que me deram, especialmente aos meus pais Sandra e Waldir e minha avó Alice, que sempre fizeram de tudo para que eu pudesse alcançar meus objetivos.

A todos que direta ou indiretamente fizeram parte de minha formação, o meu muito obrigada.

RESUMO

OLIVEIRA, Sabrina Santos Cruz de. **Uma Investigação Sobre Triplificação De Dados E Sua Aderência Aos Princípios Fair No Contexto Transdiciplinar Das Humanidades Digitais E Agricultura Digital.** 2023. 102 p. Dissertação (Mestrado em Humanidades Digitais). Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, RJ. 2023.

Com a grande escalada na datificação de recursos e da digitalização de processos na era do *Big Data* e diante da incapacidade humana de manter, por si só, a qualidade, confiabilidade e segurança dos dados, emergem diversas oportunidades que propõem, a partir da utilização de processos, princípios e novas tecnologias, alternativas capazes de prover tratamentos mais adequados aos dados, bem como formas mais eficazes para a manipulação e a disponibilização dos dados científicos. Dentre elas, surgiram a *Web* Semântica e, mais recentemente, os Princípios FAIR, que promovem, dentre outros benefícios, o uso e o reuso de dados e/ou metadados e o alinhamento ao movimento *Linked Open Data* e aos princípios do *Linked Data.* Em resumo, tais aportes consistem na difusão de práticas de interligação de dados, bem como na disponibilização deles na *Web de dados*. Atualmente, o repositório de dados da plataforma *OpenSoils* possui uma grande quantidade de dados pedológicos que são de grande importância para diversas áreas de conhecimento tais como a agricultura digital, sustentabilidade, humanidades digitais, dentre outras; mas não se encontra plenamente conectado à *Web* de dados. Esse trabalho estuda as interseções dos conceitos por detrás dos movimentos *Web* Semântica, Princípios FAIR e *Linked Open Data* no âmbito das humanidades digitais. Construímos, a partir de uma ferramenta ETL, *workflows* ETLH capazes de captar os dados da base de dados relacional do *OpenSoils*, transformá-los e gerar como saída os mesmos dados em formato de triplas. Com o arquivo gerado a partir da execução dos *workflows*, criamos uma instância em um repositório de dados em grafos, onde foi possível realizar o relacionamento entre os nós com a execução de cláusulas, que também foram utilizadas para a realização de consultas que ilustraram os experimentos realizados em um dos projetos da base de dados.

Palavras-chave: Humanidades Digitais, *Web* Semântica, Dados Interligados, Princípios FAIR, Agricultura Digital, Repositório de dados.

ABSTRACT

OLIVEIRA, Sabrina Santos Cruz de. **Transdisciplinary research on Triplification of Pedological Data and its adherence to FAIR Principles in the context of Digital Humanities and Digital Agriculture.** 2023. 102 p. Dissertation (Master Science in Digital Humanities). Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, RJ. 2023.

In the context of the Big Data era, where resources are increasingly data centric and processes are digitized, there is a growing challenge in ensuring the quality, reliability, and security of data. Human efforts alone are insufficient to address these concerns. Consequently, various opportunities have arisen, leveraging processes, principles, and emerging technologies to offer more suitable data treatments and effective ways of managing and sharing scientific data. Two prominent approaches in this context are the Semantic Web and the more recent FAIR Data Principles. These frameworks not only advocate for the utilization and reusability of data and metadata but also align with the Linked Open Data movement and its underlying principles. In essence, they facilitate the dissemination of data interconnection practices and promote their availability on the data web. Presently, the data repository of the *OpenSoils* platform contains a substantial volume of pedological data with significant relevance across multiple knowledge domains, including digital agriculture, sustainability, and digital humanities. However, this repository is not yet fully integrated into the data web. This research explores the intersections of concepts related to the Semantic Web, FAIR Principles, and Linked Open Data movements within the digital humanities domain. Using an ETL tool, we built ETLH workflows capable of capturing data from the OpenSoils relational database, transforming it and generating the same data in triple format as output. With the file generated from the execution of the workflows, we created an instance in a graph data repository, where it was possible to create the relationship between the nodes with the execution of clauses, which were also used to carry out queries that illustrated the experiments carried out in one of the database projects.

Keywords: *Digital Humanities*, *Semantic Web*, *Linked Open Data*, *FAIR Principles*, *Digital Agriculture*, *Database*.

ABREVIAÇÕES E SIGLAS

AD	Agricultura Digital
AGRIS	*International Information System of the Agricultural Science and Technology*
ARN	AGRIS *Record Number*
BD2K	*Big Data to Knowledge*
DTL	*Dutch Techcentre for Life Sciences*
ETL	*Extract, Transform, Load*
FAIR	*Findable, Accessible, Interoperable, Re-usable*
FGV	Fundação Getúlio Vargas
GO-FAIR	*Global Open* FAIR
HD	Humanidades Digitais
HDRio	Congresso Internacional em Humanidades Digitais
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
LARHUD	Laboratório em Rede de Humanidades Digitais
LaViHD	Laboratório Virtual de Humanidades Digitais
LHuD	Laboratório de Humanidades Digitais
LOD	*Linked Open Data*
N3	*Notation*3
NIH	*American National Institutes for Health*
PDI	*Pentaho Data Integration*
PICS	*Platform for Internet Content Selection*
PPGIHD	Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais
R2RML	*Relational Databases to RDF Mapping Language*
RD	Repositórios Digitais
RDF	*Resource Description Framework*
RDFS	*Resource Description Framework Schema*
SCIELO	*Scientific Eletronic Library Online*
SGML	*Standard Generalized Markup Language*
SPARQL	*SPARQL Protocol RDF Query Language*

TADiRAH	*Taxonomy of Digital Research Activities in the Humanities*
UFRRJ	Universidade Federal Rural do Rio de Janeiro
URI	*Uniform Resource Identifier*
USP	Universidade de São Paulo
W3C	*World Wide Web Consortium*
XML	*eXtensible Markup Language*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Os meios de comunicação vêm evoluindo e se transformando ao longo dos anos, encurtando as distâncias entre os povos de culturas distintas e de certa forma acelerando a produção de conhecimento, a fusão e a disseminação de dados. Desde a produção dos primeiros dados digitais até chegarmos ao *Big Data* com sua geração grandes massas de dados e aos processos de datificação (STEIN et al., 2018), percorreu-se um sinuoso caminho que culminou com a recente intensificação dos processos de transformação digital (SOUTHERTON, 2020) e estratégias de centralização de poder centradas em dados (BRONSON e SENGERS, 2022).

A construção de novos conhecimentos vem sendo realizada de diversas formas ao longo da história, percebe-se que evoluem em velocidade compatível com que ocorre a disseminação das TIC e dos meios de processamento e armazenamento de dados. Hoje, de acordo com Espíndola et al (2017), a incontável e incontrolável quantidade de dados digitais gerados diariamente na área científica escancara a urgente necessidade da compreender a importância de sua governança e curadoria.

Observa-se fenômeno assemelhado na área de Agricultura Digital (AD) (MASSRUHÁ et al., 2020), onde fusão do Big Data e Computação no domínio da produção de alimentos e apresenta como uma nova fronteira a datificação. Para que se possa tirar o melhor proveito no que diz respeito tanto na geração de produtos e processos inovadores ou novos conhecimentos quanto para que seja mantida a segurança e a privacidade de quem gera e compartilha ou utiliza os dados nos processos de tomada de decisão e de divulgação científica na AD, a utilização de serviços computacionais inteligentes e repositórios digitais confiáveis é cada vez mais essencial (ATLAN, 2021).

Embora existam muitas as maneiras de organização de dados, neste trabalho, nós investigaremos o que chamamos de Repositórios Digitais (RDs) não relacionais transdisciplinares em busca de compreender como eles podem ser explorados nas áreas de Humanidades Digitais (HD) e AD. Conceitualmente temos que:

> "Os RDs são bases de dados online que reúnem de maneira organizada a produção científica de uma instituição ou área temática. Os RDs armazenam arquivos de diversos formatos. Ainda, resultam em uma série de benefícios tanto para os pesquisadores quanto às instituições ou sociedades científicas, proporcionam maior visibilidade aos resultados de pesquisas e possibilitam a preservação da memória

> científica de sua instituição. Os RDs podem ser institucionais ou temáticos. Os repositórios institucionais lidam com a produção científica de uma determinada instituição. Os repositórios temáticos com a produção científica de uma determinada área, sem limites institucionais" (IBICT, 2012).

Visando o aproveitamento e a (re)organização de dados para a obtenção de novos conhecimentos de qualidade com vistas a melhorar os processos analíticos e de tomada de decisão na AD e, com ênfase na área de solos, é necessário que sejam levadas em consideração a representação a semântica dos dados, a modelagem de dados e sua curadoria (CRUZ et al, 2019, DE OLIVEIRA et al. (2021ª), SUBIRATS-COLL et al, 2022).

Atualmente, coexistem vários padrões de modelagem de dados da Computação que também são aplicáveis em Humanidades Digitais (FLANDERS e JANNIDIS, 2018). Por exemplo, as representações relacionais, não relacionais e semânticos de dados. No especial interesse desta dissertação, busca-se a modelagem semântica baseada de dados em estruturas que suportem dados localizáveis, acessíveis, interoperáveis e reutilizáveis formato *Resource Description Framework* (RDF), RDF é um modelo padrão para intercâmbio de dados na *Web* Semântica (W3C, 2014).

> "O RDF estende a estrutura de vinculação da *Web* para usar URIs para nomear o relacionamento entre as coisas, bem como as duas extremidades do link (isso geralmente é chamado de "triplo"). Usando este modelo simples, ele permite que dados estruturados e semiestruturados sejam misturados, expostos e compartilhados entre diferentes aplicativos." (W3C, 2014).

Nesta pesquisa aprofundaremos os estudos iniciais de Bessa (2021), De Oliveira et al. (2021a) e Cruz e Bessa (2020), Marinho, Schmitz e Cruz (2022) e Marinho et al. (2023) onde se relacionam alguns os aspectos de gestão de dados ligados a projetos voltados para área de AD. Abordaremos, dentro do contexto das HD, a importância organizacional e representativa com semântica explícita dos conjuntos de dados de projetos ligadas a área de agricultura digital (AD), mais especificamente na área de solos, tendo como referenciais teóricos os Princípios FAIR (WILKINSON *et al*., 2016), RDs e a os protocolos e padrões da *Web* Semântica e da plataforma *OpenSoils* (Cruz et al, 2018).

Destacaremos e discutiremos a importância da aderência semântica explícita do *Linked Open Data* (BIZER, VIDAL & SKAF-MOLLI, 2018) *e* padrões de metadados no âmbito interdisciplinar das HD e a AD. Utilizaremos modelos inteligentes e suas semânticas, RDF e

linguagem SPARQL para propor e construir e testar uma infraestrutura de dados triplificados legíveis tanto por humanos quanto por máquinas.

Para contextualizar e apresentar os conceitos e temas necessários ao desenvolvimento desta dissertação, ela apresentará um estudo desse problema sob a ótica das Humanidades Digitais (HD) e que merece destaque e aprofundamento da reflexão crítica envolvendo um caso de estudo o RD de um recurso de Agricultura Digital, no caso o repositório relacional de dados da plataforma *OpenSoils* (CRUZ et al., 2019 e BESSA, 2021), que possui atualmente milhares de dados de perfis de solos com uma grande relevância para a área da AD e, em última análise para a própria sociedade brasileira, e que recentemente vem sendo alvo de pesquisas na área das Humanidades Digitais conforme investigações prévias conduzida por Bessa (2021), De Oliveira (2021a) e De Oliveira (2021b).

1.1. Motivação

Humanidades Digitais (HD) são uma área do conhecimento transdisciplinar e que ainda está se transformando e se consolidando na academia, teve seu início com Roberto Busa nos anos 1960 (BUSA, 1980). No entanto, por ser uma área relativamente nova, vem influenciando e sendo influenciada por diversas áreas do conhecimento, que por sua vez influencia a formação de novos pesquisadores e demanda novos olhares e interpretações (CUARTAS, 2017).

Segundo Svensson (2012), Cuartas (2017), Flanders e Jannidis (2018), Matos et al., (2019) e Burdick et al. (2020) as questões de pesquisa envolvendo dados em HD são muito amplas, incluem desde mineração de dados, ciência de dados, algoritmos de inteligência artificial, *design* e modelagem de dados, estudos de softwares e pesquisas em ciências humanas habilitadas por meio do digital; pesquisa baseada em computador e aplicativos de computador em estudos literários, linguísticos, culturais e históricos.

Neste trabalho defendemos que o leque das HD é amplo e ainda não foi suficientemente explorado pela comunidade acadêmica brasileira no que diz respeito a aspectos sociais e técnicos mais ligados ao cenário brasileiro digitalizado da agricultura tropical ou das diversas facetas do agronegócio nacional. Adicionalmente, percebemos que a inevitável marcha dos processos de transformação digital que se aplicam a todas as áreas do conhecimento e que podem ter relações ainda pouco exploradas ou ainda não plenamente compreendidas com diversos aspectos teóricos das HD aplicadas à AD.

Por exemplo, segundo BESSA (2021), De Oliveira et al., (2021a e 2021b), ainda existem poucas investigações que visam correlacionar as HD com os aspectos sociais de algumas áreas mais consolidadas como Engenharias, Saúde, Meio Ambiente ou mesmo as Ciências Agrárias. Alguns setores da agricultura e pecuária são considerados eminentemente "datacêntricos" e tem profundos impactos sociais, econômicos, ambientais e de sustentabilidade no atual contexto brasileiro. Por exemplo, irrigação, manejo integrado de pragas, rastreabilidade e certificação de produtos, gestão de propriedades, educação no campo estão passando por intensos processos de datificação e digitalização com incorporação de construtos tradicionais da Ciência da Computação por parte de grandes coporações ou mesmo de *startups*, por exemplo: inteligência artificial, aprendizado de máquina, automação, sensores, robótica, criptografia, *blockchain* para rastreabilidade, internet das coisas, entre outros.

Segundo Bolfe et al., (2020) e Bronson (2022), as necessidades de automação e informatização dos processos ligados à agricultura (tradicional ou digital) e as crescentes demandas das pessoas e corporações que sejam habilitadas a atuar na cadeia produtiva do agronegócio vem pressionando diversos setores das sociedades, em diversos países e no Brasil não é diferente.

A agricultura brasileira é pujante e vem passando por grandes transformações tecnológicas, econômicas, sociais e ambientais nas últimas décadas[1]. Nas zonas rurais, por exemplo, a agricultura familiar, embora responsável por mais de 77% da produção agrícola nacional (IBGE, 2017), tem sofrido diversos tipos de problemas e precisa incorporar diversos tipos de transformações devido a múltiplos fatores, entre eles o avanço da mecanização (informatização), evasão do homem do campo e o não letramento digital de parte da população rural (BOLFE et al. 2020). O autor descreve que as dificuldades de se ampliar a conectividade no meio rural como um dos maiores desafios para a inclusão completa da agricultura e dos produtores no processo de transformação digital.

O fenômeno da digitalização de alguns setores do meio rural que privilegia aspectos quantitativos e tem necessidades de processamento de grandes volumes de dados digitais, cujo resultados podem ter potencial gerador de inovação nos diferentes elos na pré-produção, produção e pós-produção das cadeias produtivas (BOLFE et al., 2020). Devido a amplitude e múltiplas facetas da agricultura tropical brasileira, neste trabalho teremos um recorte,

[1] https://www.embrapa.br/en/visao/trajetoria-da-agricultura-brasileira

destacaremos a relevância de avaliar de modo transdisciplinar um dos mais importantes recursos da cadeia do agronegócio brasileiro: o solo e seus *datasets*.

De acordo nossas análises iniciais (De Oliveira et al, 2021a e De Oliveira et al, 2021b, os processos de digitalização abrem a possibilidade de novos olhares e o desenvolvimento de estudos transdisciplinares em HD e AD, tanto de um ponto vista mais focado (envolvendo análises de dados) quanto de pontos de vistas mais amplos envolvendo aspectos organizacionais (governança de dados, curadoria de dados e gestão de metadados) envolvidos no cruzamento desses dois domínios do conhecimento.

Nesse trabalho proporemos uma investigação sobre a (re)organização de alguns *datasets* usados na área de pedologia na plataforma *OpenSoils*. Consideraremos apenas o recurso dados de solos, avaliando como o arcabouço conceitual do movimento *Linked Open Data* (LOD) e os princípios FAIR podem ser incorporados à nossa pesquisa. Adicionalmente, desenvolveremos um arcabouço e experimentos computacionais que serão realizados a partir do repositório da plataforma do *OpenSoils* (CRUZ et al., 2019) com a finalidade de avaliar a proposta e disponibilizar dados linkados e legíveis por máquina para as comunidades agronômicas e afins.

Por fim, destacamos que esta dissertação se alinha com pesquisas e ações institucionais que já estão em andamento, como por exemplo o projeto *OpenSoils,* financiado pelo CNPq (Códigos 315399/2018-0 e 400044/2023-4) e, mais recentemente, com os esforços da Rede GO-FAIR Agro Brasil[2] (DE OLIVEIRA et al., 2021b).

1.2. Justificativa

A datificação não deve ser encarada apenas como a produção de grandes volumes dados, o que, em certo sentido, o ser humano vem fazendo desde a criação dos símbolos e da escrita. Pelo contrário, a datificação é um fenômeno mais contemporâneo e transdisciplinar que se refere à quantificação da vida humana por meio de informações digitais, muitas vezes considerando valores sociais e econômicos. Bronson (2022) discute este processo na área da agricultura, a autora informa que ele tem imensas consequências nas sociedades atuais pois é intensamente explorado por grandes corporações produtores de equipamentos, fertilizantes, produtos químicos e serviços. Segundo Mejias e Couldry (2019), disciplinas como economia

[2] https://www.go-fair-brasil.org/agro

política, ciência de dados, softwares, teorias jurídicas e HD consideraram importantes aspectos dessas consequências.

Segundo Southerton (2020), o processo de datificação refere-se ao processo pelo qual sujeitos, objetos e práticas são transformados em dados digitais. O processo está intimamente associado à ascensão das TIC e ao *Big Data*. Segundo o autor, muitos estudiosos argumentam que a datificação está se intensificando à medida que mais dimensões da vida social se desenrolam nos espaços digitais. A datificação processa uma gama diversificada de informações como dados quantificáveis e legíveis por máquina para fins de análise. A datificação também é usada como um termo para descrever uma lógica que vê as coisas no mundo como fontes de dados a serem "extraídas" ou "correlacionadas" ou "agregadas" ou "enriquecidas", e a partir das quais *insights* podem ser obtidos sobre o comportamento humano e sobre questões sociais.

Segata e Rifiotios (2021) consideram que a datificação e a digitalização são faces de um amplo e pervasivo processo de modelagem que visa reduzir a complexidade e as contingências da vida a meros códigos e dados. Por outro lado, Blanke e Prescott (2016) afirmam que a datificação tem um significado particular para as HD, pois "sugere que, para o trabalho de *Big Data*, dados não estruturados não são suficientes".

Nossa perspectiva de trabalho se apoia nos estudos de Blanke e Prescott (2016) e se alinha com Hawkins (2021) e Bronson (2022), que considera que quando socialmente engajada a agricultura digital e seus arquivos provenientes de processos de digitalização podem oferecer muitas novas oportunidades tanto para os humanistas digitais quanto para os produtores fazerem uso inteligente de grande quantidade de arquivos digitalizados.

Neste trabalho, pretendemos ir um pouco além desses autores, consideramos que dados digitais de HD e AD são heterogêneos, pouco documentados, muitas vezes estão distribuídos e fragmentados em *data lakes* e raramente estão semanticamente anotados com metadados e em acordo com os Princípios FAIR (WILKINSON et al. 2016). Acreditamos que ao investigar essa perspectiva traremos ao mesmo tempo um amadurecimento de pesquisas em uma área ainda pouco explorada onde grandes desafios se colocam tanto no domínio das HD quanto da AD ao explorá-las com as tecnologias da *Web* Semântica.

Operacionalmente, a AD e seus dados digitais são interdependentes (CAPMOURTERES et al., 2018, SHEPHERD et al., 2020 & BIRNER et al., 2021). Bronson

(2022), traz uma importante discussão sobre os dados agrícolas. O volume de dados digitais agrícolas cresceu exponencialmente, grande parte produzida dentro das propriedades por dispositivos operando nos campos (tratores, drones e máquinas agrícolas inteligentes), por sensores (tanto de sensoriamento remoto quanto da Internet de Coisas) e aplicativos utilizados na gestão da cadeia de produção. Esses elementos, em linhas gerais, são exemplos de processos de digitalização que resultam em enormes volumes de arquivos onde os (meta)dados nem sempre estão adequadamente presentes ou sequer linkados, compartilhados ou mesmo acessíveis para os produtores. No entanto, os dados digitais agrícolas e seus metadados são um ativo digital valiosíssimo e grande interesse das corporações produtores de equipamentos e distribuidoras de serviços agrícolas, seus usos trazem grandes impactos éticos, ambientais, sociais e econômicos tanto para os ambientes rurais e urbanos.

O *Linked Open Data* (LOD) fornece um arcabouço conceitual importante nesta pesquisa. Aqui se defende que os avanços da Inteligência Artificial (IA) nas HD e na AD (em especial na área de solos) é um processo que está em marcha e é inevitável. No entanto, os construtos de IA demandarão *datasets* cada vez de melhor qualidade. Logo, para que se obtenham esses recursos é essencial criar dados de solos linkados em arquivos legíveis por máquina, interoperáveis, reutilizáveis e extensíveis adequados para interrogação e análise usando métodos de pesquisa em Computação e HD (SILVA et al., 2018). Usando LOD, dados de arquivo (dados de catálogo, metadados, dados extraídos do conteúdo de arquivos nascidos digitais e digitalizados) podem ser facilmente incorporados na *Web*, enriquecendo e contextualizando ainda mais os dados de arquivo e tornando mais fácil descobrir, acessar e utilizar.

Destacamos que o LOD não é nenhuma novidade e nem mesmo uma exclusividade da área de computação ou das HD ou das AD, trata-se de um movimento global, consistente e crescente iniciado por Berners-Lee em 2006 (HEATH & BIZER, 2011). Ninin (2018) acredita que os princípios *Linked Data*, em consonância com os objetivos maiores da proposta da *Web Semântica* e com os princípios *Open Data*, culminaram nos LOD.

A adoção e os estudos do LOD na área de HD possibilitam a construção de bases com maior representatividade de toda variabilidade encontrada na prática para a produção de modelos que sustentarão ferramentas e pesquisas mais confiáveis e robustas. Por outro lado, as implicações do LOD em HD não se resumem apenas à área acadêmica, uma vez que (cada vez mais) os dados se fazem essenciais para absolutamente todas as áreas.

Sendo o LOD inerentemente interoperável, ele tem potencial para desempenhar um papel fundamental na implementação dos princípios FAIR que fornecem indícios de serem um meio viável de disponibilizar *datasets* de maior qualidade, criando arquivo legíveis por humanos e máquina adequados para análises usando métodos de pesquisa em humanidades digitais.

Hawkins (2021) destaca que embora um corpo crescente de estudos e práxis de arquivamento tenha explorado o LOD, seu potencial para abrir arquivos digitalizados e nascidos digitais para as HD é sub examinado. Logo, para tornar possível o acesso mais aberto a dados de pesquisa de qualidade, faz-se necessária a publicação de tais dados seguindo princípios que orientam a adoção de uma estrutura semântica legível por humanos e máquinas.

Nesse sentido, este estudo tem como justificativa identificar, sob o olhar das HD, a possibilidade de publicar dados de pesquisa oriundos da área de AD seguindo os princípios FAIR a partir da adoção das tecnologias da *Web* Semântica, por meio do LOD.

1.3. Problema

A digitalização em massa e o crescimento exponencial de arquivos nascidos digitais nas últimas décadas resultaram em um enorme volume de arquivos e dados (*datasets*) disponíveis digitalmente em RD. Isso produz fontes valiosas, mas subutilizadas, de dados digitais, prontos para serem interrogados por estudiosos e profissionais das HD. No entanto, as abordagens de digitalização atuais ficam aquém dos requisitos dos humanistas digitais para dados estruturados, integrados, interoperáveis e interrogáveis. *Linked Open Data* fornece um caminho viável de produzir tais dados, criando dados de arquivo legíveis por humanos e máquinas adequados para análise usando métodos de pesquisa de HD.

O número de RD que usam o protocolo LOD para apoiar sistemas inteligentes em HD e AD ainda pode ser modesto em comparação com outras áreas. No entanto, independentemente de usar LOD ou não, é crucial que se assegure que os RD sejam projetados especificamente para a *web* e para reutilização por humanos e máquinas (NEUROHACKWEEK & REPRONIM, 2016). Adicionalmente, em algumas áreas se observa a falta de repositórios específicos (por exemplo, Ciências Da Terra) ou mesmo falta de repositórios facilmente integráveis que possam lidar com contextos complexos (Humanidades Digitais) para localizar, integrar e reutilizar grandes dados com semânticas explícitas.

Um dos desafios enfrentados tanto pelas HD quanto pela AD consiste na construção de RD para arquivamento de dados abertos ligados confiáveis e ao mesmo promover a curadoria dessas informações. Isso porque a representatividade da variabilidade necessária depende de uma grande quantidade de dados e seus metadados, o que consequentemente demanda ações de curadoria, que se torna inviável para os pesquisadores que nem sempre possuem tempo ou recursos para desempenhá-la adequadamente (BARBEDO, 2018, BESSA, 2021).

Para superar essas limitações, uma das soluções encontradas mais utilizadas hoje e que provavelmente se manterá no futuro é investigar novas técnicas que ampliem o compartilhamento e o reuso dos repositórios e dos seus dados para apoiar ações semiautomatizadas de curadoria de dados. Entretanto, se considerarmos somente o compartilhamento de dados ela por si só traz desafios importantes, visto que os dados de solos geralmente apresentam problemas estruturais e semânticos e de governança (DE OLIVEIRA et al, 2021).

Além disso, há a necessidade de ampliar a acessibilidade e reusabilidade desses *datasets* com a adequação aos princípios FAIR (DE OLIVEIRA et al, 2021) e possibilitar a interligação mais facilitada com a *Web* de Dados e consequentemente com outros RD. Um outro problema diz respeito a natureza dos aparatos computacionais necessários para fazê-lo, este é um dos desafios adicionais quando tratamos do compartilhamento de dados e é onde consiste a contribuição deste trabalho com vista a tratar a problemática estudada nesta dissertação.

1.4. Objetivo geral

O objetivo geral desta dissertação consiste desenvolver uma investigação em HD sobre a aplicação dos conceitos do movimento *Linked Open Data* e Princípios FAIR para elaborar artefatos computacionais inteligentes voltados para o tratamento de grandes volumes dados pedológicos e criação de um repositório de dados triplificados e aberto para a plataforma *OpenSoils*.

1.4.1. Objetivos específicos

Os objetivos específicos desta pesquisa são:

1- Compreender a aplicação dos conceitos do movimento *Linked Open Data* dentro do contexto de digitalização e enriquecimentos semântico de dados na interseção entre as HD e AD;

2- Para os nossos experimentos, considerar o repositório de dados da plataforma *OpenSoil*s (CRUZ et al., 2019) como base para a execução dos experimentos baseados em *workflows* de triplificação de dados (AUER, 2009) para a aplicação dos conceitos do LOD e Princípios FAIR discutidos nesta dissertação.
3- Realizar na plataforma experimentos, de maneira a aplicar os conhecimentos teóricos discutidos durante o desenvolvimento desta dissertação, disponibilizando os resultados/produtos obtidos no repositório OpensoilsGraph do GitHub.
4- Apresentação e publicação de trabalhos em congressos ou eventos e submissão de textos para revistas indexadas a partir dos conhecimentos e análises obtidas durante o desenvolvimento da presente dissertação.
5- Registrar os produtos no Instituto Nacional de Propriedade Intelectual (INPI) como garantia da obtenção da propriedade intelectual dos produtos gerados.

1.5. Organização da Dissertação

A dissertação está organizada em capítulos. No primeiro capítulo através da Introdução, são apresentados os seguintes tópicos: motivação, justificativa, o problema, os objetivos gerais e específicos e a organização da qualificação. No segundo capítulo, é apresentado o referencial teórico da pesquisa; já o terceiro capítulo apresenta a metodologia da pesquisa adotada neste trabalho; o quarto capítulo aborda a infraestrutura de triplificação de dados de solos; o quinto capítulo traz os experimentos e discussões; o sexto e último apresenta a conclusão, bem como as limitações encontradas, as possibilidades de trabalhos futuros e as publicações realizadas durante o desenvolvimento deste trabalho; e, por fim, são disponibilizadas as referências bibliográficas.

2. REFERENCIAIS TEÓRICOS

Neste capítulo apresentaremos os principais temas correlacionados com essa pesquisa, os quais trazem conceitos necessários à compreensão de como será feita a abordagem metodológica para a encaminhar uma solução ao problema que esta dissertação pretende investigar.

2.1. Humanidades Digitais

De acordo com a literatura da área, a aplicação da computação nas Ciências Humanas teve seu início pelo padre jesuíta Roberto Busa nos anos 1960 (BUSA, 1980). Segundo Hockey (2007), Busa foi responsável por desenvolver um *index verborum* com todas as palavras contidas nas obras de São Tomás de Aquino, consistindo em mais de 11 milhões de palavras em Latim e oriundas da época medieval. Tal iniciativa resultou no primeiro prêmio da área e no reconhecimento, pela primeira vez, da aplicação da tecnologia em uma pesquisa de humanidades.

O termo Humanidades Digitais é resultado da necessidade de uma abrangência maior de escopo que anteriormente se dividia entre diferentes conceitos, como por exemplo a Computação para as Humanidades, Linguística Computacional, entre outros (TERRAS, 2014). Consiste na ligação das investigações das Humanidades e a utilização de métodos e ferramentas das tecnologias digitais (ALVES, 2016). Alguns autores consideram as HD se referem:

> "a novos modos de produção acadêmica e de unidades institucionais para a pesquisa, ensino e publicações colaborativas, transdisciplinares e permeadas pelas tecnologias computacionais. Elas são menos um campo unificado do que um feixe de práticas convergentes que exploram um universo no qual a mídia impressa não é mais o meio primário no qual o conhecimento é produzido e disseminado" (Burdick et al., 2020).

Muito se discutiu sobre uma renomeação do que anteriormente se denominava Humanidades Computacionais. Segundo Patrik (2016), a adoção do termo Humanidades Digitais tirou a computação de um status de apenas suporte dentro da área para se tornar algo necessário até mesmo para se levantar novas questões dentro das Ciências Humanas.

No Manifesto 2.0 das Humanidades Digitais, Schnapp et al. (2009) divide a história das Humanidades Digitais em ondas. Segundo o autor, a primeira onda de trabalho nas Humanidades Digitais foi quantitativa e mobilizou o poder de buscar e recuperar bancos de dados. Desta maneira, foi possível, por exemplo, automatizar a linguística de corpus.

A segunda onda, por outro lado, é qualitativa. Além disso, Schnapp et al. (2009) destaca que esta onda é também interpretativa, experiencial, emotiva e generativa em caráter, já que se utiliza de ferramentas digitais a serviço das forças metodológicas centrais das Humanidades, como o contexto histórico, a profundidade analítica, especificidade do meio, atenção à complexidade, a crítica e a interpretação.

Berry (2012) vislumbrou um caminho para a terceira onda das Humanidades Digitais. Segundo o autor, nesta onda a tecnologia digital destaca anomalias geradas em um projeto de pesquisa e leva à questionamentos de pressupostos implícitos nesta pesquisa. Na terceira onda as Humanidades Digitais testemunharam uma virada crítica com o estabelecimento de referenciais metodológicos para o campo, bem como uma dimensão interpretativa mais aprofundada que torna os projetos de HD mais conscientes dos aspectos culturais e políticos (HELMI, 2021).

A definição canônica do campo das Humanidades Digitais ainda é alvo de intensos debates. É uma área considerada democrática, interdisciplinar e não há unanimidade quanto a sua delimitação. Ramsay (2016) traz à tona o debate sobre a importância ou não do domínio tecnológico pelos profissionais da área. Koh (2014), por sua vez, explicitou a importância do contrato social da comunidade, prezando pela igualdade e não hierarquização, além de nos levar a refletir sobre as desigualdades relacionadas à raça, gênero e classe que não são levadas em consideração no contrato, herdado das Humanidades Computacionais.

Apesar da definição do termo Humanidades Digitais possuir pouco mais de uma década, no Brasil algumas iniciativas já se destacam. Criado em 2016, o Laboratório de Humanidades Digitais (LhuD) da Fundação Getúlio Vargas (FGV) contempla hoje importantes linhas de pesquisa, que consistem na literacia e acervos digitais, tecnologias textuais e de análise de som, imagem e vídeo. O Laboratório Virtual de Humanidades Digitais (LaViHD) é fruto do Grupo de Pesquisas em Humanidades Digitais da Universidade de São Paulo (USP) e desenvolve ambientes de apoio e ferramentas para acervos digitais e corpora eletrônicos.

O Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais[3] (PPGIHD) da Universidade Federal Rural do Rio de Janeiro (UFRRJ) teve início em 2019, sendo o primeiro do estado do Rio de Janeiro e um dos primeiros programas de pós-graduação em HD do Brasil. Com linhas de pesquisa de análises qualitativas e quantitativas de dinâmicas sociais,

[3] https://www.dcc.ufrrj.br/ppgihd/

métodos computacionais em políticas públicas e mineração e dados digitais, o programa já desenvolve importantes projetos dentro da área, como a implantação do Sistema Integrado de Centros de Documentação Histórica da UFRRJ, responsável pelo armazenamento digital dos acervos históricos da universidade.

Embora a digitalização de acervos seja, há algum tempo, uma área das principais facetas das HD com maior número de iniciativas, a preocupação com a interligação de dados de forma que possibilite cada vez mais o reuso, compartilhamento e a colaboração ainda é mais recente, mas já tem ganhado espaço. Alguns conjuntos de dados de humanidades são difíceis de expressar digitalmente, isso porque estão dispersos na *Web* e possuem uma grande diversidade de formatos, poucos metadados além de não estarem interligados à *Web* de Dados na maioria das vezes (MEROÑO-PEÑUELA, 2017).

O campo das HD já possui algumas iniciativas importantes na direção da transformação desses conjuntos de dados de forma que seja possível sua interligação aos conjuntos já disponíveis na *Web* de dados. Por exemplo, a *Taxonomy of Digital Research Activities in the Humanities* (TADiRAH) é uma dessas contribuições. Desenvolvida pelo Laboratório em Rede de Humanidades Digitais[4] (LARHUD) do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), trata-se de uma taxonomia desenvolvida para a compreensão das HD. Nela, são identificados objetos, métodos e técnicas e atualmente abrange, senão todos, a grande maioria da estrutura das humanidades digitais na classificação e fundamentação das atividades de pesquisa digital da área. Qualificando-se, portanto, como o principal expoente taxonômico para a área (LARHUB. *Website*).

A TADiRAH[5] surge em um momento em que o movimento Ciência Aberta demanda por colaboração e compartilhamento de produções. Além disso, a categorização e a classificação das atividades das HDs sempre foi uma área de interesse pelos profissionais deste campo de pesquisa, bem como as contínuas tentativas de definição das HDs (BOREK, 2017).

[4] http://www.larhud.ibict.br/

[5] https://vocabularyserver.com/tadirah/pt/index.php

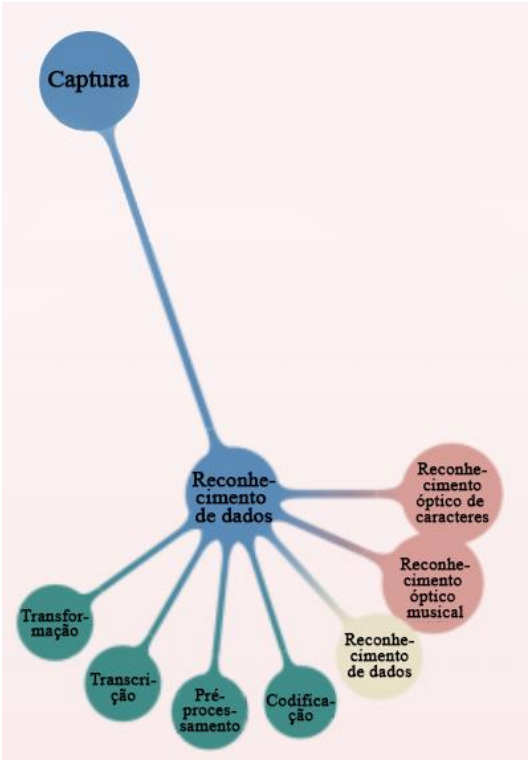


Figura 1 – Conceito "Captura" e sub conceito "Reconhecimento de dados" com seus conceitos restritos (LARHUB, website, trad. Nossa).

A estrutura do TADiRAH consiste em conceitos amplos (conceitos), de onde outros termos são derivados (sub conceitos). Dos sub conceitos também se derivam termos mais restritos, como podemos observar na Figura 1, onde o conceito "Captura" é representado, localizado no centro do mapa. Dele, oito termos são derivados como sub conceitos, são eles: gravação, transcrição, atividades de pesquisa, conversão, reconhecimento de dados, descoberta, coleta, geração de imagens, entretanto, na imagem estamos destacando apenas o sub conceito "reconhecimento de dados" e seus conceitos.

Além do objetivo de criar um recurso amplamente disponível e aplicável a diversos conceitos, fazia-se necessária uma versão legível por máquinas baseada em padrões (BOREK, 2017). Para isso, foi utilizada uma instância do *TemaTres Vocabulary Server* (FERREYRA, 2014), hospedado pela bibliografia DARIAH-DE: o núcleo SKOS, fazendo a TADiRAH disponível como dados abertos vinculados (BOREK, 2017), tema que abordaremos no tópico 2.5 deste trabalho.

Tendo em vista os processos de datificação e digitalização tão comuns ao século XXI e seus consequentemente impactos nas HD, Sorbara (2020) destaca que mudar o foco da simples publicação de dados para a análise de dados nas HD demandará que as novas pesquisas científicas enfrentem novos desafios até agora pouco discutidas nas HD, tais como como a

extração de conhecimento, a visualização agregada de dados, aprendizado de máquina e descoberta de conhecimento, entre outros. O autor pontua que novas abordagens e ferramentas com o poder de controlar, detalhar e explicitar os vocabulários e semânticas das HD podem atuar com facilitadoras para as novas pesquisas e mesmo democratizador os conhecimentos englobados pelas HD.

2.2. Agricultura Digital

A prática da agricultura surgiu cerca de 12 mil anos atrás durante o período neolítico, sendo um dos processos constitutivos das primeiras civilizações. Recentemente, a agricultura teve seus impulsos tecnológicos essenciais para sua evolução durante os séculos XIX e XX, quando surgiu então a fase que denominamos de Agricultura 1.0. A força de trabalho era exercida principalmente pela mão de obra familiar e a tecnologia mais utilizada era a tração animal nas diversas atividades realizadas no campo (MASSRUHÁ et al., 2020).

Com o advento da 1ª Revolução Industrial, o crescimento da população e consequentemente o aumento da demanda por alimentos, os processos agrícolas precisaram evoluir. Também por estes motivos, a mecanização no campo tornou-se uma tendência nos fins do século XIX e início do século XX, porém foi apenas o período que sucedeu o fim da Segunda Guerra Mundial que culminou a ampla substituição da tração animal pela força mecânica na América do Norte e Europa, dando início, portanto, a Agricultura 2.0 (MASSRUHÁ et al., 2020).

Massruhá et al. (2020) afirmam que o fortalecimento da intensificação agrícola, o investimento em pesquisa, políticas públicas e empreendedorismo; e a consequente evolução tecnológica, com máquinas e implementos desenvolvidos para uma maior eficiência em uma tendência conhecida como agricultura de precisão, culminaram na Agricultura 3.0. No Brasil, neste mesmo período, mais precisamente em 1973, a Embrapa foi criada tendo como uma das suas principais atribuições a garantia da segurança alimentar. Nossa agricultura era majoritariamente baseada na monocultura naquele momento e éramos um grande importador de alimentos (MASSRUHÁ et al., 2020).

A Agricultura 4.0, mais conhecida como Agricultura Digital (AD), é considerada a última grande revolução agrícola ainda em curso, “incorporando a conectividade e automação, o uso de máquinas, veículos, drones, robôs e animais com sensores” (ESPERIDIÃO et al. 2019). A revolução digital tem levado a uma infinidade de sistemas, sites e aplicativos móveis

que agora estão disponíveis para auxiliar o agricultor, agrônomo, investidor do agronegócio, na tomada de decisão (ALI & DAHLHAUS, 2022). Análoga à Indústria 4.0, a Agricultura 4.0 “é resultado da transformação digital do setor agrícola por meio da coleta massiva de dados para ajudar na tomada de decisão” (MASSRUHÁ et al., 2020).

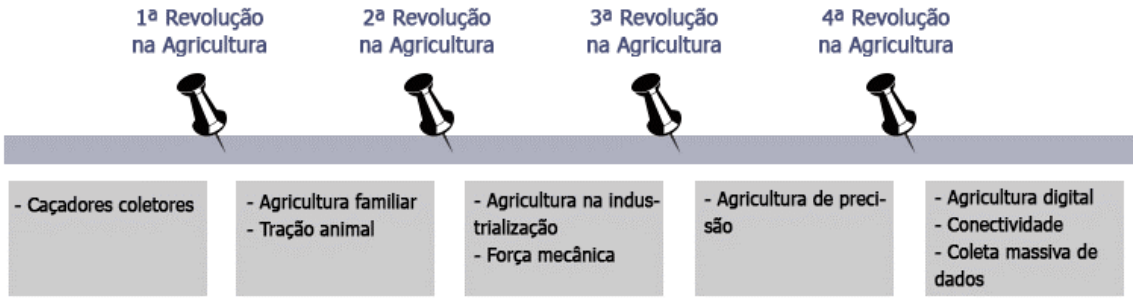


Figura 2 – As revoluções na agricultura (Elaborado pela autora).

As tecnologias digitais presentes na AD despontam como importantes facilitadoras na otimização dos processos da cadeia do agronegócio, bem como possíveis aceleradoras para o alcance de metas de sustentabilidade, para a melhoria da qualidade de vida dos trabalhadores e da população rural, além de atrair gerações mais jovens para a agricultura (BOLFE et al., 2020).

A inovação advinda da AD bem como as inserções da tecnologia no meio agrícola possibilitou a ampliação da conectividade máquina x homem x tecnologia, com resultados significativos no aumento da produtividade/rentabilidade (ESPERIDIÃO et al. 2019). Entretanto, alguns novos desafios passaram a ser observados nesta nova fase. Desafios estes que vão desde questões socioeconômicas e educacionais passando pela demanda de gestão e o monitoramento das produções vegetal e animal, até a governança de dados com a criação, manutenção e acessos as bases de dados.

Figura 3 – Fluxo de inovação tecnológico (ESPERIDIÃO et al. 2019).

A transformação digital é um outro vetor de transformação que é capaz de reduzir algumas dificuldades do homem do campo e propiciar novos valores a cadeia do agronegócio. O valor agregado na transformação digital compreende a produção, novos produtos e clientes, promovendo uma integração das tecnologias no campo (FRANÇA, 2019).

> "A aplicação da Transformação Digital no campo da agricultura moderna é capaz de permitir que as tecnologias sejam acopladas em busca de estratégias para o processo produtivo e para o crescimento do negócio. O objetivo é acelerar a produtividade rural, construir um país voltado para tecnologias limpas e harmoniosas, reduzir a distância entre os países produtores e propiciar a construção de um novo conceito de campo" (FRANÇA, 2019).

O aumento do acesso à educação de qualidade, letramento digital, e a conectividade no meio rural são os maiores desafios para a inclusão completa da agricultura no processo de transformação digital mesmo com os constantes e crescentes investimentos dos setores público e privado. Além disso, as iniciativas desses setores também buscam elevar a capacitação dos produtores e trabalhadores rurais em temas correlatos à AD, o que também serve como um estímulo para a permanência de jovens nas zonas rurais e ao surgimento de novas empresas denominadas *startups* que procuram soluções tecnológicas para o agronegócio, pensando principalmente no *modus operandi* do mercado rural de base agrotecnológica, essas empresas também são denominadas de *AgTechs*[6] (BOLFE et al., 2020).

[6] https://articlegateway.com/index.php/JHETP/article/view/3845/3657

No Brasil, o movimento de inovação no agronegócio e *AgTechs* foi particularmente intenso mesmo durante a pandemia de COVID-19. Por exemplo, em 2021 o país identificou 1.125 *Agtechs*, sendo 449 *startups* a mais em relação à 2020, ou seja, praticamente uma nova *Agtech* por dia (AGRITECH, 2021).

Os desafios de cunho eminentemente digital envolvem, por exemplo, sistemas inteligentes de nova geração e em especial de RD agronômicos; estes consistem, basicamente, na modelagem e construção de bases que possuam a representatividade de toda variabilidade encontrada na prática para a produção que sustentarão sistemas confiáveis e robustos (BARBEDO, 2018). RD agronômicos, geralmente, implicam na coleta de um número muito grande de amostras bem como a organização das informações complementares sobre o que está representando cada amostra, como local e data de coleta (BARBEDO, 2018). Chamamos estas informações complementares de metadados (VELLUCCI, 1998) e abordamos mais deste tópico na seção 2.3 deste trabalho.

No entanto, coletar e manter números elevados de amostras ao longo de muitos anos por um único grupo de pesquisa torna-se pouco viável na grande maioria das vezes (BARBEDO, 2019). Para contornar estas circunstâncias, existem duas alternativas que vêm sendo aplicadas: Ciência Cidadã (SILVERTOWN, 2009 & IRWIN, 2002) e compartilhamento e curadoria de bases de dados (BOLFE et al., 2020).

> "A ciência cidadã faz uso de voluntários não profissionais para coletar dados como parte de pesquisa científica, particularmente em ecologia e ciências ambientais. No caso da detecção de doenças em plantas, por exemplo, produtores e trabalhadores rurais poderiam coletar imagens de sintomas em campo e, após serem enviadas a um servidor, tais imagens poderiam ser rotuladas por fitopatologistas. À medida que dispositivos móveis com capacidade de imageamento tornam-se ubíquos, o desafio será encontrar mecanismos para promover a participação de voluntários" (BARBEDO, 2019).

Neste ponto vale destacar que do ponto de vista conceitual, as bases de dados tradicionais ou os novos RD no cenário agrícola possuem a mesma importância que os tradicionais acervos digitais das HD. A curadoria e o compartilhamento das bases de dados é, provavelmente, a alternativa que prevalecerá no futuro e é a que este trabalho se utilizará para oferecer sua contribuição.

Assim como BOLFE et al., (2020), acreditamos que ao disponibilizar e integrar as bases de dados geradas pelos diversos grupos de pesquisas, possivelmente alcançaremos um conjunto

de dados mais representativo desse setor. Além disso, defendemos que a adoção dos princípios FAIR (WILKINSON et al., 2016) nas Ciências Agrícolas, que discutiremos melhor na seção 2.8 deste trabalho, também citada por Bolfe (2020), é um passo complementar para a adição de valor às bases de dados da AD.

Segundo Ali e Dahlhaus (2022) e Jonquet et al. (2018) através de sua vasta literatura[7] que envolve atividade multidisciplinares em Computação, *Web* Semântica e Agronomia, se constata que os dados agronômicos são dados de difícil integração e interoperabilidade técnica pois sua semântica é complexa. Os autores defendem que os dados agrícolas e metadados devem ser transmitidos usando estruturas de dados sintáticas e semânticas cobrindo os dados brutos.

Destacamos que os dados de solos, um dos principais recursos da cadeia do agronegócio, apresentam as mesmas características e limitações supramencionadas. Segundo De Oliveira et al. (2021ª), as bases de dados pedológicos nacionais se caracterizam por seu grande volume, com dados oriundos de antigos projetos que tinham como objetivo o mapeamento de solos nos últimos 60 anos, como RADAMBRASIL[8], projeto de 1970, criado pelo Governo Federal e que foi responsável por um inventário de grande parte do país com um mapeamento dos solos em escala 1:1.000.000.

2.3. Dados pedológicos

O termo "dado" é tradicionalmente utilizado de diferentes formas em contextos diversos. Arakaki (2020) acredita que estas distinções se acentuam especialmente dentro dos cenários que envolvem as questões de *Web* Semântica, *Linked Data*, *Big Data*, curadoria digital e e-Science. O autor pontua ainda que o termo tem sido erroneamente empregado como sinônimo de metadados. O uso equivocado destes conceitos pode criar confusões e destaca a importância de se definir os termos de acordo com a área em questão.

Santos & Sant'Ana (2013) definem o termo dado como "uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação".

[7] https://www.go-fair.org/implementation-networks/overview/food-systems/
[8] https://dados.gov.br/dataset/cren_vegetacao_radambrasil

Furner (2016), abordou as diferentes definições que o termo dado possuiu na história e pontuou as diferentes abordagens (extensional, intencional, classificatória, histórica) que este conceito pode assumir, bem como suas diferentes interpretações (clássica, documental, eclesiástica, geométrica, matemática, epistêmica, informacional, computacional e diafórica).

Os metadados, por sua vez, também possuem várias definições. Riley (2017), acredita que

> "os metadados são a chave para a funcionalidade dos sistemas que mantêm conteúdo, possibilitando que os usuários encontrem itens de interesse, registrem informações essenciais sobre eles e compartilhem essas informações com outras pessoas".

Além de serem uma ferramenta de suma importância na cadeira de valor das economias do conhecimento, é importante que os matados possam descrever características relevantes e que possam fazê-lo de forma a respeitar a cultura e as diferenças de idioma (DUVAL, 2002). Existem diversas iniciativas para a definição de metadados de solos, mas a grande maioria delas não está no Brasil. É importante ressaltar que os termos e as definições variam entre os países.

Os dados do solo formam a base dos sistemas de informação do solo em todo o mundo. As necessidades de informações precisas do solo e os estudos feitas pelos pesquisadores provavelmente evoluirão em resposta aos avanços da tecnologia e do *Big Data*. Isso representa um desafio para a comunidade pedológica que já está experimentando um declínio no conhecimento e experiência do solo (ROBINSON, 2019). Com a diminuição da coleta de dados sobre o solo pelos governos, é oportuno reconsiderar como e quais informações sobre o solo devem ser fornecidas aos futuros usuários (ROBINSON, 2019).

O solo pode ser definido como a camada de materiais minerais e/ou orgânicos, formada a partir dos processos químicos, físicos e biológicos da superfície planetária (VAN ES, 2017). Trata-se de um sistema aberto de interação com outros componentes biológicos e geológicos e suas características são resultado do material parental, clima, relevo e tempo de um determinado espaço (SCRIMGEOUR, 2007). A catalogação dos solos se dá a partir de sistemas de classificação taxonômica. Por exemplo, no Brasil utilizamos o Sistema Brasileiro de Classificação de Solos (SIBCS, 2019), mas as taxonomias podem variar entre os diversos países, as diferentes classificações, do ponto de vista computacional, dificultam a integração e o reuso de dados (DA CRUZ et al, 2019).

De acordo com a Sociedade Americana de Ciência do Solo, a pedologia é “o estudo científico dos solos e seus perfis de intemperismo”. Entretanto, historicamente, os termos pedologia e ciência do solo têm sido igualados (BREVIK et al. 2015). Dentro da esfera da pedologia, pode-se incluir também o estudo do que os solos podem nos dizer sobre o presente e os processos anteriores de formação do solo, muitas vezes chamado de “memória do solo”, trazendo, portanto, um conceito geológico de tempo profundo (TARGULIAN & KRASILNIKOV, 2007).

Figura 4 - Perfil aberto para análise de solo (SOIL-NET, website).

A classificação do solo é uma área tradicional da pedologia, ela diz respeito ao agrupamento de solos com uma gama similar de propriedades químicas, físicas e biológicas, que podem ser georreferenciadas e mapeadas. Os solos contêm todos os elementos químicos naturais e combinam simultaneamente estados sólidos, líquidos e gasosos. Além disso, o número de características físicas, químicas, biológicas e suas combinações são quase infinitas (SOLOS, 2013).

A geração de dados de solos se dá partir da coleta em campo, que por sua vez é realizada a partir do se denomina de perfis ou trincheiras, que são aberturas realizadas no solo para a observação e análise dos seus horizontes. Cada perfil possui propriedades químicas, físicas, morfológicas, biológicas, mineralógicas e ambientais particulares, o que implica que as análises realizadas pelos pesquisadores ocorram em três distintos ambientes: no campo *(in situ)*, nos laboratórios *(in vitro)* e nos ambientes computacionais. (CRUZ, et al, 2018).

Os *datasets* de solos são gerados a partir dos processos analítico-experimentais supracitados, em geral, os *datasets* são volumosos e produzidos por equipes distintas e dispersas no espaço e no tempo, ocasionando, consequentemente, em *datasets* desconectados e com problemas na organização dos dados estruturais e lapsos semânticos, isso se considerarmos sob a ótica da Ciência de Dados e da *Web* Semântica. Por outro lado, se considerarmos todas as tecnologias de *Big Data* incluindo novos sistemas de arquivos de banco de dados como Hadoop, Spark e DynamoDB, bem como a proliferação de vários algoritmos de IA (DE OLIVEIRA, et al., 2021a).

Os problemas estruturais dos dados incidem, basicamente, no armazenamento de dados em arquivos de diferentes tipos e formatos ou mesmo dispersos em máquinas distintas, exibindo algumas características típicas de *data lakes* (DE OLIVEIRA et al., 2021a). Além disso, também se verifica em diversos casos lacunas e incompletude das séries de dados, irregularidades e falhas de preenchimento nos dados.

Por sua vez, os problemas semânticos englobam a falta de consistência conceitual, os usos de múltiplos termos e até mesmo erros involuntários oriundos de classificações de solos variantes no tempo em função das taxonomias de solos que sofrem constantes atualizações promovidas pelas comunidades de pesquisadores de solos, essas reclassificações não se refletem de modo automático nos *datasets* pré-existentes, induzindo ou perpetuando, mesmo que involuntariamente, os problemas semânticos pré-existentes ligados às múltiplas classificações em novas classes de solos e ou mesmo em classes extintas.

2.4. *Open Data*

Com iniciativas muito similares às que defendiam os movimentos *Open Source* e *Open Access*, o *Open Data*, ou simplesmente Dados Abertos, apresenta-se como conceito onde os dados sejam livremente disponibilizados para que possam ser utilizados e republicados por todo e qualquer indivíduo sem qualquer restrição de direito autoral ou patente (BRAUNSCHWEIG, 2012). Temos que:

> "O objetivo das iniciativas do *Open Data* é abrir todos os dados não pessoais e não comerciais, especialmente (mas não exclusivamente) todos os dados coletados e processados por órgãos governamentais." (BRAUNSCHWEIG, 2012)

É esperado que os Dados Abertos na AD (e mesmo nas HD) tragam vantagens como a estimulação à participação cidadã e ao incentivo às inovações, além de estimular a transparência

e até mesmo crescimento econômico (ZUIDERWIJK et al, 2014). A partir dessas premissas, diversos portais e plataformas de Dados Abertos foram desenvolvidos nos últimos anos para explorar esse potencial, dentre eles podemos citar, por exemplo, a Eurostat (2013), plataforma europeia de Dados Abertos. No Brasil, temos o Portal de Dados Abertos, onde são catalogados os dados (e metadados) abertos por órgãos e entidades do Poder Executivo Federal. Braunschweig (2012) aponta que para que se possa de fato avançar para uma sociedade aberta, essas plataformas precisam, necessariamente, cumprir requisitos legais e administrativos assim como requisitos técnicos.

Em relação aos pesquisadores que se utilizam de plataformas abertas para obter dados é provável que se deparem com o desafio de não encontrarem em apenas um arquivo todos os dados necessários para a abstração de informações. Isso decorre devido ao fraco acoplamento derivado das distintas e variadas políticas de publicação das organizações. O levantamento e alinhamento desses dados semanticamente é um dos mais árduos trabalhos durante a pesquisa e pode ser feito manualmente ou com o auxílio de softwares.

Braunschweig (2012) também sinaliza que dois estilos contrários para a publicação de dados abertos podem ser identificados: formato de dados legíveis para humanos e formato de dados legíveis para máquinas. Essa classificação é igualmente compartilhada pelos Princípios FAIR.

2.5. *Open Science*

A ciência está se tornando cada vez mais colaborativa e dados intensivos. Ao mesmo tempo, muitos fatores estão pressionando os cientistas a aumentar o acesso e o uso de dados ‘melhores práticas’ em lidar com dados e código. Atualmente, muitos pesquisadores em ciência do solo, AD ou mesmo HD ou quase todos os outros campos estão sendo motivados, muitas vezes por exigência de órgãos financiadores, a adotar os princípios da Ciência Aberta, ou seja, tornar a pesquisa mais acessível e transparente a toda a sociedade e, assim, aumentar a disseminação do conhecimento, levando a um maior impacto científico e social e aumentando a reprodutibilidade de seu trabalho.

Definida por Spellman (2017) como “uma coleção de ações projetadas para tornar os processos científicos mais transparentes e resultados mais acessíveis”, a *Open Science* ou Ciência Aberta, tem o “objetivo de construir uma ciência mais replicável e robusta”. Por outro lado, Fecher (2014) acredita que a *Open Science* é um termo que abrange uma infinidade de pressupostos sobre o futuro da criação e disseminação do conhecimento.

O marco decisivo em prol do acesso aberto ao conhecimento foi a Declaração de Budapeste (*Budapest Open Access Initiative*/BOAI), publicada em 2002, propondo o conceito e as estratégias para o acesso aberto por intermédio da Via Dourada e da Via Verde[9] (SANTOS et al., 2017).

Santos et al. (2017) aponta como os principais benefícios da *Open Science* a reprodutibilidade, transparência científica, velocidade de circulação da informação e o reuso de dados, o que resulta em uma ciência de maior qualidade e progressos mais rápidos.

Henning (2019) acrescenta ainda que a *Open Science* expressa um novo modelo de processo da produção e comunicação do conhecimento refletida nas relações entre ciência, tecnologia, informação e inovação, onde o cidadão é livre para usar, reutilizar e distribuir abertamente a informação assim como os dados científicos, sem restrições tecnológicas e sociais, em um ciclo de pesquisa transparente e aberto, voltado para a colaboração, onde o acesso livre é a prática comum e a restrição legal de acesso a exceção.

Santos et al. (2017) também pontua os desafios a serem enfrentados no processo de abertura e compartilhamento de dados científicos, uma vez que esta nova perspectiva pode provar mudanças radicais nos paradigmas que interferem em valores e princípios muito caros aos pesquisadores, como a autonomia e o reconhecimento, além dos impactos trazidos pela necessidade de soluções normativas e tecnológicas mais complexas.

Destacamos que os Princípios FAIR, bem como os movimentos de Ciência Aberta, Dados Abertos e LOD, surgem como propostas de soluções na longa esteira no que diz respeito à reprodutibilidade, transparência e disponibilidade de dados e que poderão ser usados de modo a agregar novos valores a esta pesquisa de mestrado e serão mais bem compreendidos nas seguintes seções desse trabalho.

2.6. *Web* Semântica

Seguramente este é um dos itens centrais desta dissertação. Com a escalada na quantidade de dados e a limitada estruturação da *Web* tornou os mecanismos de busca clássicos insuficientes para a organização e recuperação de informações. A partir dessa constatação, a

[9] Via Dourada: Acesso Aberto por meio da publicação de artigos em periódicos com Acesso Aberto (AA) (SANTOS et al., 2017).
Via Verde: Acesso Aberto por meio do depósito/auto arquivamento de artigos publicados em periódicos, anais e apresentações de conferências, revisados por pares, além de teses e dissertações, em um repositório de Acesso Aberto (SANTOS et al., 2017).

busca de uma melhor estruturação da *Web* tornou-se uma tendência. Neste contexto, segundo Alves (2005), a *Web* Semântica surgiu "com o intuito de melhorar a recuperação de recursos em ambientes informacionais como a *web*".

Proposta em 2001 por Berners-Lee, Hendler e Lassila (BERNERS-LEE et al, 2001), a *Web* Semântica tinha como objetivo original representar informações de maneira contextualizada e estruturada, mantendo uma intersecção no nível de compreensibilidade tanto pela máquina quanto pelo ser humano a partir do fornecimento semântico a esses dados. A *Web* Semântica apresenta diversos elementos teóricos, tecnologias e aplicações (científicas e comerciais) que indicam a sua consistência.

Web Semântica permite que se reúnam fontes heterogêneas de dados e se forneça significado aos dados. Além disso, segundo Hitzler et al. (2009) ela possui ainda mais uma raiz histórica: a possibilidade de construção de modelos abstratos capazes de capturar e representar as complexidades do mundo em termos simples, o que se conhece por modelagem conceitual semântico.

Atualmente, o *World Wide Web Consortium* (W3C) é o órgão principal responsável pelas padronizações presentes na *Web* Semântica. O W3C (2013b) define a *Web* Semântica como uma grande rede de dados. Neste sentido, Koivunen & Miller (2001, tradução nossa) afirmam ainda que:

> "A *Web* Semântica possibilita não apenas resultados mais precisos nas buscas por informações, ela também possibilita que saibamos integrar informações de fontes variadas, quais informações comparar e fornecer tipos variados de serviços automatizados em domínios diferentes".

Koivunen & Miller (2001) ainda destacam que com a *Web* Semântica podemos associar informações descritivas de maneira global a qualquer recurso. A partir disso, é possível realizar consultas. O *Uniform Resource Identifier* (URI) é o identificador único de um recurso na *Web* e na *Web* Semântica é atribuído à documentos, pessoas, conceitos e relacionamentos.

O termo *Web* Semântica é muito amplo e pode ter várias interpretações, nesta dissertação consideramos que representa uma *Web* de dados lincados (*linked data*). Entretanto, para que ela se torne uma realidade conforme o apregoa o W3C (2010), uma quantidade grande de dados precisa estar disponível em um formato que seja acessível e gerenciável por tecnologias adequadas, além de, primordialmente, estarem padronizados (CATARINO, 2012).

Outro aspecto fundamental de fundamental da *Web* Semântica é a modelagem e uso de ontologias. Resumidamente, uma ontologia descreve formalmente os conceitos e as relações dos recursos presentes em determinado domínio do conhecimento, permitindo que esses possam ser utilizados no processamento semântico de dados (KITAMURA & MIZOGUCHI, 2004). Essa conceitualização é essencial para que humanos e máquinas sejam capazes de compreender a semântica formal dos recursos, e assim, permitir a realização de inferências. Existem várias linguagens e abordagens para a criação de ontologias.

Destacamos que na área de solos não existem muitas ontologias na literatura, exceto pelo projeto *Global Soil Partnership* (GSP) (RODRÍGUEZ EUGENIO, 2021). Trata-se de uma rede global de partes interessadas que promove práticas sólidas de manejo de terras e solos para um sistema alimentar mundial sustentável. O GSP reconheceu a relevância das políticas globais e transnacionais para práticas sustentáveis de manejo da terra e elegeu a harmonização e troca de dados de solos como uma de suas principais linhas de ação. O GSP usa a ontologia GloSIS (uma implementação do modelo de domínio GloSIS descrita em OWL para apresentar um extenso conjunto de listas de códigos prontos para uso para descrição do solo e análise físico-química). Além disso, adota uma série de padrões da *Web* Semântica (SOSA, SKOS, GeoSPARQL, QUDT, entre outros).

A *Web Ontology Language* (OWL) (MCGUINNESS, 2004) é uma delas linguagens semânticas, sendo recomendada pela W3C para a construção de ontologias, utilizando o conceito do RDF para interligar os recursos. Hoje, existem diversas tecnologias que são utilizadas para a implementação da *Web* Semântica. Nesta pesquisa, entretanto, nos limitaremos à apenas algumas delas, como o *Resource Description Framework* (RDF) (MILLER, 1998) e o RDF *Schema* (RDFS) (BRICKLEY, 1998), a *Extensible Markup Language* (XML) (BRAY, 1997) e o SPARQL *Protocol and RDF Query Language* (SPARQL) (PRUD'HOMMEAUX, 2008), além do URI (BERNERS-LEE et al, 2001), citados nesta seção.

2.6.1. *Uniform Resource Identifier* (URI) e *Extensible Markup Language* (XML)

A identificação é a principal informação na representação de um recurso. Ramalho (2016) afirma que os URI possibilitam uma maneira única e global de nomear itens. Ou seja, o URI é o responsável por tornar um recurso como único na *Web*.

A construção do URI possui algumas particularidades. Os URI não referentes a recursos disponíveis na *Web* podem ter a mesma construção, seguindo a estrutura padrão utilizados nos endereços *Web* (FERREIRA, 2014), representada na Figura 5.

Figura 5 – Sintaxe URI (Elaborado pela autora).

Hitzler et al. (2009) consideram o esquema a principal característica de URI. O *Hypertext Transfer Protocol* (HTTP) é um dos mais conhecidos esquemas por ser utilizado nas URL e consiste, basicamente, em um protocolo de transmissão de informações na *Web*. Este esquema, entretanto, também pode ser utilizado nos URI de recursos não disponíveis na *Web*.

A XML, por sua vez, é uma linguagem de marcação genérica utilizada para descrever um conjunto de recursos de determinado cenário. Criada em 1998 tendo como base a *Standard Generalized Markup Language* (SGML), é uma linguagem especificada na ISO 8879.

Ray (2001) descreve algumas das características da XML, dentre elas, podemos citar a possibilidade de armazenamento e organização de qualquer tipo de informação em um formato capaz de ser adequado a diferentes necessidades. Além disso, Ray (2001) cita que graças a sua sintaxe simples e por possuir uma estrutura livre de ambiguidades, a XML é relativamente simples de se compreender por humanos e máquinas.

Os elementos são as unidades de armazenamento utilizadas na construção de documentos XML. Ray (2001) destaca que:

- todos os elementos precisam de *tags* de início e fim, similar ao que já estamos habituados a ver com HTML. Exemplo: <nome>Marcação</nome>;
- a representação de um elemento vazio é feita por uma *tag* única e precisa de uma barra (/) antes de ser fechada. Exemplo: <vazio/>;
- os atributos dos elementos recebem seus valores entre aspas duplas. Exemplo: <nome tipo=”objeto”>;
- não é possível sobrepor os elementos. Exemplo: <nome>UFRRJ<idade>107</idade></nome>;
- os caracteres de marcação (<,]]>, e &) não podem ser utilizados no conteúdo de marcação. Para utilizá-los, é preciso substituí-los pela sua notação equivalente. Exemplo: para representar o caractere “<”, digitamos “<”.

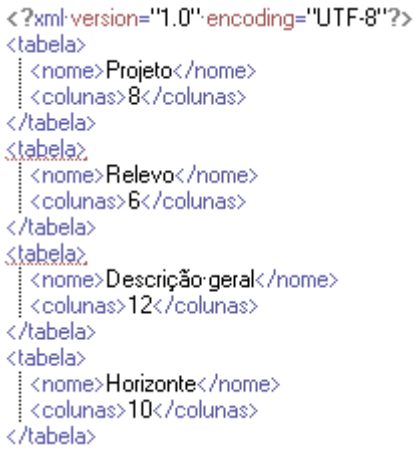


Figura 6 – Exemplo de documento XML com descrição de tabelas do *OpenSoils* (Elaborado pela autora).

A linguagem XML em conjunto com o modelo RDF são duas tecnologias fundamentais para a descrição de recursos no ambiente *Web*. Discutiremos melhor o modelo RDF na seguinte seção.

2.6.2. *Resource Description Framework* (RDF)

Recomendado pelo W3C, o RDF é um modelo de metadados e linguagem utilizado para construir uma infraestrutura de leitura por máquina para os dados na *Web* (GUTIERREZ, 2006).

Segundo Miller (1998), a história dos metadados no W3C começou em 1995 com o padrão *Platform for Internet Content Selection* (PICS), que tinham como principal objetivo possibilitar a classificação e a descrição dos conteúdos das páginas *Web*. O PICS possibilitou, por exemplo, a identificação de conteúdo impróprio como violência, nudez e conteúdo sexual a partir da descrição e uma das maiores motivações para o seu desenvolvimento foi a antecipação de restrições de conteúdos na internet nos Estados Unidos. Entretanto, lacunas foram identificadas e a necessidade de uma descrição mais abrangente das páginas foi o que impulsionou a criação de um grupo de trabalho denominado *Resource Description Framework*.

Miller (1998) define o RDF como uma infraestrutura que permite a codificação, intercâmbio e o reuso de metadados estruturados, de forma que esta infraestrutura permita a interoperabilidade de metadados a partir da concepção de mecanismos que possibilitam o suporte de convenções comuns de semântica, sintaxe e estrutura.

> "No modelo RDF, o universo a ser modelado é um conjunto de recursos onde cada um deles é identificado por um *identificador de recurso universal* (URI), a linguagem a descrevê-los é um conjunto de *propriedades*. As descrições são declarações na estrutura *sujeito-predicado-objeto*. Tanto o *sujeito* quanto o *objeto* podem ser objetos anônimos, conhecidos como *nós em branco*." (GUTIERREZ, 2006).

Um documento RDF é estruturado em forma de grafo. Isto significa que se trata de um conjunto de nós e arestas direcionados em que cada um dos elementos possuem um identificador que os distingue (FERREIRA, 2013).

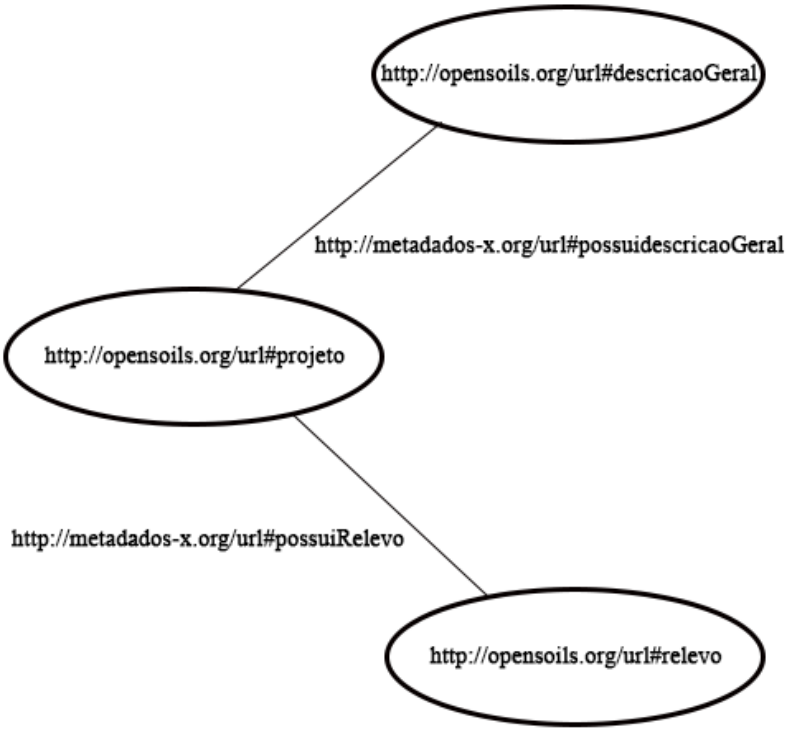


Figura 7 – Grafo RDF descrevendo o relacionamento de projeto de análise de solos, sua descrição geral e relevos (Elaborado pela autora).

A modelagem em grafos também possibilita a representação de valores literais, ou seja, valores de dados de um certo tipo de dados. Nestes casos, são utilizados retângulos, como mostra a Figura 8.

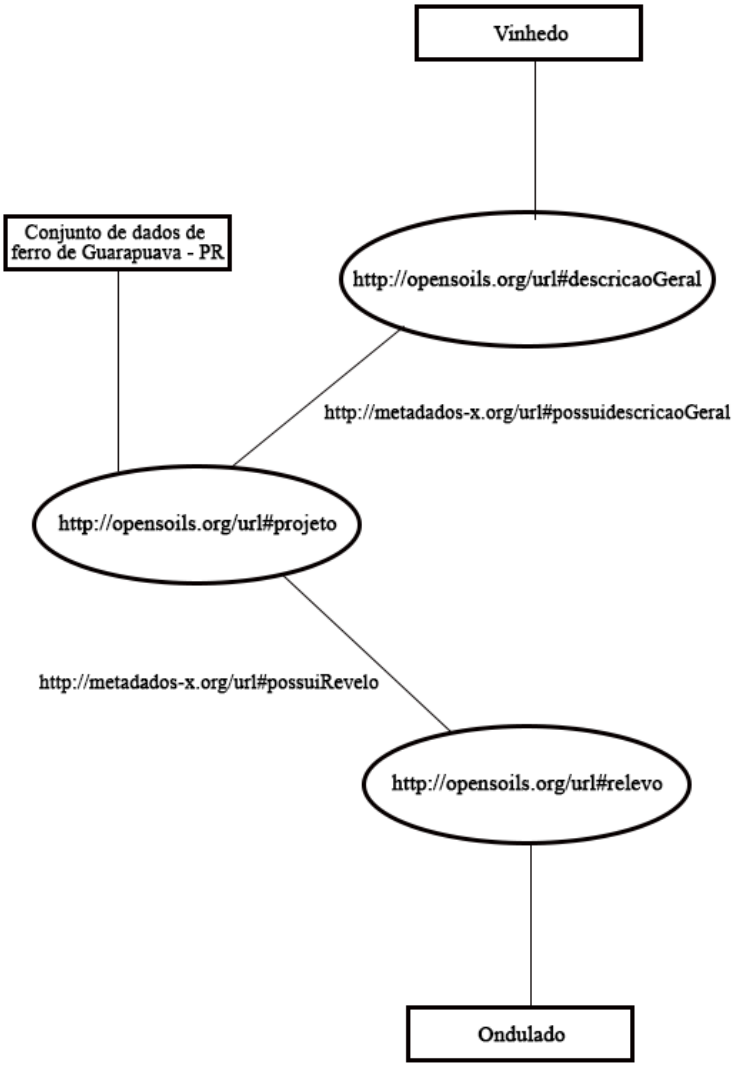


Figura 8 – Grafo RDF com literais considerando instâncias de dados do repositório de dados do *OpenSoils*.

Os URI possibilitam a nomeação e identificação de recursos abstratos, mas podem ser tratados como nomes já que a atual interpretação pretendida de um URI em particular não se dá de maneira formal. Com isso, ferramentas específicas podem possuir meios particulares de interpretar URI (HITZLER; KRÖTZSCH; RUDOLPH, 2010).

A representação do modelo RDF por meio de grafos possibilita a leitura por humanos e oferece uma maneira eficiente de modelagem conceitual. Entretanto, tratando-se de máquinas, está claramente não é uma representação viável. Isso porque, geralmente, o volume de dados não possibilita a representação visual e para isto existem as representações por cadeiras de caracteres que podem ser processadas por máquinas (FERREIRA, 2013). Para que este tipo de representação seja possível, um grafo RDF precisa ser subdividido em partes menores que podem ser armazenadas uma por uma. Esta transformação de estruturas de dados complexas

para cadeiras de caracteres é denominada serialização (HITZLER; KRÖTZSCH; RUDOLPH, 2010).

A principal sintaxe utilizada para a realização da serialização atualmente é a XML, porém, outras sintaxes como JSON, N-Triples, Turtle e Notation3 (N3) também podem ser utilizadas. A figura 9 representa a serialização indicada previamente pela figura 8 a partir da sintaxe XML.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.irg/1999/02/22-rdf-syntax-ns#"
 xmlns:cod="http://metadados-x.org/uri#">

 <rdf:Description rdf:about="http://opensoils.org/uri#projeto">

  <cod:projeto>Conjunto de dados de ferro de Guarapuava - PR</cod:projeto>

  <cod:possuidescricaoGeral>
   <rdf:Description rdf:about="http://opensoils.org/uri#descricaoGeral">
    <cod:usoAtual>Vinhedo</cod:usoAtual>
   </rdf:Description>
  </cod:possuidescricaoGeral>

  <cod:possuiRelevo>
   <rdf:Description rdf:about="http://opensoils.org/uri#relevo">
    <cod:relevoLocal>Ondulado</cod:relevoLocal>
   </rdf:Description>
  </cod:possuiRelevo>

 </rdf:Description>

</rdf:RDF>
```

Figura 9 – Serialização em XML dos grafos RDF ilustrado nas Figuras 7 e 8.

Dentre os desafios encontrados para a utilização prática do RDF pelas comunidades de descrição de recursos, podemos destacar a dificuldade de uma criação e utilização corretas de URIs que servirão de sujeitos, predicados e objetos nas declarações constituintes das descrições em RDF (FERREIRA, 2013). Além disso, uma vez que as descrições são concluídas, linguagens para consultar estas descrições se fazem necessárias para que estes dados possam ser encontrados e devidamente utilizados. Veremos na seção seguinte uma das tecnologias mais comumente utilizadas para este tipo de consulta.

2.6.3. SPARQL *Protocol* RDF *Query Language* (SPARQL)

A SPARQL é uma linguagem que foi desenvolvida para consultar dados modelados em formato RDF. O W3C (2013ª) a define como um conjunto de especificações que fornecem linguagens e protocolos para consultar e manipular os conteúdos RDF que estão disponíveis na *Web* ou em qualquer banco de dados RDF.

Como os dados RDF são representados em grafos, a SPARQL é uma linguagem de consulta baseada em grafos que tem a capacidade de realizar consultas por padrões de grafos obrigatórios e opcionais em conjunto com conjunções e disjunções (W3C, 2012).

Breslin et al. (2009) informam que a linguagem SPARQL guarda alguns aspectos do SQL[10] da *Web* Semântica, entretanto, ambas possuem diferentes abordagens, uma vez que a SQL lida com tabelas e a SPARQL com grafos. A SPARQL, inclusive, possibilita a realização de consultas mais avançadas se comparada à linguagem SQL. Breslin et al. (2009) ainda afirmam que quatro formas de consultas são utilizadas na linguagem SPARQL, são elas: SELECT, CONSTRUCT, ASK e DESCRIBE.

As operações SELECT são utilizadas com o intuito de recuperar informações baseadas padrão de triplas. O CONSTRUCT tem como objetivo a criação de documentos RDF baseados em outros documentos RDF de entrada, de forma que possa ser utilizado como um mecanismo de tradução para os dados RDF. O ASK, por sua vez, é utilizado com o intuito de identificar padrões particulares de consulta que possam ser correspondidos no grafo RDF consultado. O DESCRIBE, por fim, tem como objetivo a identificação de todas as triplas relacionadas a um objeto em específico que precisa ser descrito (FERREIRA, 2014).

2.7. *Linked Open Data*

O *Linked Open Data,* ou simplesmente Dados Abertos Interligados, consiste na prática de interligação de dados, bem como na disponibilização dos mesmos na *Web* e pode ser mais bem compreendido a partir da fundamentação dos princípios *Linked Data* ou Dados Interligados, apresentados por Berners-Lee (2006).

A aplicação de tais princípios possibilita que todas as informações relacionadas a um elemento estejam devidamente conectadas e possam, portanto, serem descobertas. Berners-Lee (2006) destaca que os dados podem estar interligados, mas não estarem necessariamente abertos. Por isso, os princípios *Linked Data* são, especificamente, sobre a interligação dos dados, estando eles abertos ou não.

> Os quatro princípios *Linked Data* (BERNERS-LEE, 2006) consistem, basicamente:
>
> 1. A utilização de URIs para a nomeação de recursos;
> 2. A utilização de HTTP URIs para que os nomes possam ser localizáveis;

[10] *Structured Query Language* (SQL) (Linguagem de Consulta Estruturada), linguagem de consulta declarativa padrão para banco de dados relacionais, com características originarias na álgebra relacional.

3. Quando uma busca por um URI for realizada, devem ser fornecidas informações úteis a partir da utilização de padrões (SPARQL, RDF);
4. A inclusão de *links* para outros URIs, possibilitando a descoberta de novos recursos a partir disso.

Os URI são elementos os responsáveis pela identificação única do recurso e o HTTP possibilita que este recurso se torne localizável na *Web* (NININ, 2018). Por outro lado, é a partir do RDF que se dá o relacionamento entre diferentes recursos e a manipulação é feita com o SPARQL (NININ, 2018).

Em contrapartida, o *Linked Open Data* consiste na aplicação dos princípios *Linked Data* de acordo com os princípios *Open Data*. Em outras palavras, ao contrário do que encontramos na definição dos princípios *Linked Data*, – onde os dados podem ou não ser abertos – os dados são, necessariamente, abertos.

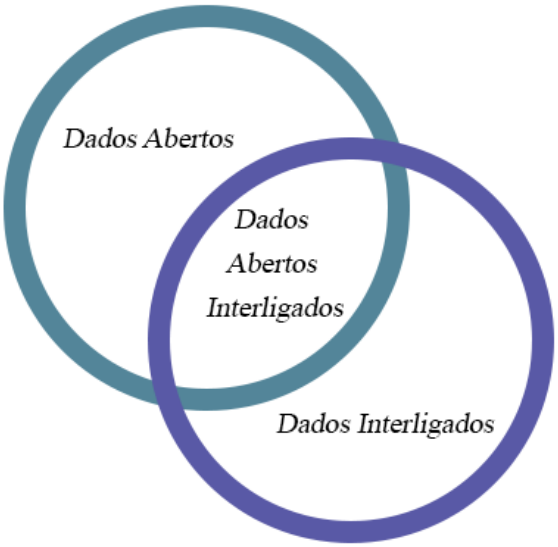


Figura 10 – Diagrama de iniciativas relativas ao Linked Data (Elaborado pela autora).

A partir do *Linked Open Data*, a iniciativa *The Linked Open Data Cloud* (*LOD-Cloud*) ganhou destaque. Com início em 2007 pelo *Semantic Web Education and Outreach Interest* e com o objetivo de converter, publicar e criar uma rede de *datasets* interligados de acordo com o *Linked Open Data* (BIZER; HEATH; BERNERS-LEE, 2009), a iniciativa encontra-se em constante crescimento.

Com *datasets* que alcançam recursos dos mais variados domínios como publicações em geral, ciências da vida, domínio geral (*cross-domain*), dados geográficos e governamentais, bem como mídia, dados de uso geral, de redes sociais e linguísticos (NININ, 2018), podemos citar, dentre eles, o conjunto de dados do *Wikipedia*, o *Dbpedia*.

O projeto *DBPedia* se concentra na tarefa de converter o conteúdo da Wikipédia em conhecimento estruturado, de modo que as técnicas da *Web* Semântica possam ser empregadas contra ela – fazendo consultas sofisticadas contra a Wikipédia, ligando-a a outros conjuntos de dados na *Web* ou criando aplicações (AUER, 2007). A vinculação criada pelo *Linked Data* fornece uma experiência de pesquisa mais imersiva, com conteúdo mais rico (BROWELL, 2015). O consumo do *Linked Open Data* por sua vez, além de possibilitar a integração, fornece informações de alta qualidade e atualizadas sobre diversos tópicos específicos (BAUER, 2011).

2.8. Princípios FAIR

Publicado em 2016, o *FAIR Guiding Principles for scientific data management and stewardship* (WILKINSON *et al*., 2016) tem como objetivo melhorar a localizabilidade, acessibilidade, interoperabilidade e reusabilidade de objetos de pesquisa digital tanto para humanos quanto máquinas. A necessidade de aprimoramento na criação e no compartilhamento do conhecimento foi o ponto de partida para um grupo de pesquisadores em 2011, na Alemanha (WILKINSON *et al*., 2016).

O cenário que antecede a publicação dos princípios FAIR em 2016 apresentava, de uma maneira geral, iniciativas que, por si só, demonstravam a necessidade de, dentre outros, uma maior eficácia no compartilhamento e na reutilização de dados de pesquisa. A conferência *Jointly designing a data FAIRPORT* de 2014 é um exemplo concreto deste cenário. O workshop reuniu 25 participantes de alto nível, representando as principais infraestruturas de pesquisa e institutos de políticas, editores, especialistas em web semântica, inovadores, cientistas da computação e cientistas experimentais.

Desde então, os princípios FAIR passaram a ser mais comumente aplicados à dados de pesquisa, porém, as ideias por trás destes princípios são igualmente relevantes para softwares de pesquisa (LAMPRECHT, 2020).

Cada um dos quatro princípios norteadores de alto nível – *Findable*, *Accessible*, *Interoperable*, *Reusable* – possui suas divisões e subdivisões, bem como suas características de aplicabilidade.

Tabela 1 – Princípio norteador – Localizável

LOCALIZÁVEL (*FINDABLE*)	
F1	(Meta)dados são atribuídos a identificadores globais únicos e persistentes.
F2	Dados são descritos com metadados ricos.
F3	Metadados clara e explicitamente incluem o identificador dos dados que descrevem.
F4	(Meta)dados são registrados ou indexados em um recurso pesquisável.

Tabela 2 – Princípio norteador – Acessível

ACESSÍVEL (*ACCESSIBLE*)	
A1	(Meta)dados são recuperáveis por seu identificador, utilizando-se de um protocolo padronizado de comunicação.
A1.1	O protocolo é aberto, gratuito e universalmente implementável.
A1.2	O protocolo permite um procedimento de autenticação e autorização, quando necessário.
A2	(Meta)dados são acessíveis mesmo quando os dados não estão mais disponíveis.

Tabela 3 – Princípio norteador – Interoperável

INTEROPERÁVEL (*INTEROPERABLE*)	
I1	(Meta)dados utilizam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento.
I2	(Meta)dados utilizam vocabulários que seguem os Princípios FAIR.
I3	(Meta)dados incluem referências qualificadas a outros (meta)dados.

Tabela 4 – Princípio norteador – Reutilizável

REUTILIZÁVEL (*REUSABLE*)	
R1	(Meta)dados são ricamente descritos com uma pluralidade de atributos precisos e relevantes.
R1.1	(Meta)dados são liberados com uma licença de uso de dados clara e acessível.

R1.2	(Meta)dados são associados à proveniência detalhada.
R1.3	(Meta)dados atendem aos padrões da comunidade relevantes para o domínio.

É possível destacar diversas adesões significativas de diversas áreas do conhecimento aos princípios FAIR, como por exemplo a *Big Data to Knowledge* (BD2K) do *American National Institutes for Health* (NIH) (SCIENTIFIC ELECTRONIC LIBRARY ONLINE, 2016). A *Scientific Electronic Library Online* (Scielo) declarou durante a conferência Jointly designing a Data FAIRPORT que além de promover os princípios FAIR, pretende adotá-los para a gestão dos seus dados científicos (SCIENTIFIC ELECTRONIC LIBRARY ONLINE, 2016).

> "[...] parece certo de que estes princípios rapidamente se tornarão uma base crucial para inovação no movimento global em direção a ambientes de Ciência Aberta." (SCIENTIFIC ELECTRONIC LIBRARY ONLINE, 2016).

Embora se reconheça a importância de adaptabilidade à tais princípios no mundo científico nacional e internacional, sua aplicação ainda é rodeada de muitas dúvidas. Pesquisas recentes identificaram inclusive que uma grande parte dos repositórios digitais da Holanda, por exemplo, possuem um baixo grau de compatibilidade aos princípios FAIR (HENNING, 2019).

A partir deste cenário é que a iniciativa *Global Open* FAIR (GO FAIR) desponta com o objetivo de promover a implantação de práticas e serviços que adotem totalmente os princípios FAIR (GO FAIR, 2018). Liderado pelo *Dutch Techcentre for Life Sciences* (DTL) e apoiado pelos governos da Holanda, Alemanha e posteriormente França, a iniciativa GO FAIR é criada. Suas atividades planejadas visam promover a utilização e reutilização de dados, bem como a elaboração de um modelo de gerenciamento global de dados de pesquisa. Além disso, a iniciativa GO FAIR se sustenta em três pilares que tem como objetivo embasar três categorias de atividades: cultura (*GO Change*), treinamento (*GO Train*) e tecnologia (*GO Build*). (DUTCH TECHCENTRE FOR LIFE SCIENCES, 2017).

1) ***GO Change*** tem como objetivo promover as mudanças culturais necessárias para tornar os princípios FAIR um padrão de trabalho na ciência de forma de que se reconheçam as atividades de ciência aberta.
2) ***GO Train***, por sua vez, promove o treinamento para que se localize, crie, mantenha e sustente o conhecimento sobre gerenciamento de dados. O objetivo é a formação de especialistas qualificados e certificados.

3) ***GO Build*** trata da construção de infraestrutura para dados interoperáveis, bem como a construção de padrões técnicos e melhores práticas para a implementação dos princípios FAIR.

No Brasil, alguns movimentos começam a ter destaque para a adesão da iniciativa GO FAIR. Esta adesão vem se consolidando por intermédio do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e deve ser estruturada com representantes segundo uma visão de domínios ou áreas de conhecimento, o que possibilitará a inserção de dados de pesquisa brasileiros nas nuvens da ciência aberta global (HENNING, 2019).

Como exemplo, podemos citar o GO FAIR Brasil[11], iniciativa nacional com o objetivo de possibilitar a adoção da estratégia de implementação dos princípios FAIR definida pela iniciativa GO FAIR em todo o território brasileiro. Além disso, a Rede Agro[12] é uma das Redes de Implementação Temáticas do GO FAIR Brasil que é coordenada pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e visa a implantação dos princípios FAIR no universo de dados da pesquisa agropecuária brasileira (DRUCKER et al., 2021).

Em geral, a FAIRificação de 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) Mapear e carregar *datasets* (dados brutos) disponíveis na *Web*;

2) Análisar dados para verificar conteúdo; conceitos representados, estruturas, as relações entre os elementos que constituem os dados;

3) Definir/escolher 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 aceder aos dados;

[11] https://www.go-fair-brasil.org/
[12] https://www.go-fair-brasil.org/agro

6) Anotar os dados com metadados e proveniência permitindo que tanto seres humanos como máquinas possam localizar os dados;

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 (DE OLIVEIRA et al, 2021).

No contexto da plataforma *OpenSoils*, Bessa (2021) conceituou um estudo inicial com vistas a automação do processo de FAIRificação em uma evidente adesão às iniciativas GO FAIR.

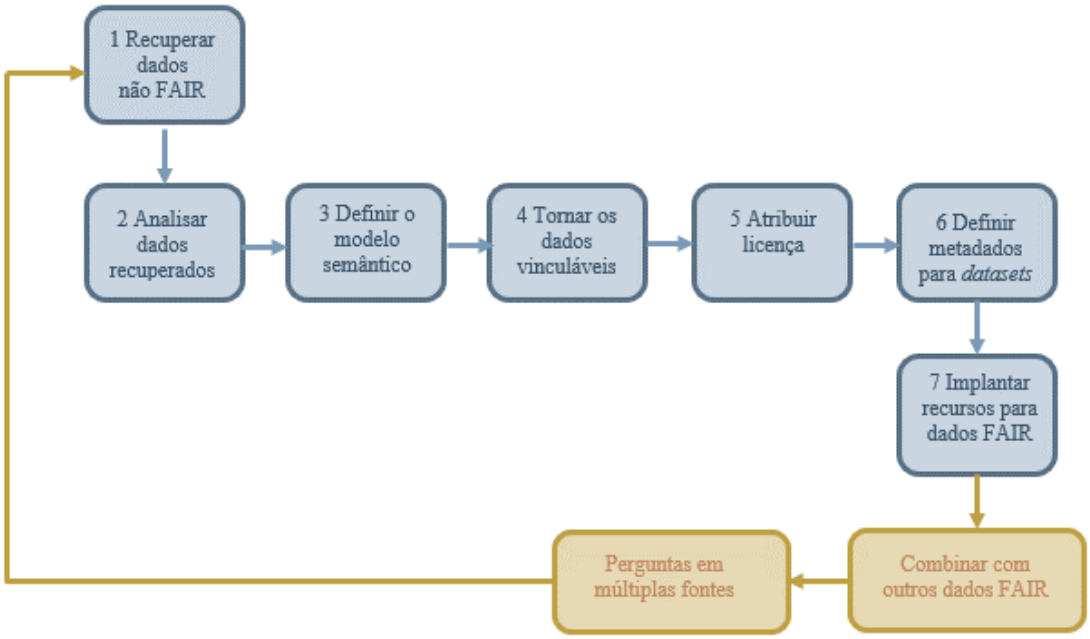


Figura 11 – Processo de FAIRificação (BESSA, 2021).

Embora dados FAIR não sejam necessariamente Dados Abertos (MONS, 2018), ao aderir aos princípios FAIR, contribui-se para a Ciência Aberta, uma vez que os dados de pesquisa poderão ser facilmente reutilizados, terão um grau de confiança maior devido a sua qualidade, além de serem facilmente localizáveis graças aos seus metadados (BESSA, 2021).

2.9. Considerações finais

A literatura nos apresenta que o processo de aplicação da computação nas Ciências Humanas ocorreu por volta dos anos 60, com o padre jesuíta Roberto Busa desenvolvendo *index verborum* com as palavras contidas em todas as obras de São Tomás de Aquino e recebeu o reconhecimento de pioneirismo na aplicação de tecnologia em uma pesquisa de humanidades.

Desde então, muito se discutiu: sobre o termo utilizado para denominar a área, que evoluiu de Humanidades Computacionais para Humanidades Digitais, a definição da área também trouxe e traz bastante discussões e ainda não existe um consenso sobre a sua delimitação. Apesar disso, uma característica marcante da área é a digitalização. Existem exemplos importantes na Europa de iniciativas inclusive fomentadas pelo governo para a realização da digitalização de acervos.

O processo de datificação também é característico nas discussões sobre Humanidades Digitais, que é justamente a geração de dados a partir de ações cotidianas, por exemplo. Esse é um processo da era do Big Data e está ocorrendo o tempo inteiro em todas as áreas e na agricultura isso não é diferente.

A agricultura também teve seu processo de evolução. Passou pela Agricultura 1.0, que tinha como principal característica a mão de obra familiar, a Agricultura 2.0 com o início da utilização da força mecânica, a Agricultura 3.0 que trouxe a agricultura de precisão e finalmente a Agricultura 4.0, que é a que vivemos atualmente e também é comumente chamada de Agricultura Digital.

Os dados que gerados pela agricultura, de uma maneira geral, são dados difíceis de serem integrados, além de apresentarem problemas semânticos e estruturais muitas vezes devido ao compartilhamento. E os dados de solos ou dados pedológicos sofrem desses mesmos problemas.

A pedologia é o estudo científico do solo e nesse contexto a classificação do solo passa por uma série de análises de propriedades físicas, químicas e morfológicas. A geração dos dados se dá justamente a partir da coleta em campo, uma vez coletadas essas amostras passam por análises que geram os dados das propriedades químicas, físicas, morfológicas, que geralmente são dados bastante volumosos principalmente se pensarmos que em um projeto que é realizado, são coletadas comumente diversas amostras.

Os dados pedológicos que estamos utilizando nesse trabalho são dados abertos, ou seja, dados que são livremente disponibilizados e que podem ser utilizados e republicados sem restrições. O objetivo das iniciativas do *Open Data* é abrir todos os dados que não sejam dados pessoais ou dados comerciais, principalmente os dados coletados e processados por órgãos governamentais. Espera-se que os Dados Abertos tragam vantagens como o estímulo a participação cidadã, o estímulo a transparência e também fomentem o crescimento econômico.

O *Open Science*, por sua vez, tem o objetivo de construir uma ciência mais replicável e robusta e para isso o movimento conta com uma série de ações que foram projetadas para tornar os processos científicos mais transparentes e resultados mais acessíveis. Os Princípios FAIR e o movimento *Linked Open Data* surgem, inclusive, como propostas de soluções no processo do *Open Science*.

A *Web* Semântica foi proposta por Berners-Lee em 2001 e surgiu com o objetivo de representar as informações na web de maneira contextualizada e estruturada, isso porque com a escalada da quantidade de dados, os mecanismos de buscas se tornaram insuficientes para recuperar informações. Como vimos anteriormente, a *Web* Semântica é um termo bastante amplo, pode ter diversas interpretações, mas nesse trabalho consideraremos que ela representa uma *web* de dados lincados.

O *Linked Open Data* é a prática da interligação de dados e a disponibilização desses dados na web. A interligação dos dados possibilita que as informações relacionadas a um elemento estejam devidamente conectadas e possam ser descobertas. O LOD não é uma exclusividade da área de computação, é um movimento global. A disponibilização de dados ligados favorece o enriquecimento social e o aumento dos *datasets* com aderência ao *Linked Open Data* pelos anos é resultado do incentivo não apenas a estruturação e a lincagem de dados, mas também do incentivo aos dados abertos.

Os Princípios FAIR foram publicados pela primeira vez em 2016, FAIR é um acrônimo para Findable, Accessible, Interoperable e Reusable. Surgem de uma necessidade de uma maior eficácia no compartilhamento e na reutilização de dados de pesquisa. Os Princípios FAIR têm sido mais comumente aplicados a dados de pesquisa, mas a ideia por trás dos Princípios é relevante na mesma proporção para softwares de pesquisa. Por serem Princípios que direcionam pra uma maior qualidade no compartilhamento e na reutilização dos dados, nós nos apoiamos nos Princípios FAIR para desenvolvimento desse trabalho.

3. METODOLOGIA DE PESQUISA

A presente pesquisa pode ser identificada como teórico aplicada. Segundo Gerhardt & Silveira (2009), este tipo de pesquisa tem como finalidade a geração de conhecimentos para aplicação prática e são dirigidas a soluções de problemas específicos.

Nesta pesquisa, analisou-se a partir da perspectiva *Linked Open Data*, o repositório digital da plataforma *OpenSoils* e se propõe a possibilidade da aplicação de tecnologias inteligentes da *Web* Semântica para a adequação do repositório supracitado às características do movimento *Linked Open Data*, *Open Data e* princípios FAIR.

Do ponto de vista da aplicação, se pretende utilizar tecnologias da *Web* Semântica e Ciência de dados no repositório digital da plataforma *OpenSoils* com o intuito de alcançarmos os princípios apresentados pelo movimento *Linked Open Data*, que segundo Berners-Lee (2006), eles consistem basicamente em:

- A utilização de URIs para a nomeação de recursos;
- A utilização de HTTP URIs para que os nomes possam ser localizáveis;
- Quando uma busca por um URI for realizada, devem ser fornecidas informações úteis a partir da utilização de padrões (SPARQL, RDF);
- A inclusão de *links* para outros URIs, possibilitando a descoberta de novos recursos a partir disso.

É importante destacar ainda que a pesquisa possui características exploratórias, cujo principal recorte de escopo é compreender e contribuir com a explicitação das intersecções entre as facetas das Humanidades Digitais e Agricultura Digital, tema com pouco conhecimento acumulado até o presente momento (BESSA, 2021).

Destacamos que devido ao escopo da pesquisa e do recorte metodológico adotado, não se está interessado em desenvolver novas ontologias da área de solos (SILVA et al, 2006, ANDRÉS et al, 2017, DEB et al, 2020, HELFER et al, 2021, entre outros). Nossos estudos preliminares indicam que essas ontologias são muito heterogêneas tanto em termos de construção e acesso aos arquivos fonte, quanto em termos de representação semântica dos seus conceitos e relações.

Foi realizado ainda o levantamento bibliográfico em bases de dados especializadas com o objetivo de investigarmos a utilização de tecnologias para a aplicação dos conceitos levantados no capítulo anterior desta pesquisa. A partir disso, detalharemos nas seguintes subseções a revisão sistemática da literatura (RSL) realizada para o levantamento bibliográfico

bem como a definição e escolha dos principais trabalhos relacionados. Após a RSL, apresentamos a proposta dos workflows ETLH e execução de experimentos computacionais para a validação dos artefatos propostos.

3.1. Revisão Sistemática na Literatura

Nesta subseção apresentamos uma RSL para avaliar as evidências relacionadas com as estratégias computacionais disponíveis na literatura. Nossa revisão busca identificar os artigos mais relevantes que contribuam para um melhor entendimento do problema tratado nesta pesquisa. A RSL foi conduzida seguindo um protocolo que consiste em uma versão adaptada das recomendações propostas por Kitchenham & Charters (2007). Além disso, a ferramenta Parsifal[13] foi utilizada para a sua estruturação.

3.1.1. Objetivo e perguntas da pesquisa

O objetivo dessa subseção é identificar as iniciativas de aplicação da *Web* Semântica em estudos que correlacionam Humanidades Digitais e Agricultura Digital. A partir desse ponto, quatro questões de pesquisa (P) foram definitas com o objetivo de guiar a seleção dos trabalhos, bem como a extração de dados:

- **P1:** Como a web semântica é aplicada em problemas das humanidades digitais?
- **P2:** Quais são as iniciativas existentes de dados interligados no contexto das humanidades digitais?
- **P3:** Como a agricultura digital e as humanidades digitais se relacionam?
- **P4:** Quais as iniciativas existentes de dados interligados no contexto das humanidades digitais e da agricultura digital?

3.1.2. Estratégias de busca e fontes de dados

As fontes de buscas utilizadas neste trabalho foram as seguintes bases de dados: *ACM Digital* <https://dl.acm.org/>; *Google Scholar* <https://www.scholar.google.com.br/>; *IEEE Xplore* <https://ieeexplore.ieee.org/>; *Scopus* <https://www.scopus.com/>; *Science Direct* <https://www.sciencedirect.com>; *Springer* <https://link.springer.com/>. Além disso, foram incluídos manualmente alguns trabalhos que não foram retornados pela *string* de busca.

Os termos aplicados nas bases de dados foram: "*digital agriculture*", "*linked open data*", "*semantic web*", "*digital humanities*", "*fair data principles*", "*agriculture* 4.0" e suas

[13] https://parsif.al/

correspondências em português. Com base nos termos definidos anteriormente, eles foram relacionados gerando a seguinte *string* de busca:

((digital AND agriculture) OR (agricultura AND digital) OR (agriculture AND 4.0) OR (agricultura AND 4.0)) AND ((semantic AND web) OR (web AND semantica)) AND ((fair AND data AND principles) OR (principios AND fair)) AND ((digital AND humanities) OR (humanidades AND digitais)) AND ((linked AND open AND data) OR (dados AND interligados)). Excepcionalmente, durante a busca na plataforma Science Direct, a *string* original precisou ser reelaborada devido a limitação de 8 operadores por busca.

Devido a isso, optou-se pela eliminação dos termos em português da *string*, bem como os operadores que conectavam as palavras dos termos, por exemplo: ao invés de "digital AND agriculture", foi utilizada, simplesmente "digital agriculture". Nesta 55lataforma, a *string* foi a seguinte: ((digital agriculture) OR (agriculture 4.0)) AND (semantic web) AND (fair data principles) AND (digital humanities) AND (linked open data).

Tabela 5 – Resultados retornados a partir da busca com a string nas bases de dados

BASE DE DADOS	**NÚMERO DE ARTIGOS – n (%)**
Google Scholar	11 (18,9)
IEEE Xplore	0 (0)
Scopus	5 (8,6)
ScienceDirect	27 (46,5)
Springer	0 (0)
ACM Digital	11 (18,9)
Manual	4 (6,8)

Os critérios de exclusão aplicados para os documentos coletados nas bases utilizando as *strings* supracitadas foram, respectivamente: (a) estudos duplicados (2); (b)estudos anteriores a 2012 (15); (c) estudos em qualquer idioma que não seja inglês ou português (1); (d) literatura cinza (manuais e relatórios) (10) e (e) estudos fora do escopo desta pesquisa. (22).

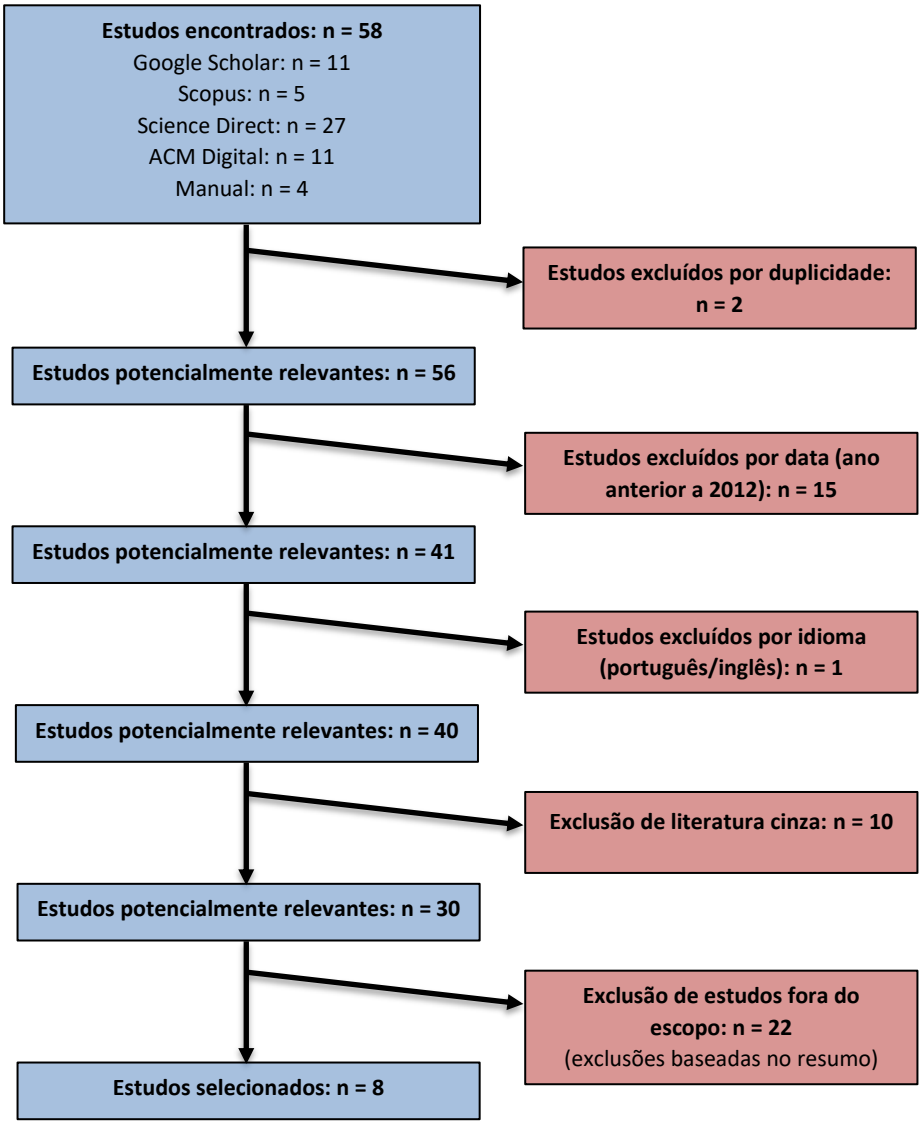
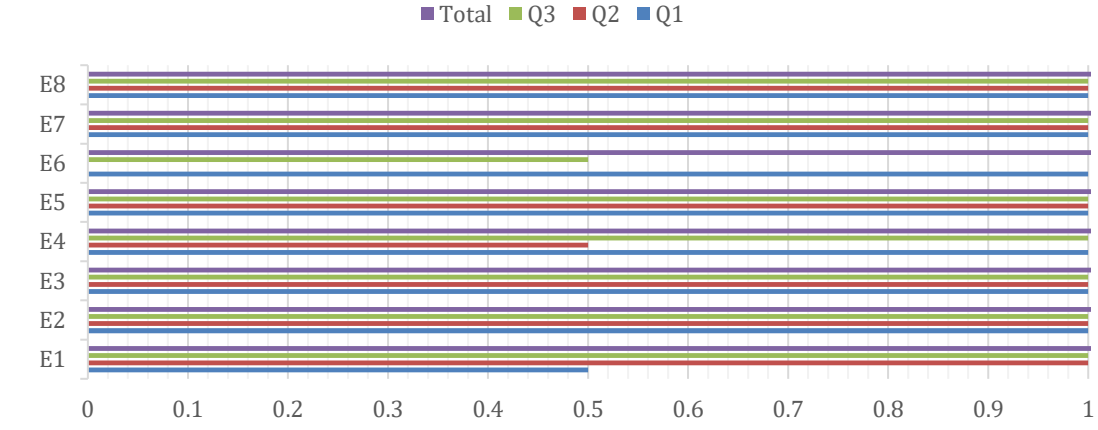


Figura 12 - Fluxograma do processo de seleção de estudos (Elaborado pela autora).

Os critérios de inclusão, por sua vez, foram definidos em: (a) estudos em que os dados interligados são aplicados em problemas das humanidades digitais; (b) estudos que envolvem humanidades digitais e agricultura digital; (c) estudos que envolvem web semântica e humanidades digitais.

Uma vez selecionados, os trabalhos (E) passaram por 3 critérios de qualidade que levam em consideração a descrição das limitações do estudo, a avaliação de um experimento computacionalmente bem descrito e de um objetivo igualmente claro, são eles:

- Q1: As limitações do estudo estão descritas?
- Q2: O estudo realizou um experimento computacionalmente bem descrito para avaliar a proposta?
- Q3: O objetivo da pesquisa está claramente descrito?



	E1	E2	E3	E4	E5	E6	E7	E8
Total	2.5	3	3	2.5	3	1.5	3	3
Q3	1	1	1	1	1	0.5	1	1
Q2	1	1	1	0.5	1	0	1	1
Q1	0.5	1	1	1	1	1	1	1

Figura 13 – Critérios de qualidade aplicados em cada um dos trabalhos selecionados (Elaborado pela autora).

Os trabalhos selecionados (E) passaram pelos critérios de qualidade e poderiam pontuar: 0 (em não conformidade), 0,5 (parcialmente em conformidade), 1 (totalmente em conformidade). Dessa maneira, cada trabalho poderia pontuar no máximo 3 e no mínimo 0.

3.1.3. Análise e discussão dos artigos selecionados

Jayaraman (2015) desenvolveu o Phenonet, um caso de uso da plataforma OpenIoT[14] (Internet das Coisas Aberta) com o objetivo de, além de propor uma solução para os desafios encontrados pela agricultura industrial, dentre eles a segurança alimentar, demonstrar como a agricultura digital pode se beneficiar com iniciativas assemelhadas.

Além dos procedimentos de integração de sensores, foi desenvolvida a ontologia denominada Phenonet, com o intuito de disponibilizar os dados capturados pelos sensores na nuvem RDF, como podemos observar na seguinte figura.

[14] Plataforma baseada na tecnologia IoT (Internet das coisas), desenvolvida em conjunto com o consórcio EU FP7 OpenIoT.

Fonte: (JAYARAMAN, 2015)

```
<rdf:Description rdf:about="http://sensordb.csiro.au/
  phenonet/experiment/kirkegaard-and-danish/plot/7002">
    <phenonet:plotBlock rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">3</phenonet:plotBlock>
    <phenonet:plotColumn rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">7</phenonet:plotColumn>
    <phenonet:plotRow rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">2</phenonet:plotRow>
    <dcterms:modified rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">2014-05-23</dcterms:
      modified>
    <phenonet:withinSite rdf:resource="http://sensordb.
      csiro.au/id/site/ges-creek-range"/>
    <dcterms:created rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">2014-05-23</dcterms:
      created>
    <dcterms:identifier rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">537f549984ae3ab43a06dc67</
      dcterms:identifier>
    <phenonet:plotID rdf:datatype="http://www.w3.org
      /2001/XMLSchema#string">7002</phenonet:plotID>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/
      XMLSchema#string">Revenue_C07R02</rdfs:label>
    <rdf:type rdf:resource="http://sensordb.csiro.au/
      ontology/phenonet#Plot"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/
      XMLSchema#string">imidacloprid + impact</rdfs:
      comment>
  </rdf:Description>
```

Figura 14 – Descrição do gráfico Phenonet em RDF.

NGO et al. (2020), por sua vez, apresenta uma arquitetura para a *Ontology-Based Knowledge* (OAK), uma ontologia criada para representar o conhecimento extraído de tarefas de mineração de dados na agricultura digital. Na arquitetura proposta estão destacados três componentes chave: *Knowledge Miner* (Minerador de conhecimento), *Knowledge Wrapper* (Concentrador de conhecimento) e *Knowledge Management System* (Sistema de gestão de conhecimento).

O Minerador de Conhecimento é o componente responsável por extrair o conhecimento dos dados. O Concentrador de conhecimento fica responsável por transformar o conhecimento extraído pelo Minerador de Conhecimento para o sistema de gestão de conhecimento, coletando resultados da mineração, identificando o tipo de tarefa e algoritmo de mineração de dados, bem como listando conceitos, funções de transformações correlativas e estados. Por fim, é gerada uma representação *Knowledge Map*, convertendo-a em triplas RDF, que serão então importadas pelo sistema de gestão de conhecimento.

O sistema de gestão de conhecimento consiste em um servidor de banco de dados gráfico que suporta o armazenamento de triplas RDF e o protocolo SPARQL para a realização de consultas (NGO et al., 2020).

Moreira et al. (2015) acreditam no potencial das tecnologias da *Web* Semântica na contribuição para a padronização da representação dos recursos provenientes da agricultura. Neste sentido, os autores fizeram um levantamento sobre as tecnologias semânticas utilizadas no *International Information System of the Agricultural Science and Technology* (AGRIS).

O AGRIS[15] é resultado de uma iniciativa liderada pela ONU nos anos 70 com o intuito de viabilizar uma ampla cooperação para o acesso compartilhado em ciência e tecnologia agrícola. Cada registro indexado no sistema do AGRIS é identificado por um AGRIS *Record Number* (ARN), ou seja, um URI. Os metadados dos conteúdos são formatados em arquivos XML e armazenados na base de dados AGRIS XML. Para possibilitar as ligações semânticas dos recursos indexados, os arquivos XML são convertidos para o formato RDF, que são então armazenados na base AGRIS *Record*. A definição de vocabulários é feita nesta mesma etapa e são utilizados a ontologia *Friend of a Friend* (FOAF) (BRICKLEY & MILLER, 2007) além dos padrões *Dublin Core* (BAKER, 2000), por exemplo, um tesauro específico da agricultura, denominado AGROVOC (MOREIRA et al., 2015).

> "O AGROVOC é um vocabulário controlado agrícola multilíngue com mais de 32.000 conceitos traduzidos para mais de 20 idiomas, cobrindo assuntos na agricultura, pecuária, sistemas florestais e pesca. Atuando como um "*backbone*", o AGROVOC permite realizar interligações dos recursos AGRIS com bases de dados externas, tais como *Dbpedia, World Banck, Google Custom Search API, Nature Open Search, FAO Geopolitical Ontology – Country profiles, Global Biodiversity Information Facility, International Food Policy Research Institute* – IFPRI, FAO *Fisheries and Aquaculture fach sheets API e Bioversity International*." (MOREIRA et al., 2015)

Para consultar os recursos indexados foi desenvolvida uma plataforma baseada na tecnologia SPARQL, denominada *Open* AGRIS. Trata-se de um *webservice* integrado ao AGRIS que possibilita a visualização de gráficos, estatísticas, dados do DBPedia, indicadores, entre outros (MOREIRA et al. 2015).

Subirats-Coll (2022), por sua vez, descreve o presente status do AGROVOC e como foi o processo para que este se tornasse o *Linked Data Concept Hub* para alimentação e agricultura,

[15] http://agris.fao.org/

que surgiu com o objetivo de aumentar a acessibilidade e a visibilidade dos produtos de pesquisa nos países membros da *Food and Agriculture Organization* (FAO).

O AGROVOC teve sua primeira publicação realizada pela FAO nos anos de 1980 para descrever documentos e outras informações em uma linguagem controlada para indexação e pesquisa. Desde então, o vocabulário evolui de catálogos impressos para a tecnologia de *web* semântica (SUBIRATS-COLL, 2022).

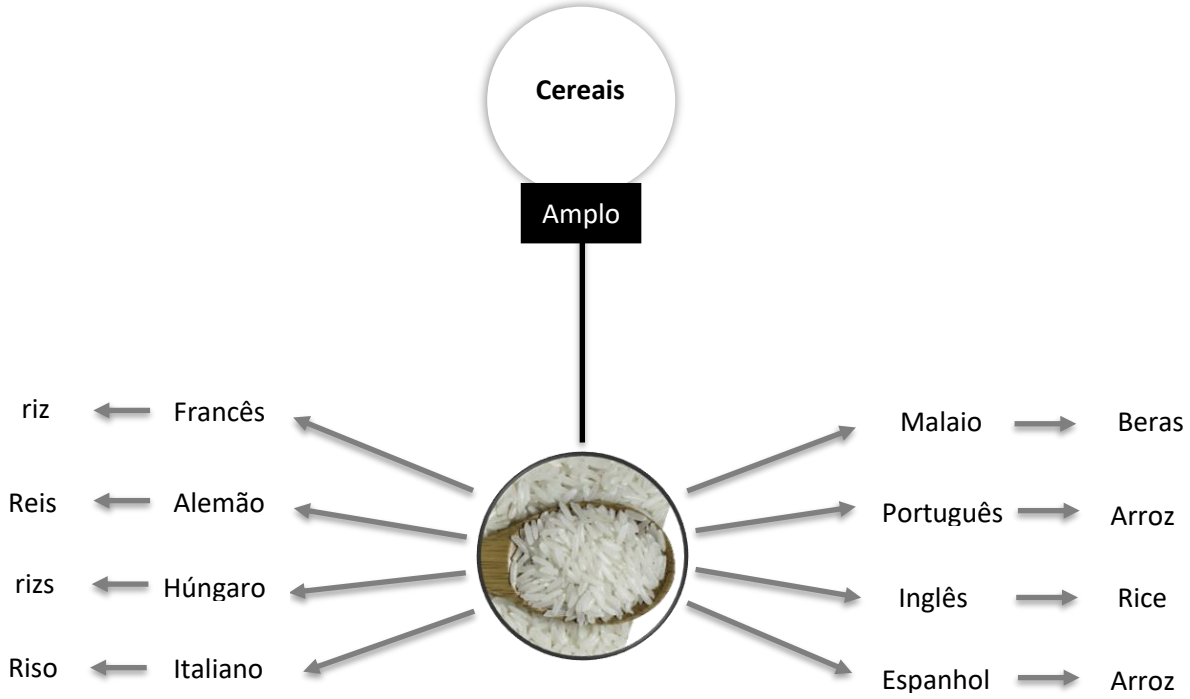


Figura 15 - Modelo conceitual do AGROVOC

Com um conteúdo acessível para humanos e máquinas, o AGROVOC é atualmente o tesauro mais utilizado na agricultura e é um dos recursos mais populares para indexar e marcar recursos de informação em países de baixa renda. Nele, conceitos e termos podem ser pesquisados manualmente navegando por uma hierarquia, pode ser baixado como um conjunto de dados RDF, acessado por meio de serviços da *web* ou pesquisado por meio de consultas SPARQL, usando um *endpoint* SPARQL público.

O autor pontua ainda que as parcerias são a maior sustentação deste trabalho, uma vez que o AGROVOC conta com uma rede internacional de editores e instituições parceiras. Além disso, o fortalecimento dos alinhamentos com outros tesauros é de grande importância para melhorar a coerência e coesão nos sistemas de conhecimento, o que consequentemente melhorará as abordagens de big data facilitando na tomada de decisão e prestação de contas,

bem como um compartilhamento mais eficaz de conhecimento e tecnologias em todo o mundo (SUBIRATS-COLL, 2022).

Outra iniciativa recente é o trabalho de Turki (2021) que apresenta quatro aspectos que permitem que o *Wikidata* – uma base de conhecimento colaborativa interdisciplinar, multilíngue e aberta de mais de 90 milhões de entidades conectadas por bem mais de um bilhão de relacionamentos – sirva como base de conhecimento para informações gerais sobre a pandemia do COVID-19, são eles: modelo de dados flexível, seus recursos multilíngues, seu alinhamento a vários bancos de dados externos e sua organização multidisciplinar. Embora não seja um trabalho no contexto da agricultura digital como os mencionados anteriormente até aqui, as análises realizadas a partir dos quatro aspectos inicialmente apresentados demonstram importantes conclusões sobre a importância da colaboração para o enriquecimento de bases de conhecimento de uma maneira geral.

Também em uma direção analítica e propositiva, o trabalho de Zhang & Yang (2018) tem o objetivo de transformar gradualmente a informação de dados públicos em uma nova propriedade pública social, construindo assim, uma futura sociedade do conhecimento através do desenvolvimento da *web* semântica. Um estudo sobre reutilização de dados públicos e propriedade de informações francesas é realizado, fomentando um papel de "partilha". O autor defende que a web semântica é capaz de realizar o reuso das informações dos dados públicos e que os dados interligados, atualmente, não são apenas as estruturas de informação de links de dados, mas se tornaram a principal maneira de desenvolver conhecimento e modos de inovações econômicas.

Em um trabalho de grande relevância para as humanidades digitais, no campo da história socioeconômica, Hoekstra (2018) apresenta a plataforma dataLegend, que possibilita aos pesquisadores a publicação de seus conjuntos de dados, além de vinculá-los a vocabulários existentes e a outros conjuntos de dados, contribuindo para uma coleção crescente de conjuntos de dados interligados.

Os elementos chave da dataLegend são o Qber – aplicação web que possibilita que pesquisadores disponibilizem, convertam e conectem dados "limpos" à datasets e vocabulários existentes na plataforma sem comprometer o detalhe e a heterogeneidade dos dados originais – e o grlc, um servidor que possibilita métodos de acesso à dados interligados (SPARQL, fragmentos de dados vinculados, RDFa, etc). A proposta da dataLegend é proporcionar um ecossistema viável para *Linked Humanities Data*. O autor destaca como um dos maiores

desafios a transformação dos dados, que estão, em sua maioria, em diferentes fontes tabulares como planilhas, bancos de dados e arquivos CSV. Além disso, o autor pontua que algumas melhorias ainda podem ser realizadas na ferramenta, como por exemplo a implementação de URIs.

3.1.4. Comparativos

Os trabalhos que receberam a pontuação máxima possuem um objetivo bem definido e tiveram uma clara descrição das limitações técnicas, bem como dos desafios do desenvolvimento. Além disso, apresentaram os experimentos que sustentam a proposta de maneira detalhada.

NGO et al. (2020), por exemplo, apresentaram pontos importantes sobre a descoberta de conhecimento na agricultura na última década e destacaram principalmente os estudos que geraram os *datasets* de solos que, dentre outras, possibilitam o monitoramento de características dos solos sob efeitos de outros fatores. Além disso, destacaram as diversas abordagens existentes para a construção de mapas de conhecimento e a melhor escolha para possibilitar a conversão para triplas RDF.

É importante ressaltar que todos os trabalhos que alcançaram a pontuação máxima possuem similaridades com esta presente dissertação, interseccionando principalmente a agricultura digital com a geração de dados triplificados.

Nossa revisão completa da literatura revelou que a relação que existem pontos de entrelaçamento entre sistemas agrícolas digitais, dados de solos e os princípios FAIR e apesar do reconhecermos seu valor, os estudos sobre o tema ainda são raros e estão em sua infância e por esses motivos acreditamos que essa dissertação poderá contribuir para sanar algumas das lacunas discutidas até aqui.

4. INFRAESTRUTURA DE TRIPILICAÇÃO DE DADOS DE SOLOS

4.1. *OpenSoils*

Diante dos problemas expostos, propôs-se um estudo, no contexto das Humanidades Digitais, sobre os princípios do movimento *Linked Open Data* e dos Princípios FAIR. Para isso, utilizamos o repositório de dados do *OpenSoils* como caso de estudo para a aplicação e exemplificação das tecnologias de *Web* Semântica, realizando experimento de triplificação de dados através de *workflows* descritos por Oliveira *et al.* (2021), a fim de atendermos aos princípios do movimento LOD.

Resumidamente, temos que o *OpenSoils* é uma infraestrutura computacional aberta, elástica, distribuída, multiusuário, multicamada e orientada para armazenar dados primários, secundários de solos e sua proveniência (DA CRUZ et al., 2018).

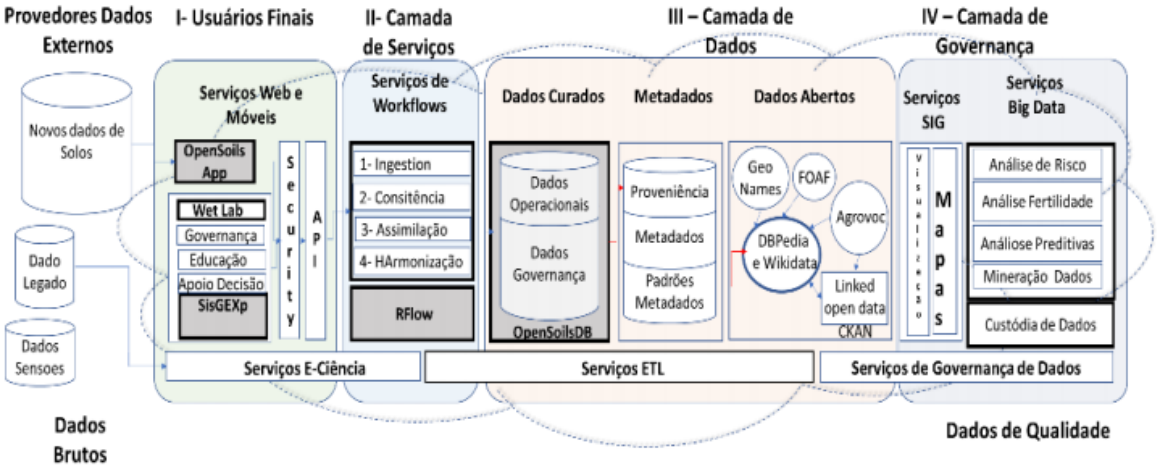


Figura 16 - Esquema de camadas da plataforma *OpenSoils* (DA CRUZ et al., 2018)

O objetivo da plataforma é oferecer melhorias nas condições e na organização do armazenamento de dados de levantamento de solos, bem como fornecer aos profissionais e pesquisadores da área uma alternativa ao modo de colheita de dados em campo.

A base de dados relacional do *OpenSoils*, hospedada em um serviço de nuvem da Amazon AWS, conta atualmente com cerca de 850 mil registros, sendo mais de 9 mil observações (perfis), com cerca de 32 mil horizontes.

O *schema* relacional[16], é composto por 46 tabelas. É possível visualizar que a partir da definição de um projeto na plataforma, torna-se possível a definição de seus relevos, que por sua vez possuem uma descrição geral. As descrições gerais podem ter uma ou mais observações e cada uma das observações podem possuir um ou mais horizontes. Por fim, os horizontes podem possuir ou não as informações de propriedade química, física, morfológica, análise de pasta saturada, dentre outras. Além disso, é possível correlacionar os dados produzidos por processos observacionais de campo com os dados produzidos em laboratórios (DE OLIVEIRA, 2021b).

4.2. *Workflow ETLH* de triplificação de dados pedológicos

A triplificação dos dados do *OpenSoils* possibilitará, além da disponibilização na *Web* de Dados com a integração total com as diferentes bases de dados já disponíveis, a diminuição de gargalos relacionados à alta dispersão e baixa integração de dados que resultam em pouca transparência e limitadas informações sobre proveniência, bem como o ganho em acessibilidade, interoperabilidade e o reuso de dados pedológicos.

Entretanto, para que seja possível, faz-se necessária a construção de um processo de automação que possa realizar a extração dos dados, realizar as transformações e adaptações necessárias que possibilitem a geração de triplas que possam ser, em seguida, carregadas em um repositório compatível com RDF.

As investigações preliminares, assim como no trabalho de Hoekstra (2018), tem indícios de que um dos grandes desafios é a combinação de diferentes bases de dados tabulares para o refinamento dos dados.

A carga dos dados legados para os repositórios de triplas, após sua transformação, está baseada em *workflows* ETLH, originalmente descritos por Cruz et al. (2021). Baseados nos processos de Extração, Transformação e Carga (ETL) que são utilizados para a extração de dados de uma base dados, sendo também responsável pelo processamento, modificação e inserção desses dados em uma outra base de dados (FERREIRA, 2010), os *workflows* ETLH engloba além da carga de dados brutos, a análise de dados e suas estruturas, a definição de modelos semânticos e a harmonização e identificação de dados e *datasets*.

[16] https://drive.google.com/file/d/1FCKy9lyp9uIuper4HVD0HHzajXZvb0lx/view?usp=sharing

A Figura 17 ilustra o relacionamento entre os *workflows* ETLH e de FAIRificação de dados. Cruz et al. (2021) destacam que a plataforma *OpenSoils* atende apenas em parte ao menos uma ou mais subdivisões de cada um dos princípios FAIR e citam pontos de melhoria, como por exemplo a adequação gradual aos princípios FAIR, FAIRificação de dados e a interligação dos dados na *web* de dados, possibilitando a interligação com diferentes bases de dados já disponíveis.

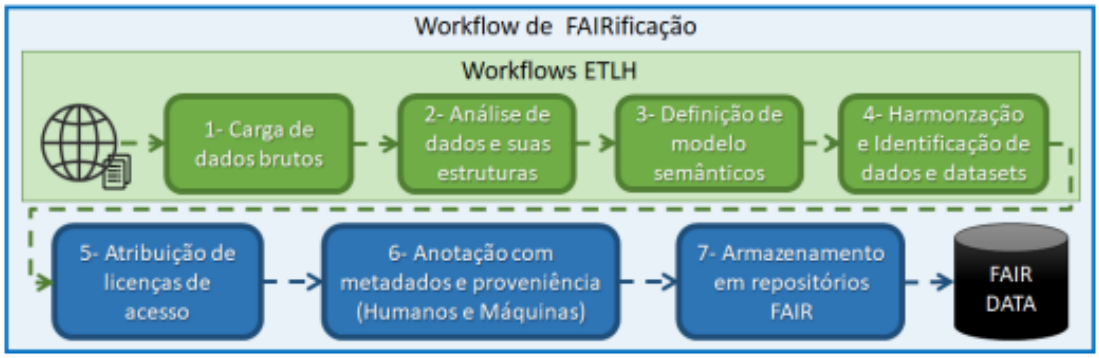


Figura 17 – Representação conceitual das etapas do processo ETLH e FIRiri e FAIRificação de dados (Fonte: DE OLIVEIRA, 2021)

Para a realização da triplificação, desenvolvemos e utilizamos *workflows* de Extração, Transformação, Carga e Harmonização (*Extract, Transform, Load, Harmonization* – ETLH), construídos a partir da ferramenta *Pentaho Data Integration* (PDI). O ETL é um processo oriundo da área de *Business Inteligence,* que é utilizado para a extração de dados de uma base dados, sendo também responsável pelo processamento, modificação e inserção desses dados em uma outra base de dados (FERREIRA, 2010).

É importante ressaltar que o PDI não apresenta, por padrão, tarefas que possibilitam a transformação de dados em triplas RDF. Entretanto, por ser um software de código aberto, a ferramenta possibilita a instalação de plugins que são capazes de personalizar e tornar a ferramenta ainda mais robusta. Para realizar a triplificação de dados, foram utilizados os Plugins Jena[17], desenvolvidos em 2021 pela Evolved Binary e DeveXE como parte do Projeto OMEGA para os Arquivos Nacionais do governo do Reino Unido. O pacote dos Plugins Jena conta atualmente com 5 tarefas, sendo uma delas a responsável pela criação de modelos, uma responsável pela combinação desses modelos e uma capaz de realizar o merge entre os modelos. O pacote conta ainda com uma tarefa responsável pela serialização dos modelos em 3 diferentes

[17] https://github.com/nationalarchives/kettle-jena-plugins

formatos à escolha do usuário, sendo eles: Turtle, N3, N-Triples, e RDF/XML. Por fim, o pacote possui também uma tarefa capaz de realizar a validação dos modelos criados.

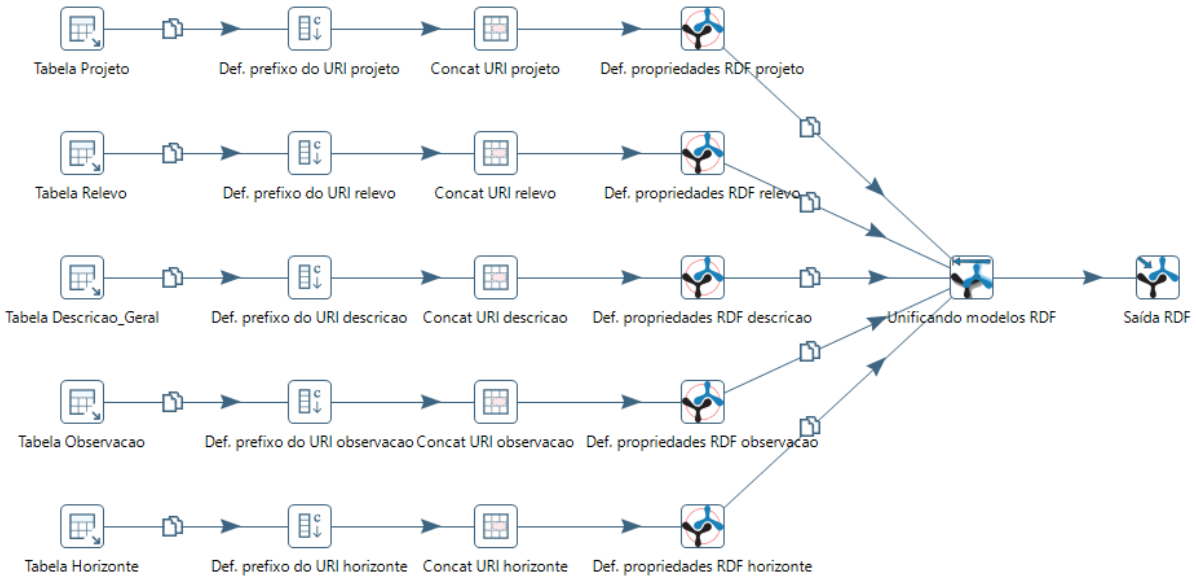


Figura 18 - *Workflow* ETLH com saída RDF

A primeira tarefa do *workflow* da Figura 18 é um fragmento que ilustra trechos do código responsável pela extração dos dados que iremos triplificar. Uma consulta SQL, como mostra a Figura 21, é realizada diretamente em cada uma das tabelas do banco de dados previamente conectado ao PDI.

Em seguida, na tarefa de transformação de inclusão de valores constantes, definimos o prefixo do URI, o qual concatenamos, na tarefa de transformação de concatenação, ao dado extraído na primeira tarefa. Por fim, utilizaremos as tarefas dos Plugins Jena, responsáveis pelas definições das propriedades RDF e pela serialização dos dados para a saída RDF.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:os="http://os.com/#">
 <rdf:Description rdf:about="http://opensoils.org/project_24">
  <os:identificador>24</os:identificador>
  <os:municipio>Eliseu Martins</os:municipio>
  <os:estado>PI</os:estado>
  <os:nome>Conjunto de dados do levantamento de reconhecimento 'Levantamento de Reconhecimento dos Solos do Núcleo Colonial de Gurguéia.'</os:nome>
  <os:aberto>1</os:aberto>
 </rdf:Description>
 <rdf:Description rdf:about="http://opensoils.org/project_36">
  <os:identificador>36</os:identificador>
  <os:nome>Conjunto de dados do levantamento exploratório 'PROJETO RADAMBRASIL - Levantamento de Recursos Naturais. Volume 5.'</os:nome>
  <os:aberto>1</os:aberto>
 </rdf:Description>
 <rdf:Description rdf:about="http://opensoils.org/project_12">
  <os:identificador>12</os:identificador>
  <os:nome>Conjunto de dados do 'Anais da III Reunião de Classificação, Correlação de Solos e Interpretação de Aptidão Agrícola.'</os:nome>
  <os:aberto>1</os:aberto>
 </rdf:Description>
 <rdf:Description rdf:about="http://opensoils.org/project_140">
  <os:identificador>140</os:identificador>
  <os:municipio>Rio de Janeiro</os:municipio>
  <os:estado>RJ</os:estado>
  <os:nome>Conjunto de dados do levantamento detalhado 'Projeto Parque Frei Veloso - Levantamento Detalhado dos Solos do Campus da Ilha do Fundão UFRJ'</os:nome>
  <os:aberto>1</os:aberto>
 </rdf:Description>
```

Figura 19 – Fragmento do arquivo de dados triplificados gerado pelo *workflow* ETLH.

A Figura 19, por sua vez, consiste na saída gerada pelos *workflows ETLH*, com os dados serializados em formato RDF/XML.

4.3. Repositórios de dados triplificados

Uma vez triplificados, armazenamos as triplas RDF em um banco de dados de grafos. É importante destacar que o processo ETL criado para a geração das triplas é um processo independente e destacado da criação da instância do banco de dados de grafos que descreveremos na presente seção. O arquivo gerado como saída dos *workflows* ETLH é carregado (Figura 20), manualmente, no projeto OpensoilsGraph, criado no Neo4j, plataforma de manipulação de bancos de dados de grafos.

Figura 20 - Importação do arquivo de triplas para o projeto no Neo4j

Os recursos do Neo4j incluem gerenciamento de armazenamento de dados, relatórios, marcação semântica e visualização de dados e permite integração facilitada com outros SGBDs. A escolha pelo Neo4j se deu por alguns motivos. Além de gratuito, as instâncias criadas já recebem um URI que pode ser acessado online juntamente com mais algumas credenciais de acesso. Adicionalmente, possui uma interface amigável, de compreensão simples e muito bem

documentada e que pode ser facilmente utilizada mesmo por usuários, produtores ou tomadores de decisão com poucos conhecimentos em Computação.

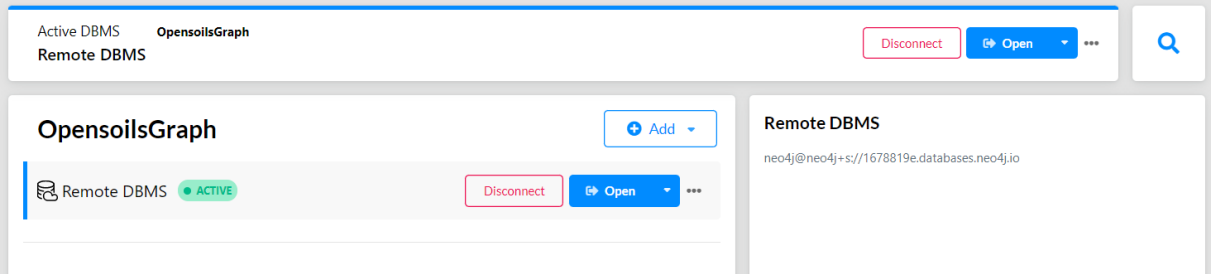


Figura 21 - Instância de banco de dados de grafos criada a partir do arquivo gerado pelos workflows

O projeto OpensoilsGraph foi criado e uma conexão remota estabelecida. Um URI[18] foi gerado pelo Neo4j na criação da instância. A partir dela, com as credenciais de login, é possível acessar o banco de dados tanto localmente, quanto online. A conexão pode ser estabelecida a partir do endereço[19] que provê os acessos aos *workspaces* gerados.

A partir daí, os dados passaram a ser manipulados a partir do Neo4j. É importante destacar que, traçando um paralelo entre um banco de dados relacional e um banco de dados de grafos, o que conhecemos como tabela nos bancos relacionais, é tratado como um nó no banco de grafos. Na Figura 22, é possível visualizar os nós que foram gerados.

[18] neo4j+s://1678819e.databases.neo4j.io
[19] https://workspace-preview.neo4j.io/workspace

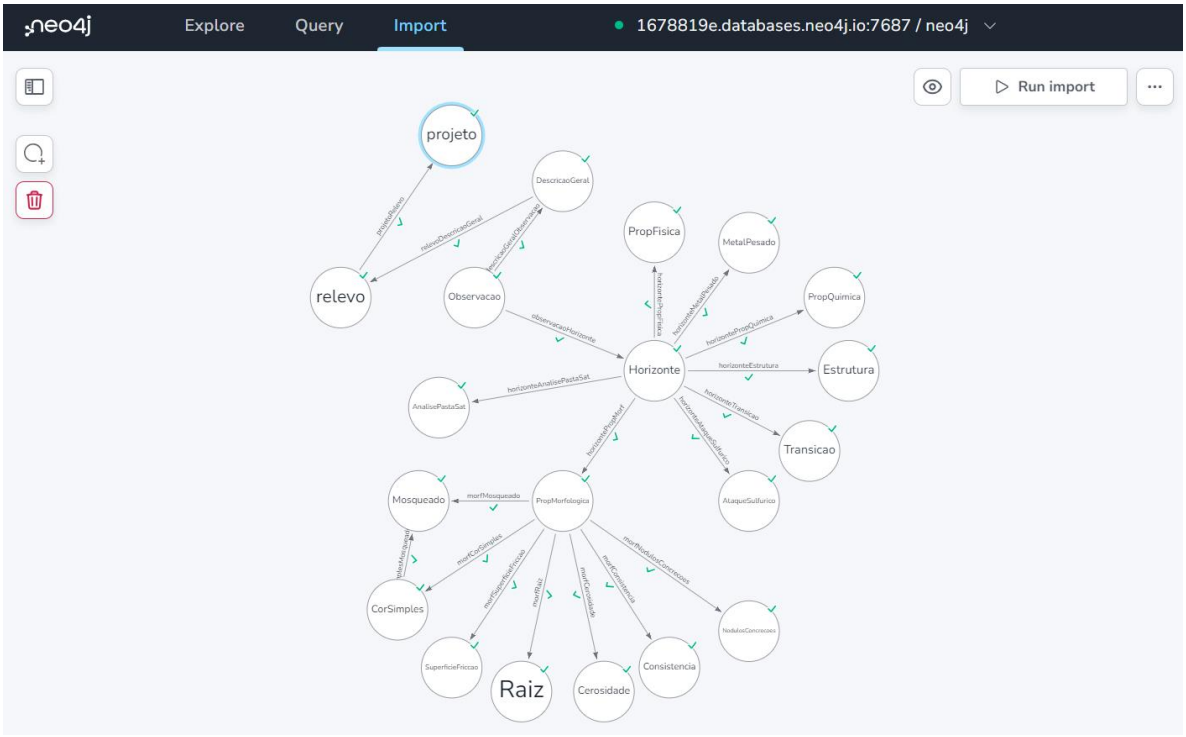


Figura 22 - Nós gerados a partir das tabelas do banco de dados relacional da plataforma *OpenSoils.*

O Neo4j utiliza a *Cypher* como linguagem de consulta, que assim como o SPARQL, é declarativa e possibilita consultas expressivas de dados em grafos de propriedades. Os relacionamentos entre as tabelas foram criados a partir da cláusula *MATCH*, responsável por especificar padrões que serão buscados no banco de dados.

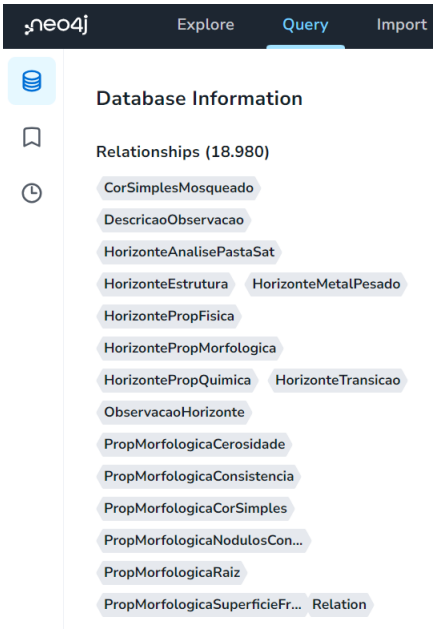


Figura 23 – Relacionamentos construídos a partir de cláusulas MATCH

A Figura 23 exibe as relações que foram construídas e salvas a partir da execução das cláusulas *MATCH*. A construção de uma *query Cypher* se dá pelo encadeamento de várias cláusulas.

As cláusulas podem ser de leitura e de gravação, onde as de leitura retornarão um resultado com base nos dados já existentes no banco e as de gravação determinarão alguma alteração real no banco, sendo com a inclusão ou a exclusão de dados, por exemplo. É importante citar que em uma mesma *query Cypher*, as cláusulas de leitura e gravação podem se revezar.

```
1 MATCH (d:projeto), (o:relevo)
2    WHERE d.idProjeto = o.idProjeto
3 CREATE (d)-[: ProjetoRelevo]→(o)
4 RETURN d,o
```

Figura 24 – Estrutura de uma query utilizando as cláusulas MATCH, WHERE e CREATE

Na Figura 24 temos a estrutura utilizada na criação das relações entre os nós. A cláusula MATCH na linha 1 referência as duas tabelas que serão relacionadas. Por sua vez, a cláusula WHERE na linha 2 identifica os campos responsáveis pelo relacionamento entre ambas as tabelas da primeira linha e por fim, com a cláusula CREATE na linha 3 é possível nomear o relacionamento e identificar a direção dele.

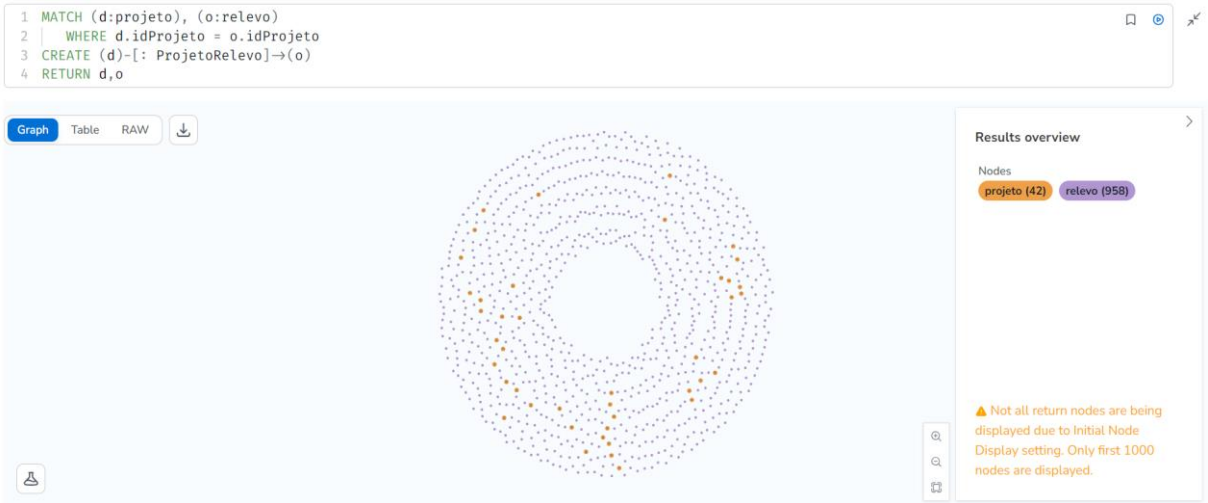


Figura 25 – Visualização gráfica da Query indicada na Figura 24

A cláusula RETURN definirá o que será exibido. No exemplo acima, ambos os dados das tabelas projeto e relevo foram exibidos no resultado. É possível, ao ampliar a visualização, expandir e observar a relação que se deu entre as duas tabelas.

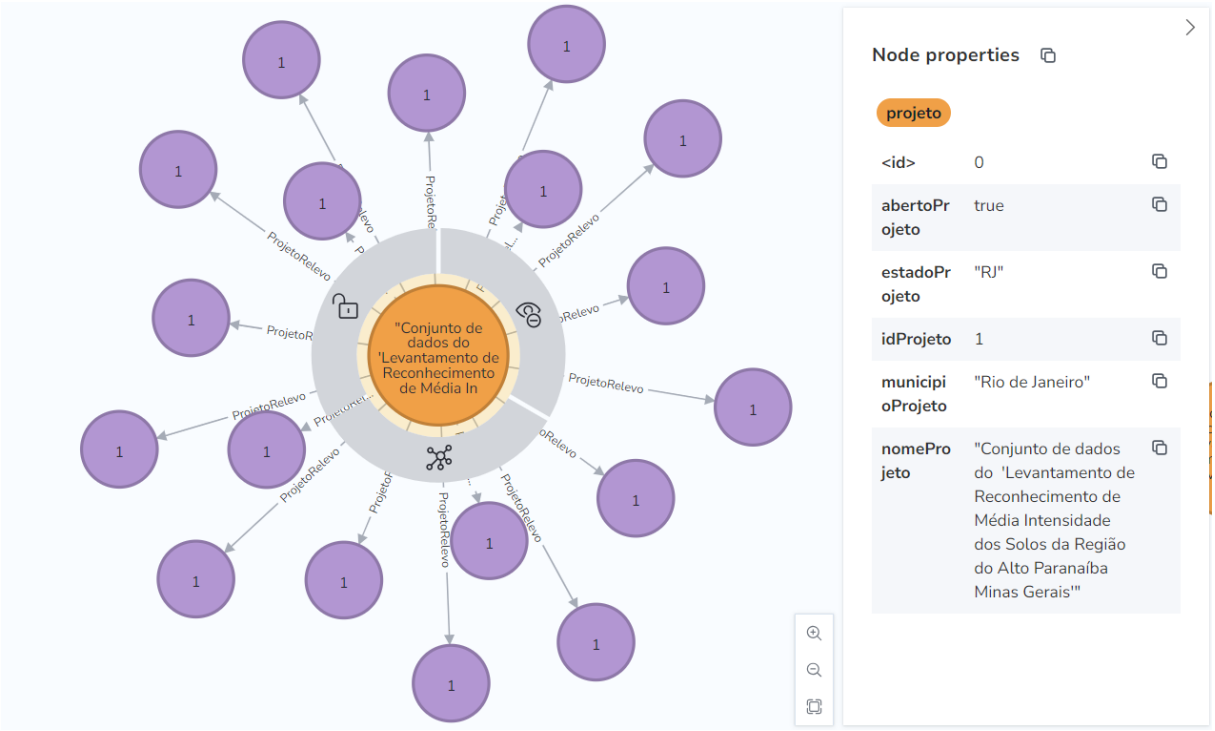


Figura 26 – Visualização da relação entre um projeto e seus tipos de relevo.

Na seção 5.2 detalharemos melhor o processo de mapeamento dos campos, bem como a construção dos demais relacionamentos utilizando como exemplo o projeto de Levantamento Detalhado de solos da área da Universidade Federal Rural do Rio de Janeiro, presente no banco de dados relacional da plataforma *OpenSoils*.

5. EXPERIMENTOS E DISCUSSÃO

5.1. Delimitação dos dados e da área do experimento

Devido a grande quantidade de dados da base de dados do *OpenSoils*, a realização de experimentos em todo a base seria inviável por questões de tempo. Logo, os experimentos realizados nesta seção são relativos ao projeto 90 presente naquela base de dados. O projeto 90 diz respeito ao mapeamento de solos da área da Fazendinha Agroecológica do Km 47[20] (Figura 27).

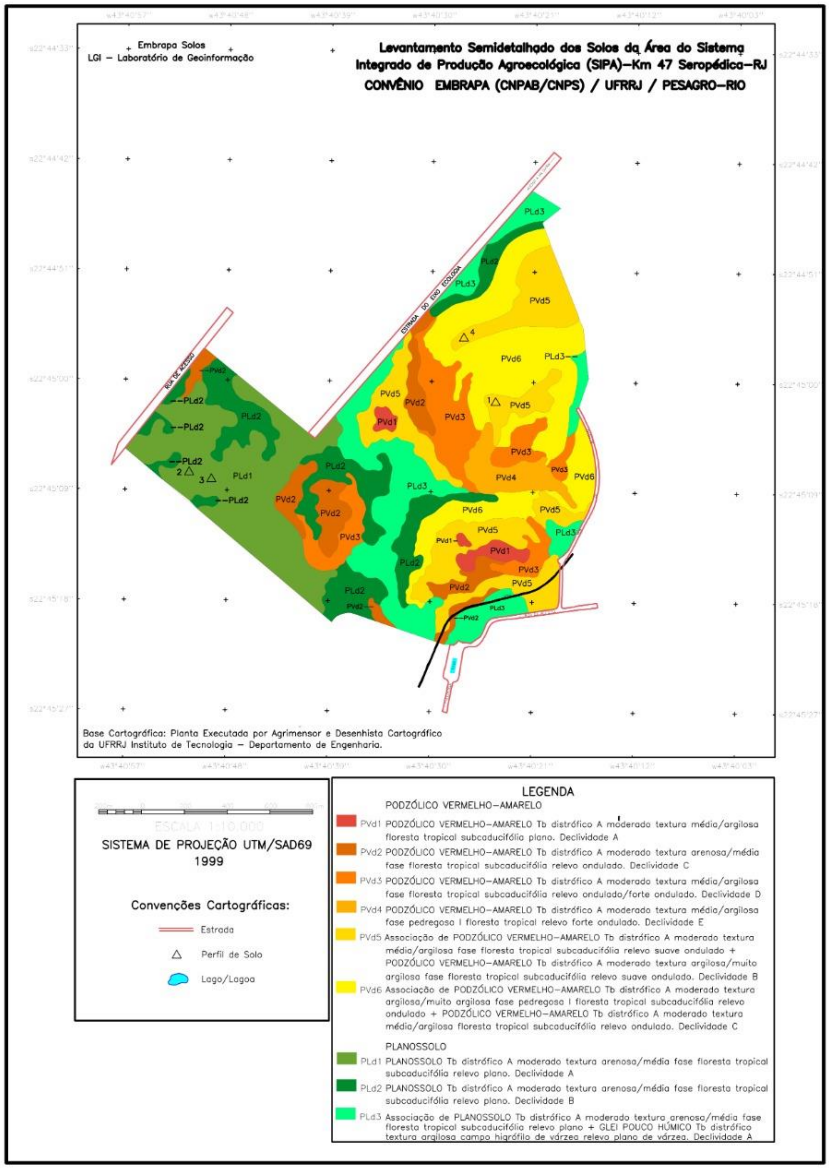


Figura 27 - Localização e tipos de solos da Fazendinha Agroecológica do Km 47

[20] https://institucional.ufrrj.br/fazendinha/historia/

A Fazendinha foi estabelecida em 1993, é uma parceria interinstitucional formada pelo Colégio Técnico da UFRRJ (CTUR), Empresa Brasileira de Pesquisa Agropecuária – Embrapa Agrobiologia, Empresa de Pesquisa Agropecuária do Estado do Rio de Janeiro (PESAGRO-RIO) e a Universidade Federal Rural do Rio de Janeiro (UFRRJ).

Através da Fazendinha pesquisadores e professores se reúnem para desenvolver o Sistema Integrado de Produção Agroecológica com vistas a ser um espaço de ensino, pesquisa e capacitação para o exercício da agroecologia e da agricultura orgânica em bases científicas, com base no enfoque sistêmico e na parceria interinstitucional. Logo, nossos experimentos computacionais optaram pela área da Fazendinha devido a esta ser uma ativo muito importante para a formação de novos alunos e pesquisadores no Brasil.

5.2. Construção dos *workflows ETLH* e integração com o *OpenSoils*

Antes mesmo da construção dos *workflows* responsáveis pela transformação dos dados da base relacional para dados triplificados, foi realizada a integração do *OpenSoils* com o Pentaho, possibilitando o acesso a base de dados para que as transformações fossem executadas. Para isso, uma nova conexão do tipo MySQL foi criada na ferramenta, com as informações de host, nome, porta, usuário e senha, como mostra a Figura 27.

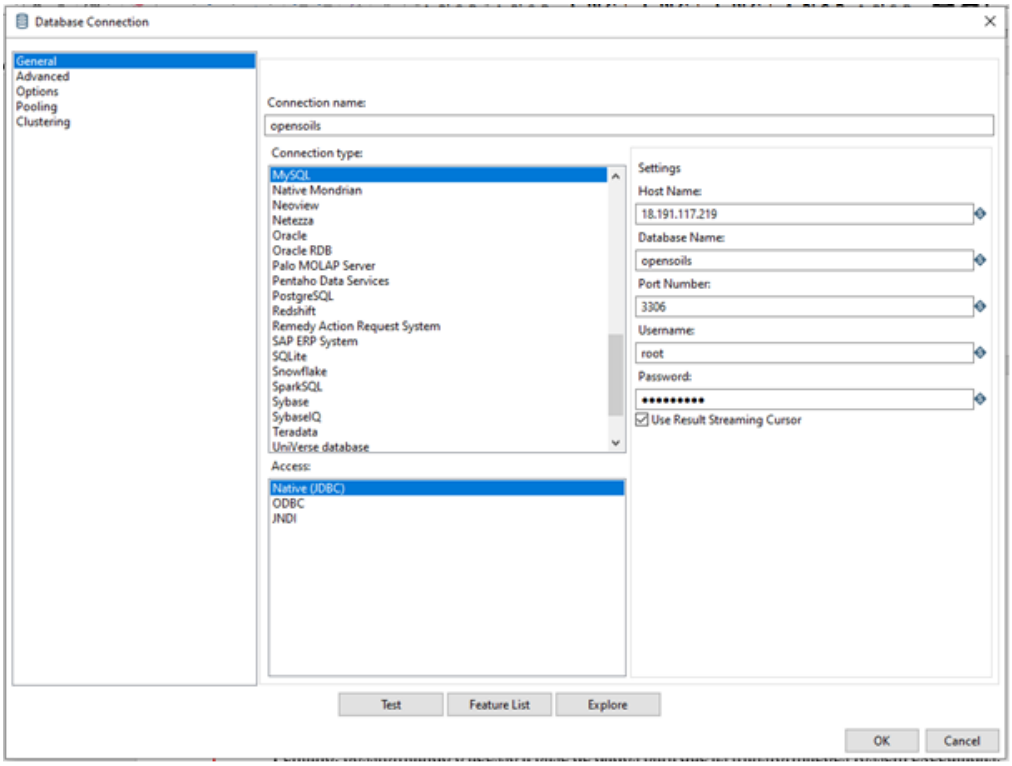


Figura 28 – Conexão do Pentaho com o banco de dados relacional do *Opensoils*.

A partir disso, o acesso às tabelas e aos dados tornou-se possível, possibilitando que as primeiras etapas do *workflow*, ou seja, a entrada de dados, pudesse ser iniciada.

A entrada de dados no *workflow*, por sua vez, foi realizada por um componente da ferramenta Pentaho que possibilita o acesso aos dados por consultas SQL. Com isso, utilizamos a conexão criada (Figura 27) para recuperar os dados de cada uma das tabelas, como mostra a Figura 28.

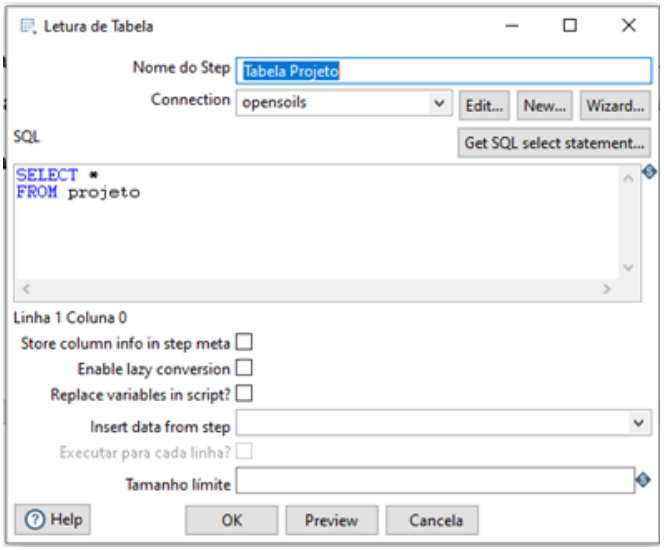


Figura 29 – Recuperando os dados da tabela “projeto”.

O passo seguinte foi a definição de um URI para os registros retornados. Na Figura 29 é possível visualizar que o URI escolhido para os dados da tabela projeto foi o "http://opensoils.org/projeto_", padrão que se repetirá para todas as demais tabelas.

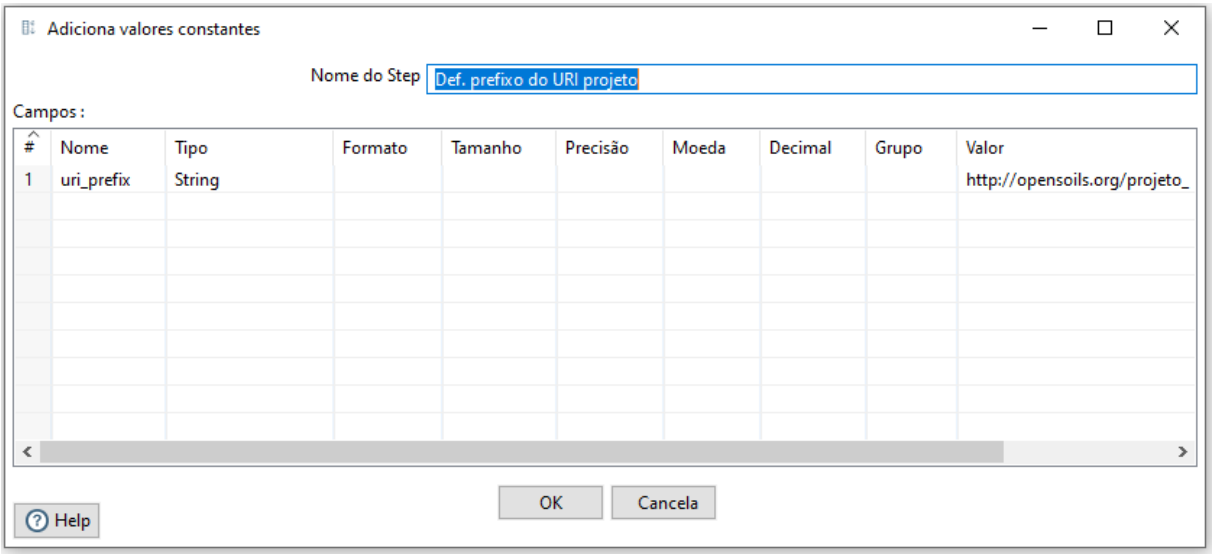


Figura 30 - Definindo o prefixo do URI.

Utilizamos então mais um componente do Pentaho para concatenar o URI definido com o ID dos projetos (Figura 30). É importante ressaltar que esse ID é oriundo de uma das colunas da tabela projeto, a qual recuperamos os dados no primeiro componente do presente *workflow* (Figura 28).

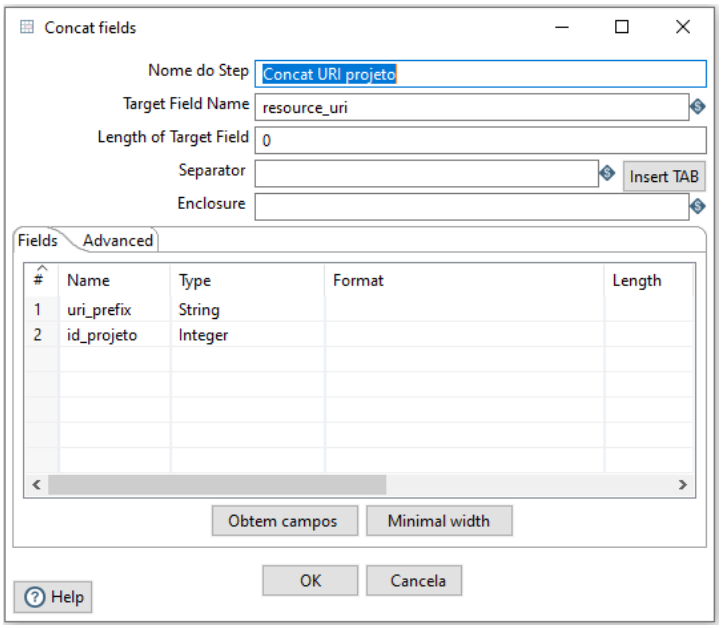


Figura 31 - Concatenando o URI com valor do identificador único do registro.

O seguinte passo já envolve um dos componentes do Plugin Jena, citado em 4.2. Esse componente fica responsável pela definição do modelo de saída. Nele são definidos prefixos, URIs e as propriedades RDF para os campos.

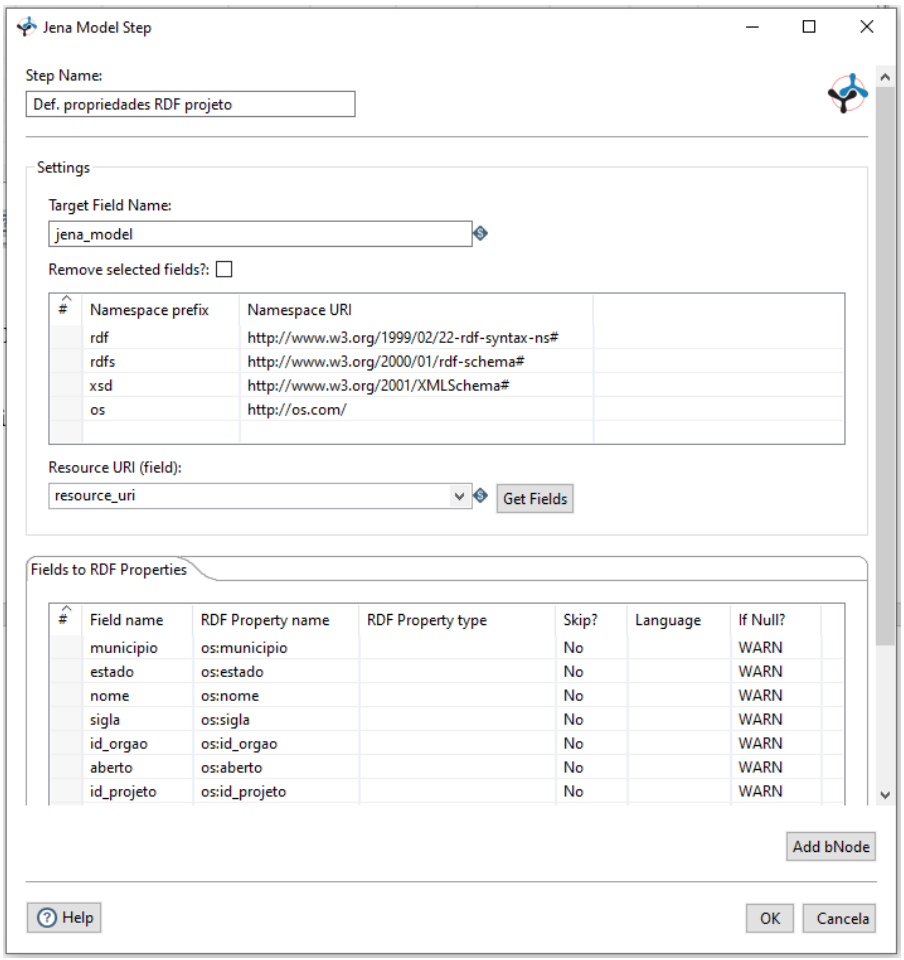


Figura 32 - Definindo propriedades RDF para os metadados.

Repetindo todo o processo para cada uma das tabelas do *OpenSoils*, ao final, utilizamos um segundo componente do Plugin Jena, responsável por unificar os modelos gerados por cada uma das tabelas em apenas um.

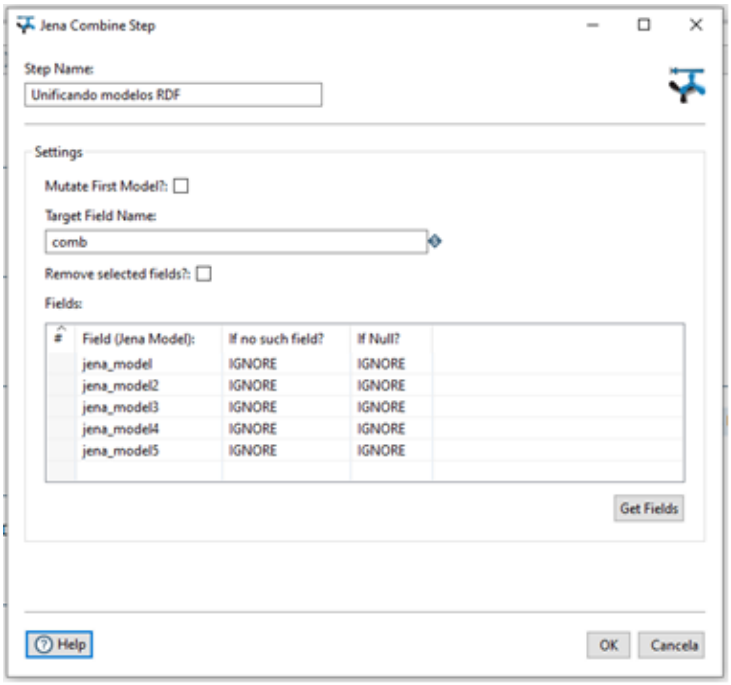


Figura 33 – Unificando os modelos RDF gerados a partir das tabelas.

Por fim, um último componente Jena será responsável por serializar os modelos unificados no componente anterior em um modelo único, gerando a saída dos dados já triplificados.

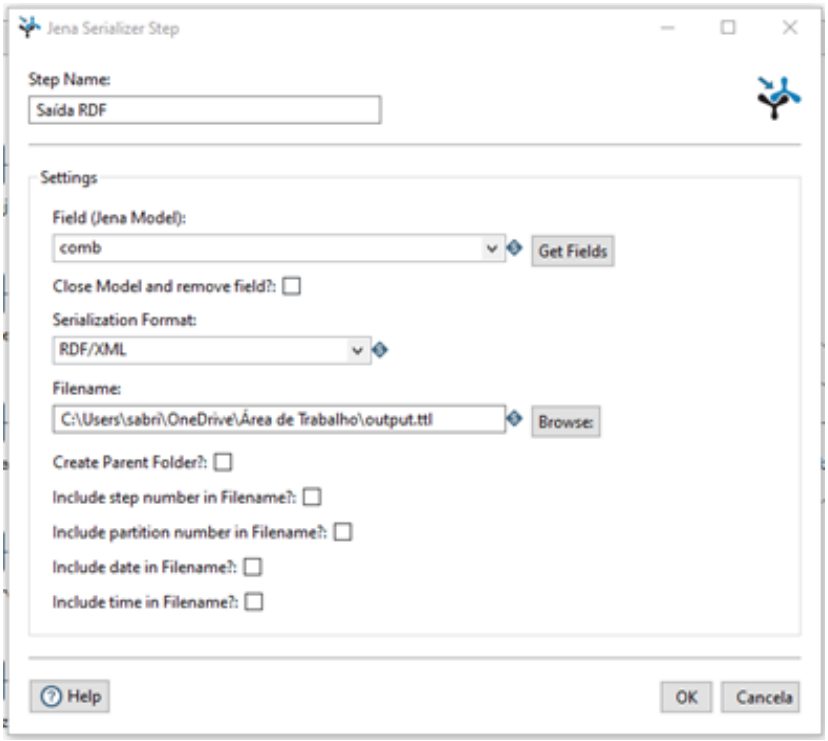


Figura 34 – Configurando a saída final dos dados triplificados.

O tempo de execução do *workflow* é variável de acordo com a máquina que o executa, a velocidade da internet que está sendo utilizada, entre outros fatores. A ferramenta, entretanto, oferece informações sobre o tempo de execução de cada uma das etapas, que também variam muito de acordo, principalmente, com a tabela do banco relacional que é tratada, uma vez que cada uma delas possui variações com relação a quantidade de registros.

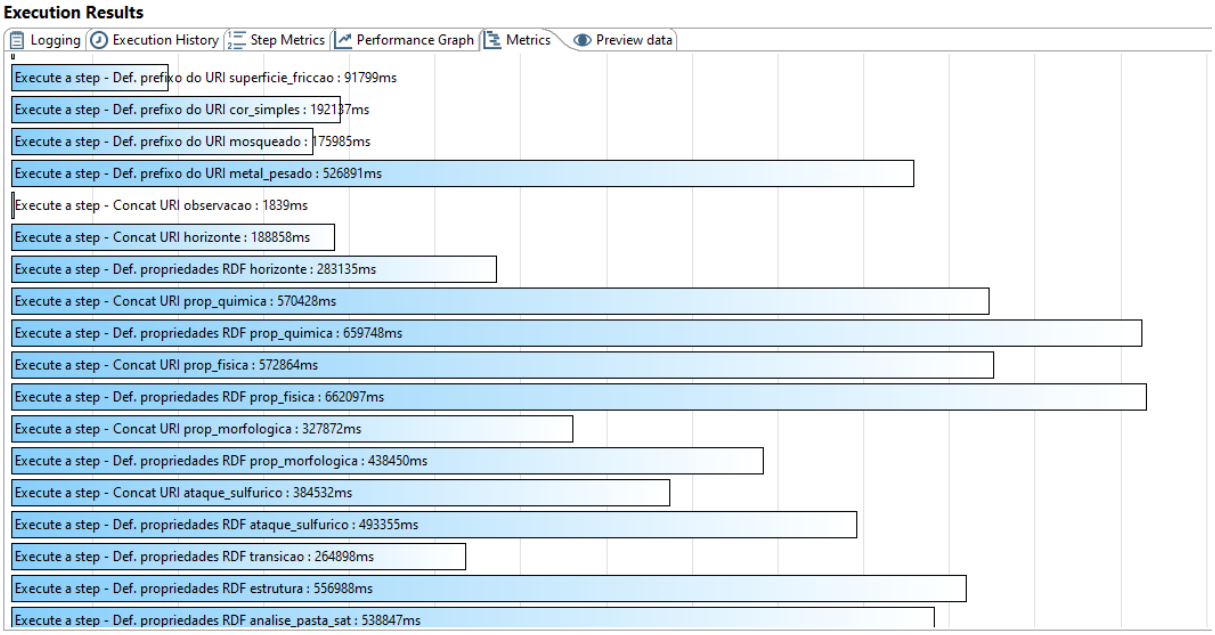


Figura 35 - Fragmento referente às métricas de execução do workflow ETLH.

5.3. Manipulando os dados triplificados no Neo4j

Como anteriormente demonstrado na seção 4.3, uma instância de banco de dados dentro do projeto OpenSoilsGraph foi criada no Neo4j. Um URI foi gerado pela própria ferramenta, possibilitando o acesso dos dados localmente e no *workspace* online da plataforma.

Uma vez carregados e estabelecidos todos os relacionamentos necessários entre os nós, escolhemos um dos projetos de levantamento de solos realizados na UFRRJ, denominado projeto de Levantamento Detalhado de solos da área da Fazendinha da Universidade Federal Rural do Rio de Janeiro[21], presente no banco de dados da plataforma *OpenSoils* e disponibilizado inicialmente a partir da digitalização do acervo, sendo uma das práticas iniciais, mais comuns e importantes das Humanidades Digitais. Inicialmente, executamos uma *query*

[21] https://www.alice.cnptia.embrapa.br/alice/handle/doc/336519

para retornar as informações iniciais referentes ao projeto, como pode ser visualizado na Figura 35.

Figura 36 - Query executada para retornar informações iniciais do projeto.

A próxima etapa consiste em retornar os relevos relativos a este projeto. Para isso executamos uma consulta que retornará os registros do nó de relevo conectados ao registro do nó de projeto. Na Figura 35 é possível visualizar pelas propriedades do nó que o identificador do projeto (idProjeto) é o número 90. Por isso, a partir daqui utilizaremos este identificador em nossas *querys* com o objetivo de otimizar o processo.

```
1 MATCH (p:projeto), (d:relevo)
2 WHERE p.idProjeto = d.idProjeto AND p.idProjeto = 90
3 RETURN p,d
```

Figura 37 - Query para retorno dos registros dos nós de projeto e relevo relacionados ao projeto 90.

Com os 6 relevos que são retornados, podemos visualizar além dos registros dos nós de projeto e relevo, o relacionamento entre eles, denominado “ProjetoRelevo”. Para cada relevo que possua como identificador do projeto o número 90, temos uma relação entre ambos os nós, como pode ser observado na Figura 37.

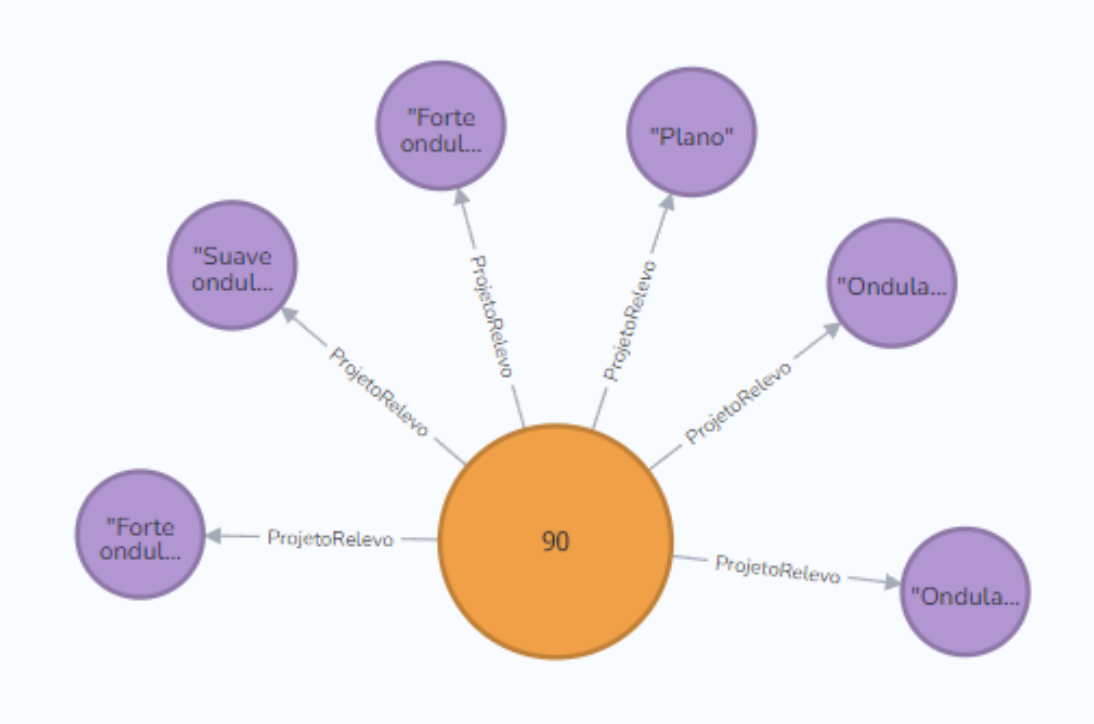


Figura 38 - Relacionamento entre os nós projeto e relevo referentes ao projeto 90

Um processo similar pode ser realizado para que possamos exibir os registros dos demais nós, uma vez que os relacionamentos já foram previamente determinamos a partir da execução das consultas com as cláusulas MATCH e CREATE. Por exemplo, os registros do nó de descrição geral armazenam informações mais detalhadas quanto aos relevos. Portanto, existe um relacionamento entre os nós relevo e descrição geral, o qual estabelecemos executando a *query* da Figura 38.

```
1 MATCH (r:relevo), (d:DescricaoGeral)
2   WHERE r.idRelevo = d.idRelevo
3 CREATE (r)-[: RelevoDescricaoGeral]→(d)
4 RETURN r,d
```

Figura 39 - Estabelecendo relacionamento entre os nós relevo e descrição geral.

A nó de observação, por sua vez, possui uma dupla relação, já que está diretamente ligado tanto ao nó de descrição geral quanto ao nó de projeto. Na Figura 39 é possível visualizar o relacionamento entre todos os nós citados até aqui.

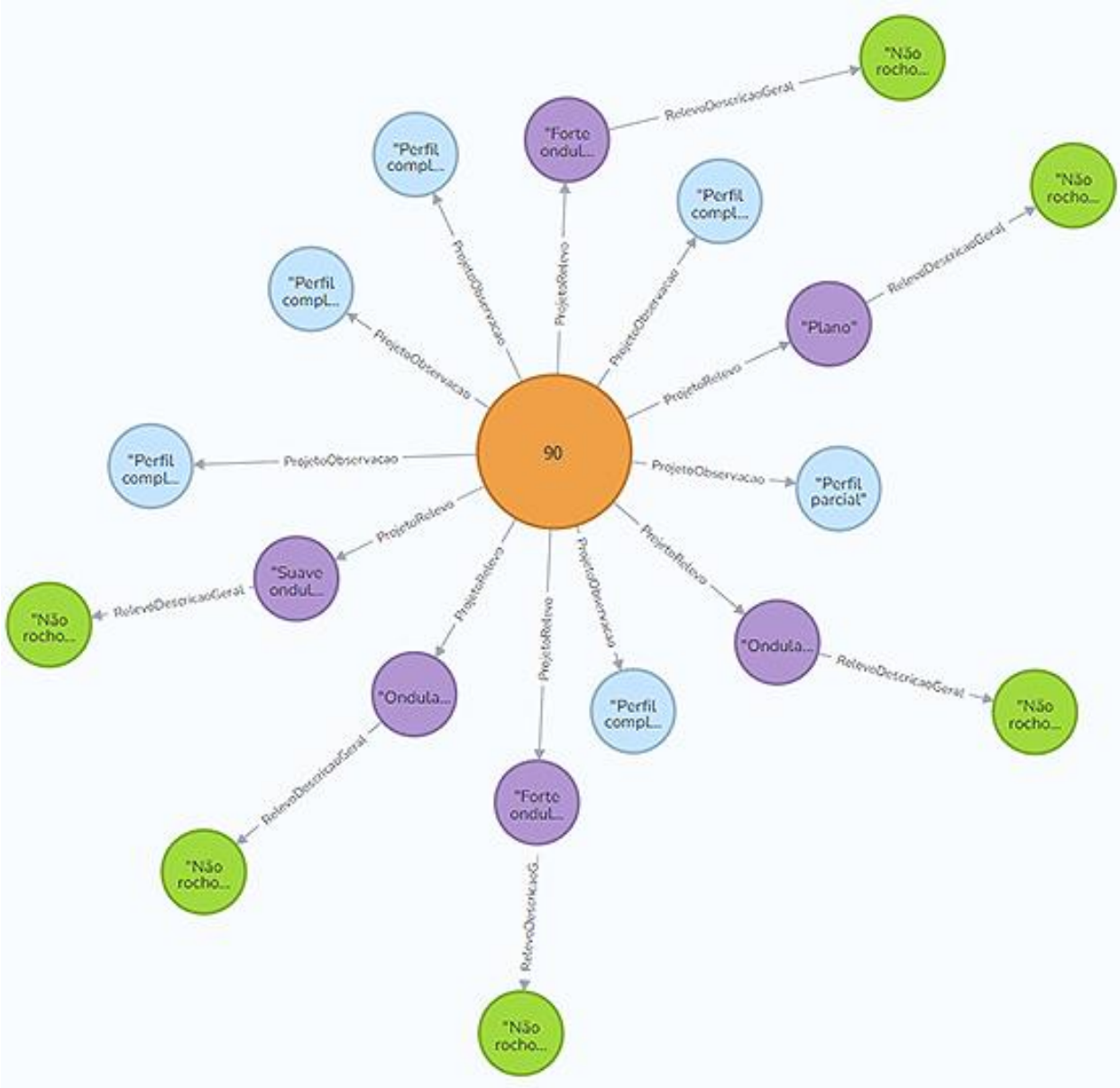


Figura 40 - Relacionamentos entre os nós de projeto, relevo, descrição geral e observação referentes ao projeto 90.

O estabelecimento dos relacionamentos foi realizado entre todos os 48 nós que compõem o banco de dados. Para visualizarmos todas as informações que compõem o projeto de Levantamento Detalhado de solos da área da Fazendinha, utilizamos um recurso da interface do Neo4j e expandimos todos os nós com seus relacionamentos, como pode ser visualizado na Figura 40.

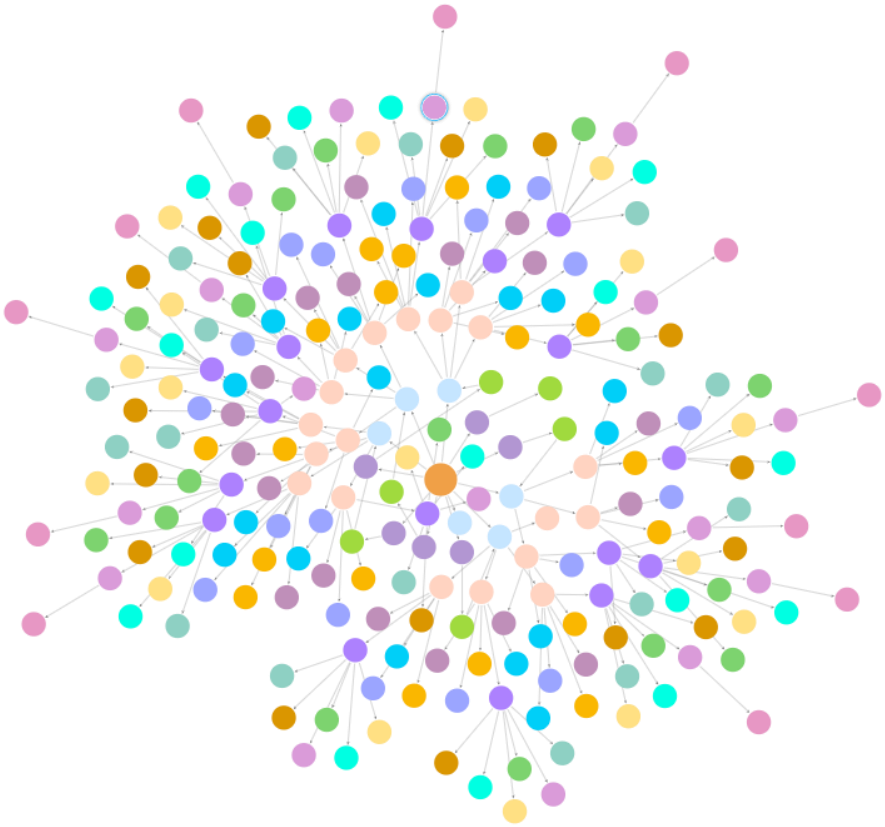


Figura 41 – Visualização de todos os registros do projeto 90 expandidos em um grafo.

6. CONCLUSÃO

É importante destacar que o desenvolvimento das pesquisas em humanidades digitais no Brasil ainda está em fase de expansão, portanto é necessário esclarecer as diferenças entre a pesquisa em HD nacional e as estrangeiras.

Diante dos desafios enfrentados pela HD e a AD buscamos estudar a ampliação da reusabilidade de *datasets* de solos do Brasil. Demonstramos de maneira prática que, a preocupação com a governança, curadoria, qualidade e compartilhamento de dados deve ser considerado como um importante segmento dentro das diversas vertentes das Humanidades Digitais, uma vez que os dados são capazes de, dentre outras coisas, impactar socialmente a partir da disseminação do conhecimento.

As origens desta pesquisa remontam nossa participação no grupo PET-SI da UFRRJ e do grupo de pesquisas do CNPq. No entanto, este trabalho vai muito além dos seus experimentos, traz como contribuição a disponibilização de cerca de 800 mil registros de dados de solos em triplas, no formato RDF/XML, o que torna possível a conexão com a *web* de dados. Além disso, discutimos e apresentamos os aparatos computacionais utilizados na triplificação de dados no contexto da digitalização, bem como as suas origens e vantagens.

Adicionalmente, discutimos a interseção das HD e AD em um cenário de poucas iniciativas e trabalhos nesse sentido; e utilizamos, como caso de estudo, uma ferramenta multicamada capaz de armazenar dados primários e secundários de solos, bem como sua proveniência.

Foram produzidos *workflows* responsáveis por conectar à instância atual do banco de dados relacional do *OpenSoils* à ferramenta de triplificação de dados. Os *workflows* são capazes por realizar a transformação da estrutura relacional atual e gerar dados triplificados. Estes, por sua vez, foram incluídos em um banco de dados de triplas e podem ser acessados a partir do repositório OpensoilsGraph[22] no GitHub.

Além disso, os dados triplificados pelos *workflows* também deram origem à um banco de dados de grafos disponível para acesso a partir do URI e das credenciais de login e senha. Com os dados carregados e disponibilizados nessa plataforma, conseguimos observar seu funcionamento utilizando como exemplo um projeto pedológico realizado na UFRRJ.

[22] https://github.com/OliveiraSabrina/OpensoilsGraph

Com os estudos e análises realizadas durante o desenvolvimento da presente dissertação, foram apresentados e publicados três trabalhos em eventos e congressos, apresentados na seção 6.2 e os produtos gerados, apresentados na seção 6.1, serão registrados no INPI.

6.1. Produtos acadêmicos e de inovação

Além dos *workflows* ETLH construídos (Figura 17), o resultado da execução deles é o principal produto deste trabalho, que consiste em um arquivo com os dados do *OpenSoils* triplificados.

Os arquivos gerados pelos *workflows* ETLH foi incorporado à um novo repositório de dados triplificados que pode ser facilmente integrado com a plataforma *OpenSoils*, gerando um URI que possibilita o seu acesso online, sendo este também um dos produtos oriundos desse trabalho.

Um repositório[23] no GitHub foi gerado contendo os dados triplificados, os *workflows* construídos e utilizados para a triplificação e o repositório de triplas do Neo4j no formato JSON.

Destacamos ainda que, um registro do INPI foi realizado pela Agência de Inovação da UFRRJ para garantir os direitos de propriedade intelectual dos *workflows* ETLH oriundos desta dissertação, estabelecido a partir da petição de número 870230093490 e número de processo 512023003187-9.

6.2. Trabalhos publicados

Durante o período de desenvolvimento deste trabalho, o conhecimento obtido através de leituras e experimentos possibilitou a elaboração, submissão e apresentação de três trabalhos em três congressos, são eles:

> DE OLIVEIRA, Sabrina Santos Cruz et al. Integração de Data Lakes Pedológicos através de Workflows ETLH. In: **Anais da VII Escola Regional de Sistemas de Informação do Rio de Janeiro**. SBC, 2021. p. 48-55.

[23] https://github.com/OliveiraSabrina/OpensoilsGraph

DE OLIVEIRA, Sabrina Santos Cruz et al. Uso de workflows ETLH para integrar datasets pedológicos: estudo para adequação aos princípios FAIR. In: **Anais do XIII Congresso Brasileiro de Agroinformática**. SBC, 2021. p. 348-357.

DE OLIVEIRA, Sabrina Santos Cruz et al. Um estudo integrador dos corpora das Humanidades Digitais com os *datasets* da Agricultura Digital sob os princípios de dados FAIR. In: **HDRio**. 2023.

Além disso, temos um texto em avaliação e que estamos aguardando o resultado, que é o seguinte:

DE OLIVEIRA, Sabrina Santos Cruz et al. Uma investigação integradora dos *datasets* da Agricultura Digital com corpora das Humanidades Digitais sob os princípios de dados FAIR. In: **Revista Brasileira em Humanidades Digitais**. 2023.

6.3. Limitações

Toda pesquisa possui limitações de escopo, a nossa não é exceção. Ela foi executada durante a pandemia de SARS-Cov 2 e sem financiamentos externos. Durante a transformação dos dados para triplas RDF, não foram utilizados vocabulários já existentes, pois não existem vocabulários ou ontologias voltadas para a área de dados pedológicos no Brasil.

A utilização de um vocabulário ou ontologias possibilitaria a padronização dos dados para a integração deles na *web* de dados juntamente com outros dados de solos que porventura já houvessem sido triplificados por outras equipes de pesquisa. Porém, além de não termos encontrado um vocabulário atualizado, não teríamos tempo de avaliá-lo tecnicamente e conhecê-lo em detalhes para a sua correta utilização.

6.4. Trabalhos futuros

A adequação dos dados triplificados a um vocabulário ou alinhamento com ontologias da área de solos que atenda o contexto dos dados de solos do Brasil é, certamente, o próximo passo a ser dado. Uma vez concluída essa etapa, a integração dos dados à *web* de dados também se tornará viável, podendo essa ser também mais uma sugestão de trabalhos futuros.

Além disso, a estruturação de um software com um *front-end* amigável em conjunto com um desenvolvimento *back-end* diretamente ligado aos *workflows*, possibilitaria que os mesmos trabalhassem de uma maneira genérica para todos os tipos de dados, gerando dados triplificados independente do seu segmento.

REFERÊNCIAS

AGRITECH. **Dados 2020/2021**. Disponível em: <https://radaragritech.com.br/dados-2020-2021/>.

ALI, Basharat; DAHLHAUS, Peter. The role of FAIR data towards sustainable agricultural performance: A systematic literature review. **Agriculture**, v. 12, n. 2, p. 309, 2022.

ALMEIDA, Maurício Barcellos. Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares. **Ciência da informação**, v. 31, p. 5-13, 2002.

ALVES, R. C. V. Metadados como elementos do processo de catalogação. 2010. 132 f. **Tese (Doutorado em Ciência da Informação)** – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em: <http://repositorio.unesp.br/handle/11449/103361>. Acesso em: 22 ago. 2017.

ALVES, D. As Humanidades Digitais como uma comunidade de práticas dentro do formalismo académico: dos exemplos internacionais ao caso português, **Ler História**, 69 | 2016, 91-103.

ARAKAKI, Ana Carolina Simionato; ARAKAKI, Felipe Augusto. Dados e metadados: conceitos e relações: concepts and relationships. **Ciência da Informação**, v. 49, n. 3, 2020.

ARAKAKI, F. A. Metadados administrativos e a proveniência dos dados: modelo baseado na família PROV. 2019. 139 f. **Tese (Doutorado)** - Doutorado em Ciência da Informação, Universidade Estadual 2019. Disponível em: https://repositorio.unesp.br/handle/11449/180490. Acesso em: 8 dez. 2021.

ATLAN. What Is Data Curation? How Does It Intersect with Governance?. Disponível em: https://atlan.com/what-is-data-curation/. Acesso em: 25 mai. 2022.

AUER, Sören et al. Dbpedia: A nucleus for a web of open data. In: **The semantic web**. Springer, Berlin, Heidelberg, 2007. p. 722-735.

AUER, Sören et al. Triplify: light-weight linked data publication from relational databases. In: **Proceedings of the 18th international conference on World wide web**. 2009. p. 621-630.

BARBEDO, J. G. A.; KOENIGKAN, L. V. Perspectives on the use of unmanned aerial systems to monitor cattle. **Outlook on Agriculture**, v. 47, n. 3, p. 214-222, June 2018. DOI: 10.1177/0030727018781876.

BARBEDO, J. G. A. Detection of nutrition deficiencies in plants using proximal images and machine learning: A review. **Computers and Electronics in Agriculture**, v. 162, p. 482-492, July 2019b. DOI: 10.1016/j.compag.2019.04.035.

BAUER, Florian; KALTENBÖCK, Martin. Linked open data: The essentials. **Edition mono/monochrom**, Vienna, v. 710, 2011.

BAKER, Thomas. A grammar of Dublin Core. **D-lib magazine**, v. 6, n. 10, p. 3, 2000.

BERARDI, Rita Cristina Galarraga. Design Rationale in the Triplification of Relational Databses. 2015. **Tese de Doutorado**. Programa de Pós-Graduação em Informática of the Departamento de Informática, PUC-Rio.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific american**, v. 284, n. 5, p. 34-43, 2001.

BERNERS-LEE, T. Linked Data: design Issues. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 4 dez. 2021.

BERNERS-LEE, T. Semantic Web road map. 1998. Disponível em: <http://www.w3.org/DesignIssues/Semantic.html>. Acesso em: 17 jan. 2022.

BERRY, David M. Introduction: Understanding the digital humanities. In: **Understanding digital humanities**. Palgrave Macmillan, London, 2012. p. 1-20.

BESSA, Alessandra. Open data tripping in a soil data repository: a case study in the context of Digital Humanities. 2021. 67 p. Dissertation (Master Science in Digital Humanities). **Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro**, Nova Iguaçu, RJ. 2021.

BIOLCHINI, J. et al. Systematic Review in Software Engineering. Technical Report ES, **COPPE / UFRJ**, mai. 2005. Disponível em: <ftp://161.24.19.221/ele/ivo/Leitura/biolchini_2005.pdf>. Acesso em: 7 mar. 2022.

BIRNER, Regina; DAUM, Thomas; PRAY, Carl. Who drives the digital revolution in agriculture? A review of supply-side trends, players and challenges. **Applied Economic Perspectives and Policy**, v. 43, n. 4, p. 1260-1285, 2021.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: the story so far. **International Journal on Semantic Web and Information Systems**, EUA, v. 5, n. 3, p. 1-22, 2009.

BIZER, Christian; VIDAL, Maria-Esther; SKAF-MOLLI, Hala. Linked open data. 2017. Disponível em: <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>. Acesso em: 13 jan. 2022.

BLANKE, T.; PRESCOTT, A. Lidando com Big Data. In: **Griffin G, Hayler M (eds) Métodos de pesquisa para leitura de dados digitais nas humanidades digitais**. Edinburgh University Press, Edinburgo, pp 184–205, 2016.

BOLFE, E. L. et al. Desafios, tendências e oportunidades em agricultura digital no Brasil. **Embrapa Agricultura Digital-Capítulo em livro científico (ALICE)**, 2020.

BOREK, Luise et al. TaDiRAH: a case study in pragmatic classification. 2017.

BRAUNSCHWEIG, Katrin et al. The state of open data. **Limits of current open data platforms**, 2012.

BRAY, Tim et al. Extensible markup language (XML). **World Wide Web Journal**, v. 2, n. 4, p. 27-66, 1997.

BRESLIN, John G.; PASSANT, Alexandre; DECKER, Stefan. The social semantic Web. **New York: Springer**, 2009.

BREVIK, Eric C.; WEINDORF, David C.; STILES, Cynthia. Pedology. **Researchgate**, [S.L], jan. 2015. Disponível em: <http://www.oxfordbibliographies.com/view/document/obo-9780199363445/obo9780199363445-0017.xml.>. Acesso em: 25 abr. 2022.

BRICKLEY, D.; MILLER, L. FOAF: Friend-of-a-Friend. **Retrieved July**, v. 24, p. 2009, 2007.

BRICKLEY, Dan. Resource Description Framework (RDF) Schema Specification. 1998.

BRONSON, Kelly. The immaculate conception of data: agribusiness, activists, and their Shared Politics of the future. **McGill-Queen's Press-MQUP**, 2022.

BRONSON, Kelly; SENGERS, Phoebe. Big tech meets big ag: Diversifying epistemologies of data and power. **Science as Culture**, v. 31, n. 1, p. 15-28, 2022.

BROWELL, Geoff. From linked open data to linked open knowledge. **Digital Information Strategies: From Applications and Content to Libraries and People**, p. 87, 2015.

BUENO-SOLER, Juliana; CARNIELLI, Walter. e-Reasoning: Between Digital Humanities and e-science. In: **2014 IEEE 10th International Conference on e-Science**. IEEE, 2014. p. 33-35.

BUNEMAN, Peter; KHANNA, Sanjeev; WANG-CHIEW, Tan. Why and where: A characterization of data provenance. In: **International conference on database theory**. Springer, Berlin, Heidelberg, 2001. p. 316-330.

BURDICK, Anne et al. Um breve guia para as Humanidades Digitais. **TECCOGS Revista Digital de Tecnologias Cognitivas**, 2020.

BUSA, Roberto. The annals of humanities computing: The index thomisticus. **Computers and the Humanities**, p. 83-90, 1980.

CAPMOURTERES, Virginia et al. Precision conservation meets precision agriculture: A case study from southern Ontario. **Agricultural systems**, v. 167, p. 176-185, 2018.

CATARINO, Maria Elisabete; SOUZA, Terezinha Batista de. A representação descritiva no contexto da web semântica. **Transinformação**, v. 24, p. 77-90, 2012.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. Europeana no Linked Open Data: conceitos de Web Semântica na dimensão aplicada das Humanidades Digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 22, n. 48, p. 88-99, jan./abr. 2017. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2017v22n48p88>. Acesso em: 08 out. 2021.

CEDDIA, M.B.; CRUZ, S. M. S. ; MIRANDA, R. ; CRUZ, P. V. ; CRUZ, S. S. O. ; RIZZO, G. S. . OpenSoils DB. 2018.

CRUZ, Sergio Manuel Serra da; DUARTE, A. ; KLINGER, F. ; Pedro Vieira Cruz ; MARINHO, E. C. ; MENDES, J. ; SCHMITZ:, E. A. . OpenSoils: Uma Plataforma de Apoio à Agricultura Digital. In: Congresso Brasileiro de Agroinformática - SBIAgro 2019, 2019, Indaiatuba/SP. **Anais da XII Congresso Brasileiro de Agroinformática - SBIAgro** 2019. Brasilia: Embrapa, 2019.

CRUZ, S. M. S.; NASCIMENTO, J. A. P. . Towards integration of data-driven agronomic experiments with data provenance. **COMPUTERS AND ELECTRONICS IN AGRICULTURE**, v. 161, p. 14-28, 2019.

CUARTAS-RESTREPO, Juan Manuel. Humanidades digitais, deixe-as ser. **Revista Colombiana de Educación**, n. 72, p. 65-78, 2017.

DEBRUYNE, Christophe; O'SULLIVAN, Declan. R2RML-F: towards sharing and executing domain logic in R2RML mappings. In: **LDOW@ WWW**. 2016.

DA CRUZ, Sérgio Manuel Serra et al. Towards an e-infrastructure for Open Science in Soils Security. In: **Anais do XII Brazilian e-Science Workshop**. SBC, 2018.

DA CRUZ, Sérgio Manuel Serra; KLINGER, F. ; Pedro Vieira Cruz ; MARINHO, E. C. ; SCHMITZ:, E. A. . Desenvolvendo Sistemas Agrícolas Próxima Geração: Um Estudo em Ciências de Solos. In: X Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais, 2019, Belém - PA. **Anais do XXXIC Congresso da Sociedade Brasileira de Computação**. Porto Alegre: Sociedade Brasielira de Computação, 2019.

DE OLIVEIRA, Sabrina Santos Cruz et al. Integração de Data Lakes Pedológicos através de Workflows ETLH. In: **Anais da VII Escola Regional de Sistemas de Informação do Rio de Janeiro**. SBC, 2021a. p. 48-55.

DE OLIVEIRA, Sabrina Santos Cruz et al. Uso de workflows ETLH para integrar datasets pedológicos: estudo para adequação aos princípios FAIR. In: **Anais do XIII Congresso Brasileiro de Agroinformática**. SBC, 2021b. p. 348-357.

DEB, Chandan Kumar; MARWAHA, Sudeep; PANDEY, R. N. Ontology Learning Algorithm for Development of Ontologies from Taxonomic Text and USDA Soil Taxonomy Ontology. **Journal of the Indian Society of Agricultural Statistics**, v. 1, p. 74, 2020.

DOKUCHAEV, V.V. 1883. Russian Chernozem. In: **V.V. Dokuchaev**. Selected Papers, 1: 14-419. (Translated into English by N. Kander – Jerusalem: Israel Program for Scientific Translations).

DRUCKER, Debora P. et al. Implantação da Rede Temática GO-FAIR Agro Brasil: Primeiros Passos. In: **Anais do XIII Congresso Brasileiro de Agroinformática**. SBC, 2021. p. 164-171.

DUVAL, Erik et al. Metadata principles and practicalities. **D-lib Magazine**, v. 8, n. 4, p. 1-10, 2002.

ESPERIDIÃO, T. L.; SANTOS, T. C.; AMARANTE, M. S. Agricultura 4.0: Software de Gerenciamento de Produção. **Mogi das Cruzes: Pesquisa e Ação V5 N4**, 2019.

ESPÍNDOLA, P. L.; SALM JUNIOR, J. F.; ROSA, F.; JULIANI, J. P. Governança de dados aplicada à ciência da informação: análise de um sistema de dados científicos para a área da saúde. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, SP, v. 16, n. 3, p. 274–298, 2018. DOI: 10.20396/rdbci.v16i3.8651080. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8651080. Acesso em: 25 maio. 2022.

EUROPEAN, C. Eurostat. Your key to European statistics, 2013. Disponível em: http://epp.eurostat.ec. europa.eu/portal/page/portal/eurostat/home.

FECHER, Benedikt; FRIESIKE, Sascha. Open science: one term, five schools of thought. In: **Opening science**. Springer, Cham, 2014. p. 17-47.

FERREIRA, João et al. O processo etl em sistemas data warehouse. In: **INForum**. 2010. p. 757-765.

FERREIRA, Jaider Andrade et al. O modelo de dados Resource Description Framework (RDF) e o seu papel na descrição de recursos. **Informação & Sociedade: Estudos**, v. 23, n. 2, 2013.

FERREIRA, Jaider Andrade. Wikis semânticos: da Web para a Web Semântica. 2014. 131 f. Dissertação (Mestrado em Ciência da Informação). **Faculdade de Filosofia e Ciências – Universidade Estadual Paulista**, Marília, 2014. Disponível em: <http://hdl.handle.net/11449/108380>.

FLANDERS, Julia; JANNIDIS, Fotis (Ed.). **The shape of data in digital humanities: modeling texts and text-based resources.** Routledge, 2018.

FRANÇA, Renata et al. Transformação digital na agricultura moderna: pilares e proposta de modelo para o futuro da inovação agrícola. In: **Anais do Congresso Internacional de Conhecimento e Inovação–ciki**. 2019.

FURNER J. "Data": The data. In: KELLY, M.; BIELBY J. (Eds.) **Information Cultures in the Digital Age**. Wiesbaden: Springer VS, 2016.

GERHARDT, T. E.; SILVEIRA, D. T. (Orgs). Métodos de pesquisa. Porto Alegre: **Editora da UFRGS**, 2009. (Série Educação à Distância).

IBICT/MCTI. Instituto Brasileiro de Informação em Ciência e Tecnologia. Fonte: IBICT: http://sitehistorico.ibict.br/informacao-para-ciencia-tecnologia-e-inovacao%20/repositorios-digitais/sobre-repositorios-digitais. 2012.

GILLILAND, A. J. Setting the Stage. In: **BACA, M.** (Ed.). Introduction to metadata. 3. ed. Los Angeles: Getty Research Institute, 2016.

GUTIERREZ, Claudio; HURTADO, Carlos A.; VAISMAN, Alejandro. Introducing time into RDF. **IEEE Transactions on Knowledge and Data Engineering**, v. 19, n. 2, p. 207-218, 2006.

HEATH, Tom; BIZER, Christian. Linked data: Evolving the web into a global data space. **Synthesis lectures on the semantic web: theory and technology**, v. 1, n. 1, p. 1-136, 2011.

HAWKINS, Ashleigh. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. **Archival Science**, p. 1-26, 2021.

HELFER, Gilson Augusto et al. Tellus-Onto: uma ontologia para classificação e inferência de solos na agricultura de precisão: Tellus-Onto: an ontology for soil classification and inference in precision agriculture. In: **XVII Brazilian Symposium on Information Systems**. 2021. p. 1-7.

HELMI, Laila CA. DIGITAL HUMANITIES: A PARADIGM FOR THE 21ST CENTURY. **BAU Journal-Society**, Culture and Human Behavior, v. 2, n. 2, p. 5, 2021.

HENNING, Patrícia Corrêa et al. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389-412, 2019.

HITZLER, P.; KRÖTZSCH, M.; RUDOLPH, S. Foundations of Semantic Web technologies. **Boca Raton: CRC Press**, 2010

HOCKEY, S. The history of humanities computing. In: (Hockey, 2007) Susan Schreibman, Ray Siemens e John Unsworth, eds., **Companion to Digital Humanities**. Oxford, Blackwell, 2004, pp. 15-16.

HOEKSTRA, Rinke et al. The dataLegend ecosystem for historical statistics. **Journal of Web Semantics**, v. 50, p. 49-61, 2018.

HYVÖNEN, Eero. Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. **Semantic Web**, v. 11, n. 1, p. 187-193, 2020.

IRWIN, A. Citizen science: a study of people, expertise and sustainable development. **London: Routledge**, 2002. 216 p.

JAYARAMAN, Prem Prakash et al. Addressing information processing needs of digital agriculture with OpenIoT platform. In: **Interoperability and Open-Source Solutions for the Internet of Things**. Springer, Cham, 2015. p. 137-152.

KITAMURA, Yoshinobu et al. Deployment of an ontological framework of functional design knowledge. **Advanced Engineering Informatics**, v. 18, n. 2, p. 115-127, 2004.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, **Departament of Computer Science Keele University**, Keele, 2007.

KOH, A. Niceness, Building, and Opening the Genealogy of the Digital Humanities: Beyond the Social Contract of Humanities Computing. **Differences**, 25(1), 93–106 (2014). doi:10.1215/10407391-2420015.

KOIVUNEN, Marja-Riitta; MILLER, Eric. W3C Semantic Web Activity. **Cambridge: W3C**, 2001. Disponível em: <http://www.w3.org/2001/12/semweb-fin/w3csw>. Acesso em: 21 fev. 2022

LAMPRECHT, Anna-Lena et al. Towards FAIR principles for research software. **Data Science**, v. 3, n. 1, p. 37-59, 2020.

MANUEL SERRA DA CRUZ, S.; DEOLIVEIRA, A.; FIRMINODE FARIA, F. Evolutionary Scientific Workflows. In: **2018 IEEE Congress on Evolutionary Computation (CEC)**, 2018, Rio de Janeiro. 2018 IEEE Congress on Evolutionary Computation (CEC), 2018. v. 1. p. 1.

MARINHO, Élton Carneiro; SCHMITZ, Eber Assis; DA CRUZ, Sérgio Manuel Serra. Provenance, Blockchain, and Smart Contracts as a Traceable Soil Mapping Solution. **In: Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados**. SBC, 2022. p. 89-97.

MARINHO, Élton Carneiro et al. OpenSoils: Uma Plataforma de Apoio à Agricultura Digital Brasileira. **In: Anais Estendidos do XIX Simpósio Brasileiro de Sistemas de Informação**. SBC, 2023. p. 90-92.

MARTINS DA SILVA, D.; QUARESMA MARTINS, R.; PINTO LIMA, H. Estudo de caracterização físico-quimico das terras pretas arqueológicas da região de Caxiuanã no âmbito do projeto: ocupação humana no Delta Amazônico.

MASSRUHÁ, S. M. F. S.; LEITE, M. A. de A.; OLIVEIRA, S. R. de M. *et al.* (Ed. Téc.). Agricultura Digital: pesquisa, desenvolvimento e inovação nas cadeias produtivas. Brasília, DF: **EMBRAPA**, 2020. Disponível em:< https://bit.ly/3uYefIP>.

MATOS, J. C. M.; JACINTHO, E. M. D. S. B.; ALVAREZ, E. B. Humanidades digitais e a simbiose entre humano e máquina: algumas reflexões comparativas entre a interpretação e a mineração de textos. **Logeion: filosofia da informação**, v. 6, n. 1, p. 126-145, 2019. DOI: 10.21728/logeion.2019v6n1.p126-145 Acesso em: 19 ago. 2023.

MCGUINNESS, Deborah L. et al. OWL web ontology language overview. **W3C recommendation**, v. 10, n. 10, p. 2004, 2004.

MEJIAS, Ulises A.; COULDRY, Nick. Datafication. **Internet Policy Review**, v. 8, n. 4, 2019.

MEROÑO-PEÑUELA, Albert et al. Digital humanities on the Semantic Web: accessing historical and musical linked data. **Journal of Catalan Intellectual History**, v. 1, n. 11, p. 144-149, 2017.

MILLER, Eric. An introduction to the resource description framework. **D-lib Magazine**, 1998.

MONS, Barend. **Data stewardship for open science: Implementing FAIR principles**. CRC Press, 2018.

MOREIRA, Fábio Mosso et al. Tecnologias da Web Semântica para a recuperação de dados agrícolas: um estudo sobre o International Information System of the Agricultural Science and Technology (AGRIS). **Em Questão**, v. 21, n. 1, p. 1-20, 2015.

MUÑOZ, Andrés; SORIANO-DISLA, José Martín; JANIK, Leslie J. An Ontology-Based Approach for an Efficient Selection and Classification of Soils. In: **Intelligent Environments (Workshops)**. 2017. p. 69-78.

MURRAY-RUST, Peter. Open data in science. **Nature Precedings**, p. 1-1, 2008.

NEUROHACKWEEK; REPRONIM. Introduction to the Web of Data. 2016. Disponível em: <http://www.repronim.org/module-FAIR-data/01-Web-of-Data/>. Acesso em 09 abr. 2023.

NGO, Quoc Hung; KECHADI, Tahar; LE-KHAC, Nhien-An. OAK: Ontology-Based Knowledge Map Model for Digital Agriculture. In: **International Conference on Future Data and Security Engineering**. Springer, Cham, 2020. p. 245-259.

NININ, Débora Marroco. Linked Open Data em coleções de patrimônio cultural: aspectos da representação da informação para Humanidades Digitais. 2018.

PRIYATNA, Freddy; CORCHO, Oscar; SEQUEDA, Juan. Formalisation and experiences of R2RML-based SPARQL to SQL query translation using morph. In: **Proceedings of the 23rd international conference on World Wide Web**. 2014. p. 479-490.

RAMALHO, Rogério Aparecido Sá. Web Semântica: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação. 2006. 120 f. **Dissertação (Mestrado em Ciência da Informação)**. Faculdade de Filosofia e Ciências – Universidade Estadual Paulista, Marília, 2006. Disponível em: <http://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/ramalho_ras_me_mar.pdf>. Acesso em: 10 abr. 2022.

RAMSAY, Stephen. Who's in and who's out. In: **Defining digital humanities**. Routledge, 2016. p. 255-258.

RAY, Erik. Aprendendo XML. Rio de Janeiro: Campus; O'Reilly, 2001.

REICHE, R. ; FERREIRA, J. ; Kelli de Faria Cordeiro ; CRUZ, Sergio Manuel Serra da ; Maria Claudia Cavalcanti ; CAMPOS, Maria Luiza Machado . LOP - Capturing and Linking Open Provenance on LOD Cycle. **In: 5th International Workshop on Semantic Web Information Management - SWIM**, 2013, New York , USA. 2013 ACM International Conference on Management of Data (SIGMOD 2013), 2013.

RILEY, J. Understanding Metadata: what is metadata, and what is it for? **National Information Standards Organization (NISO)**, 2017. Disponível em:<http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf>. Acesso em: 01 ago. 2022.

ROBINSON, Nathan J. et al. Testing the public–private soil data and information sharing model for sustainable soil management outcomes. **Soil use and management**, v. 35, n. 1, p. 94-104, 2019.

RODRÍGUEZ EUGENIO, Natalia. The Global Soil Partnership: Tackling Global Soil Threats Through Collective Action. **International Yearbook of Soil Law and Policy 2019**, p. 197-221, 2021.

SANTOS, Cristina P.; DA SILVA, Denílson Rodrigues; CARDOSO, Gleidson Antônio. ONIAQUIS–Uma Ontologia para a Interpretação de Análise Química do Solo. **VI Simpósio de Informática da Região Centro do Rio Grande do Sul. Unifra**, 2007.

SANTOS, P. L. V. A. da C.; SANTANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. Ciência da Informação, v. 42, n. 2, jan. 2013. Disponível em: http://revista.ibict.br/index.php/ciinf/article/view/228. Acesso em: 8 dez. 2021.

SCIENTIFIC ELECTRONIC LIBRARY ONLINE. Princípios reitores FAIR publicados em periódico do Nature Publishing Group. **SciELO em Perspectiva**, [S.l], 2016.

SCHNAPP, Jeffrey et al. Digital humanities manifesto 2.0. Retrieved September, v. 23, p. 2012, 2009.

SCRIMGEOUR, Charlie. Handbook of Soil Analysis. Mineralogical, Organic and Inorganic Methods. By M. Pansu and J. Gautheyrou. Berlin, Heidelberg, New York: Springer (2006), pp. 993,£ 191.50. ISBN 978-3-540-31210-9. **Experimental Agriculture**, v. 43, n. 3, p. 401-401, 2007.

SIBCS, 2019. Disponível em: https://www.embrapa.br/solos/sibcs.

SEGATA, J.; RIFIOTIS, T. Digitalização e dataficação da vida. **Civitas - Revista de Ciências Sociais**, v. 21, n. 2, p. 186-192, 24 ago. 2021.

SHEPHERD, Mark et al. Priorities for science to overcome hurdles thwarting the full promise of the 'digital agriculture'revolution. **Journal of the Science of Food and Agriculture**, v. 100, n. 14, p. 5083-5092, 2020.

SILVA, Luciana Candida; SEGUNDO, JOSÉ EDUARDO SANTAREM; SILVA, MARCEL FERRANTE. PRINCÍPIOS DE FAIR E MELHORES PRÁTICAS DO LINKED DATA NA PUBLICAÇÃO DE DADOS DE PESQUISA. In: **XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XIX ENANCIB)**. 2018.

SILVERTOWN, J. A new dawn for citizen science. **Trends in Ecology & Evolution**. v. 24, n. 9, p. 467-471, 2009. DOI: 10.1016/j.tree.2009.03.017.

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

SORBARA, Agostino. Digital Humanities and Semantic Web: The New Frontiers of Transdisciplinary Knowledge. **Journal of Higher Education Theory and Practice**, v. 20, n. 13, p. 206-210, 2020.

SOUTHERTON, C. Datafication. In: Schintler L., McNeely C. (eds) Encyclopedia of Big Data. Springer, Cham, 2020. https://doi.org/10.1007/978-3-319-32001-4_332-1

SPELLMAN, Bobbie; GILBERT, Elizabeth; CORKER, Katherine S. Open science: What, why, and how. 2017.

STEIN, Mari-Klara et al. Datification and the pursuit of meaningfulness in work. **Journal of Management Studies**, v. 56, n. 3, p. 685-717, 2018.

SUBIRATS-COLL, Imma et al. AGROVOC: The linked data concept hub for food and agriculture. **Computers and Electronics in Agriculture**, v. 196, p. 105965, 2022.

SVENSSON, Patrik. Humanities computing as digital humanities. In: **Defining Digital Humanities**. Routledge, 2016. p. 175-202.

SVENSSON, Patrik. Beyond the big tent. Debates in the digital humanities, v. 36, p. 49, 2012.

SVENSSON, Patrik. Humanities computing as digital humanities. **Defining Digital Humanities: A Reader**, v. 159, 2013.

TANG, Shihao et al. A conception of digital agriculture. In: **IEEE international geoscience and remote sensing symposium**. IEEE, 2002. p. 3026-3028.

TARGULIAN, V. O.; KRASILNIKOV, P. V. Soil system and pedogenic processes: Self-organization, time scales, and environmental significance. **Catena**, v. 71, n. 3, p. 373-381, 2007.

TEIXEIRA, Gerson. O Censo Agropecuário 2017. **Revista NECAT-Revista do Núcleo de Estudos de Economia Catarinense**, v. 8, n. 16, p. 8-39, 2019.

TURKI, Houcemeddine et al. Representing COVID-19 information in collaborative knowledge graphs: the case of Wikidata. **Semantic Web**, n. Preprint, p. 1-32, 2021.

VAN ES, Harold. A new definition of soil. **CSA News**, v. 62, n. 10, p. 20-21, 2017.

VELLUCCI, Sherry L. Metadata. **Annual Review of Information Science and Technology (ARIST)**, v. 33, p. 187-222, 1998.

W3C. Resource Description Framework (RDF). Disponível em: https://www.w3.org/RDF/

W3C. R2RML: RDB to RDF Mapping Language. W3C, 2012. Disponível em: <https://www.w3.org/TR/r2rml/>. Acesso em: 26 fev. 2022.

W3C. SPARQL 1.1 Overview. Cambridge: W3C, 2013a. Disponível em: <http://www.w3.org/TR/sparql11-overview>. Acesso em: 28 jul. 2013.

W3C. W3C Semantic Web Activity. Cambrigde: W3C, 2013b. Disponível em: <http://www.w3.org/2001/sw>. Acesso em: 24 fev. 2022.

WILDING, L. P. 1994. Factors of soil formation: Contributions to pedology. In Factors of soil formation: A fiftieth anniversary retrospective: Proceedings of a symposium cosponsored by the **Council on the History of Soil Science (5205.1) and Division S-50 of the Soil Science Society of America**, Held in Denver, CO, 28 October 1991. Edited by R. Amundson, J. Harden, and M. Singer, 15–30. SSSA Special Publication 33. Madison, WI: Soil Science Society of America.

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

WOODS, William. Os solos e as ciências humanas: Interpretação do passado. As Terras Pretas de Índio da Amazônia: sua caracterização e uso deste conhecimento na criação de novas áreas, p. 62-71, 2009.

ZHANG, Xiao; YANG, Deling. Reuse of Public Data and Information Property and Building of Knowledge Society. In: **Proceedings of the 2018 International Conference on Information Science and System**. 2018. p. 266-272.

ZUIDERWIJK, Anneke; JANSSEN, Marijn; DAVIS, Chris. Innovation with open data: Essential elements of open data ecosystems. **Information polity**, v. 19, n. 1-2, p. 17-33, 2014.