UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Anderson Juan Rocha Menezes

Avaliação da Qualidade de Dados Abertos Educacionais em Instituições Públicas de Ensino Superior

Recife

Outubro de 2024

Anderson Juan Rocha Menezes

Avaliação da Qualidade de Dados Abertos Educacionais em Instituições Públicas de Ensino Superior

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientadora: Roberta Macêdo Marques Gouveia
Coorientadora: Maria da Conceição Moraes Batista

Recife
Outubro de 2024

Avaliação da Qualidade de Dados Abertos Educacionais em Instituições Públicas de Ensino Superior

Anderson Juan Rocha Menezes[1], **Roberta Macêdo Marques Gouveia**[1], **Maria da Conceição Moraes Batista**[1], **Gabriel Alves de Albuquerque Júnior**[1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

{anderson.juan,roberta.gouveia,maria.cmbatista,gabriel.alves}@ufrpe.br

Abstract. *Open government data provides important information about the various scenarios in a region and can guide public policy decisions, even between countries. Its publication has an impact on transparency and economic potential, especially when discussed in the educational context. Although current frameworks are aimed at facilitating the availability of open data on the Web, it is observed that there is still significant room for improvement in terms of ensuring the quality of published data. This aspect reveals the importance of a more rigorous and systematic assessment of the quality of these data. In this context, there is a need to assess the quality of open educational data after its availability, where scores can be established for each criterion, for example, credibility. The definition of the dimensions (criteria) with the formulas created in this work could be seen as determinants for the variation of the same for each institution evaluated, therefore it is important to understand the bases for its use, in addition to the quality of the metadata generated having a significant weight in the result.*

Resumo. *Os dados abertos governamentais trazem informações importantes sobre os diversos cenários de uma região e podem nortear politicas públicas decisórias até mesmo entre países. Sua publicação traz impactos em transparência e potencial econômico, especialmente quando se é falado no contexto educacional. Embora os frameworks atuais sejam voltados para facilitar a disponibilização de dados abertos na Web, observa-se que ainda há um espaço significativo para melhorias no que tange à garantia da qualidade dos dados publicados. Esse aspecto revela a importância de uma avaliação mais rigorosa e sistemática da qualidade desses dados. Nesse contexto, surge a necessidade de avaliação da qualidade dos dados abertos educacionais depois de sua disponibilização onde se possa estabelecer pontuações para cada critério, por exemplo, credibilidade. A definição das dimensões (critérios) com as fórmulas criadas neste trabalho puderam ser vistas como determinantes para a variação da mesma por cada instituição avaliada, portanto sendo importante o entendimento sobre as bases para sua utilização, além da qualidade dos metadados gerados terem peso significativo no resultado.*

1. Introdução

Nos últimos anos, a disponibilização de dados abertos tem desempenhado um papel fundamental no contexto da transparência pública e no fortalecimento das decisões baseadas

em dados, fornecendo à sociedade um meio de monitorar a ação governamental e, consequentemente, aprimorar a qualidade das políticas públicas. Na área da educação, os Portais de Dados Abertos se destacam como canais de comunicação entre as instituições de ensino e a sociedade, reunindo informações relevantes para diversos *stakeholders*. A acessibilidade dos dados abertos fornecidos pelas universidades federais, necessários para o cálculo de variados indicadores, por exemplo a Taxa de Permanência (TAP), entre outros, foram contemplados no trabalho de Lima (2022) onde é criado uma Classificação das Instituições de Ensino Superior (IES) públicas segundo a Taxa de Acesso aos Dados Abertos.

No entanto, a qualidade dos dados abertos ainda é uma questão premente a ser abordada, como destacado por Zuiderwijk et al. (2012) em seu estudo sobre os desafios dos dados governamentais abertos. Eles afirmam que, apesar dos avanços significativos na disponibilização de dados, a qualidade muitas vezes não atende às necessidades dos usuários, impactando negativamente a utilidade desses dados para decisões estratégicas e operacionais. Um estudo complementar, como o que se propõe neste trabalho, torna-se necessário para avaliar a qualidade dos dados disponibilizados e estabelecer critérios sólidos de avaliação. Pode-se observar também que a questão da qualidade dos dados não está somente ligada aos dados em si mas também ao mau gerenciamento de metadados pois afetam a qualidade dos dados como demonstrado em Ryu et al. (2006), no qual se norteou uma nova perspectiva no estado da arte cujo teve a preocupação da qualidade de trabalhos anteriores como de Kahn et al. (2002), apud Ryu et al. (2006) se preocupavam no conteúdo do dado e pouco na estrutura que se descreve sobre esses dados (metadados).

Atualmente o problema de avaliação de qualidade dos dados já se encontrou importância em diversas áreas como no agronegócio de onde surge o estudo de Junior and Dorneles (2021) que levou em consideração metadados já gerados pelas bases que o trabalho avaliou, na situação atual deste trabalho fez-se necessário gerar essa estrutura de metadados para fazer a avaliação dos critérios posteriormente, um outro bom exemplo do estado da arte é a inconsistências de qualidade de dados em dados governamentais brasileiros apontado nos resultados de Oliveira et al. (2023), entretanto se utiliza a abordagem de comparação ferramental onde se opta pelo GE (Great Expectations) para gerar as métricas e por tanto chegar a conclusão, segundo o próprio Great Expectation (2024), "o GE ajuda você a obter insights sobre seus dados com mais rapidez, colaborar de forma mais eficaz e capacitar equipes técnicas e não técnicas para trabalhar em seus estilos preferidos e, ao mesmo tempo, atingir seus objetivos comuns. As expectativas criam os pontos de referência centrais que são fundamentais para a qualidade dos dados".

Nesse sentido, vale destacar que a avaliação da qualidade dos dados abertos é um desafio importante. O estudo de Naumann and Rolker (2005) apresenta valiosas contribuições para a área de qualidade dos dados em si. Nele, são identificadas três classes de critérios de qualidade de informação, cada uma com diferentes possibilidades de avaliação, detalhadas de forma minuciosa. Além disso, o trabalho também aborda medidas de confiança para os métodos de avaliação, aspecto fundamental para garantir a precisão e credibilidade dos resultados. Munido do atual estado da arte e de conhecimento prévio dos trabalhos até aqui citados, o presente trabalho visa avaliar a qualidade dos dados para isso foi necessário gerar uma estrutura dos metadados e ainda sim criar métricas que tornem possíveis a mensuração do que representa essa melhoria na qualidade, como

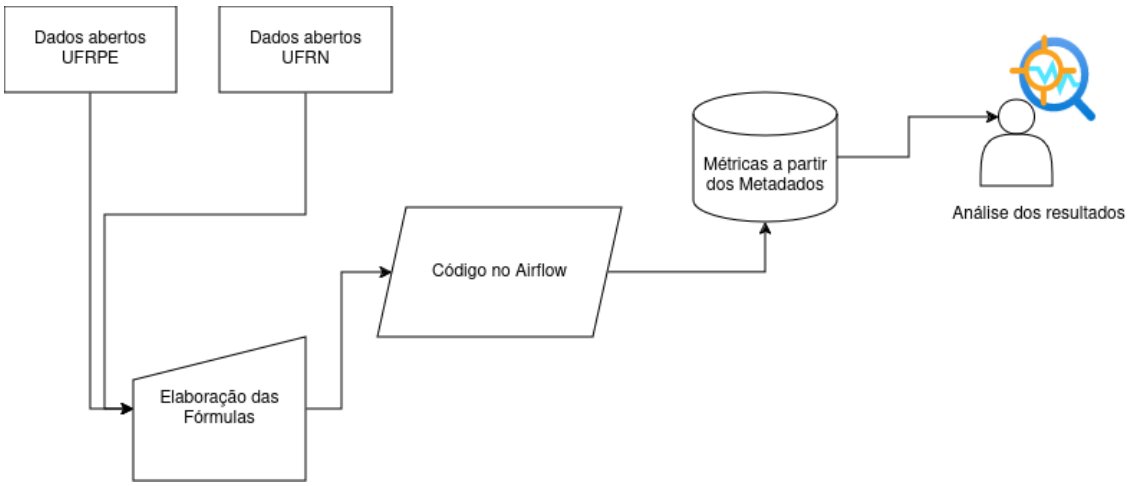
trabalhos anteriores se utilizam de ferramentas como o GE para já gerar essas métricas, a abordagem do presente trabalho é ligeiramente diferente pelo fato de propor a tornar metrificável por meio da avaliação dos critérios (dimensões) o desafio de interoperabilidade e alinhamento semântico citadas no estudo de Penteado et al. (2021).

Sendo assim, diante da crescente disponibilidade de dados abertos na educação e da necessidade de garantir a qualidade desses dados, surgem duas perguntas de pesquisa, são elas: (i) O gerenciamento de metadados atendendo aos critérios de estado da arte atual implica em uma maior qualidade dos dados no contexto de DAE (Dados Abertos Educacionais)? (ii) Os critérios (dimensões) de qualidade escolhidos acabam sendo quanto percentualmente melhor quando altera-se o tipo de tabela analisada?

A primeira pergunta envolve os problemas de gerenciamento de metadados e qualidade dos dados, e sua resposta dá um salto de paradigma em possível correlação existente entre eles. Já a segunda pergunta envolve os problemas de quais dimensões afetam mais e quais podem afetar menos a base de dados baseada em sua natureza. Este estudo lida com bases de dados educacionais, e a resposta a esta pergunta é importante para compreender melhor tais bases.

Dito isto, o presente trabalho tem como propósito avaliar com critérios (dimensões) de qualidade estabelecidos pelos autores, tornando metrificável por meio de equações, a qualidade dos dados abertos educacionais de bases disponíveis por meio do portal de dados abertos de IES. Com três objetivos específicos sendo o **(i) de gerar metadados que cumpram os requisitos do experimento** e contemplem requisitos mínimos da geração dos metadados sendo eles: número de colunas, número de linhas, tipagem das colunas, valores nulos por colunas, valores distintos por coluna. Para isto foi utilizado uma automação para extração desses metadados com um método de validação a partir de comparação com padrões estabelecidos para metadados educacionais ou similares. **(ii) Comparar os critérios de qualidade**, para isto foram selecionados os seguintes critérios (dimensões) baseado nos requisitos mínimos dos metadados: Precisão, Completude, Credibilidade, Consistência. O método comparativo para avaliação utilizado foi o de cálculo percentual para cada critério em relação ao total de conjuntos de dados avaliados além de uma análise estatística para identificar diferenças significativas entre as instituições. **(iii) Criar fórmulas matemáticas capazes de descrever a realidade das dimensões escolhidas (ou contexto de caso especifico)**, no qual culminou no desenvolvimento de fórmulas para incorporação de variáveis ponderadas com base na importância de cada critério (dimensão) para a qualidade dos dados. A forma de validação foram testes das fórmulas em conjuntos de dados não utilizados na fase de desenvolvimento e que já tenham passado por estudos relacionados, além de comparação dos resultados das fórmulas com avaliações subjetivas de especialistas na área educacional, este processo pode ser visualizado na Figura 1.

Figura 1. Fluxograma do processo citado como objetivos específicos



Fonte: autor (2024)

2. Trabalhos Relacionados

Dado a contextualização é necessário abordar o que de fato se sabe sobre o problema levantado, partindo de uma tese de pesquisa de Alves et al. (2010) na qual se propõe a integração estratégica entre as tecnologias de informática e os métodos de Tratamento Descritivo da Informação (TDI) na Ciência da Informação como meio de consolidar a construção padronizada e consistente de metadados. A hipótese sugere que teorias, princípios, fundamentos e métodos de catalogação, em meio a atualizações frente às mudanças tecnológicas, orientam essa construção padronizada em padrões de metadados. Partindo dessa premissa o trabalho se relaciona com o atual a partir do momento em que se fez necessário a geração e padronização dos metadados justamente com finalidade de garantia da unicidade e formas de recuperação desta informação.

Quando foi realizado o levantamento da existência de inconsistências na qualidade de dado abertos governamentais se depara com o estudo de Oliveira et al. (2023) onde se faz um estudo de caso nos dados abertos governamentais no estado de Minas Gerais, entretanto diferente do que se foi explicitado até agora não é utilizada a lógica dos metadados como chave dessa medição mas sim já usar ferramentas da indústria próprias para isso, com isto se compara 8 destas e optam pelo GE para analisar dados de licitações públicas e despesas públicas. A análise trás que o GE identifica efetivamente problemas na qualidade dos dados que podem impactar a construção de aplicações finais que utilizam esses dados o que também é um fim para o nosso objetivo, visto que dados com maior qualidade garante melhoria e aproximação da realidade para métricas importantes para o pais como abordado por Lima (2022).

Como esta abordagem de fins comparativos entre ferramentas e avaliação da qualidade dos dados é muito bem sanada no artigo a ideia do presente trabalho é então não repetir as comparações ferramentais com um trabalho técnico mas sim tentar expandir com um trabalho científico semelhante ao que Ryu et al. (2006) fez no passado onde trás a luz as possibilidade de avaliação com metadados e suas amplas riquezas de possibilidades. Em resumo, o artigo enfatiza a importância da qualidade de dados em dados

governamentais, introduz o GE como uma ferramenta adequada para análise e propõe uma nova métrica de qualidade para facilitar a inspeção manual prioritária de tabelas com problemas identificados.

No trabalho de Naumann and Rolker (2005) é definido os critérios (dimensões) de qualidade da informação, por vezes neste trabalho também é utilizado critério como sinônimo de dimensão, com isso para Naumann and Rolker (2005) são classificados classes, critérios e seus métodos de avaliação conforme Quadro 1.

Quadro 1: Tradução da Classificação de QI (Qualidade da Informação) Critérios de Metadados Naumann and Rolker (2005)

Classe de Avaliação	**QI Critério**	**Método de Avaliação**
Critério Subjetivo (ou do sujeito)	Credibilidade	Experiencia do usuário
	Representação Concisa	Amostragem do Usuário
	Interpretabilidade	Amostragem do Usuário
	Relevância	Avaliação continua do usuário
	Reputação	Experiência do usuário
	Compreendibilidade	Amostragem do usuário
	Valor-Adicionado	Avaliação continua do usuário
Critério Objetivo (ou do objeto)	Completude	Análise, Amostragem
	Suporte ao cliente	Análise, Contrato
	Documentação	Análise
	Objetividade	Opinião de especialistas
	Preço	Contrato
	Confiabilidade	Avaliação Continua
	Segurança	Análise
	Pontualidade	Análise
	Verificabilidade	Opinião de especialistas
Critério do Processo	Precisão (Exatidão)	Amostragem, técnicas de limpeza
	Quantia de dados	Avaliação Continua
	Disponibilidade	Avaliação Continua
	Representação Consistente	Análise
	Latência	Avaliação Continua
	Tempo de resposta	Avaliação Continua.

Fonte: tradução de autor (2024) a partir de Naumann and Rolker (2005)

Para Naumann and Rolker (2005) os critérios escolhidos pelo presente trabalho

podem ser configurados como observado no Quadro 2

Quadro 2: Tradução e localização para o português dos Critérios de Qualidade de Dados Naumann and Rolker (2005).

Dimensão (Critério)	**Tipo**	**Significado**	**Sinônimos**
Precisão	Processo	Quociente entre o número de valores corretos na fonte e o número total de valores na fonte.	qualidade dos dados (em oposição à qualidade da informação), taxa de erro, correção, integridade
Completude	Objeto	Quociente entre o número de itens de resposta e o número de itens do mundo real.	cobertura, escopo, granularidade, abrangência, densidade, extensão
Credibilidade	Sujeito	Grau em que a estrutura da informação corresponde à própria informação.	taxa de erro, credibilidade, confiabilidade
Consistência	Sujeito	Grau em que a informação é aceita como correta.	granularidade de atributos, identificabilidade de ocorrências, consistência estrutural, adequação, precisão de formato

Fonte: tradução de autor (2024) a partir de Naumann and Rolker (2005)

A tradução para português do critério de Consistência definido por Naumann and Rolker (2005) pode por vezes ser chamado de Representação Concisa entretanto está mais próximo de Consistência estrutural e para evitar algumas confusões com estudos em português foi preferido por resumir a Consistência.

Considerando o avanço em estudos de trabalhos anteriores que se relacionam com o presente trabalho chega-se então ao trabalho de Penteado et al. (2021) que propõe um modelo de infraestrutura para a publicação de Dados Abertos Educacionais Conectados (DAEC) que consiste em basicamente a garantia de que os dados abertos educacionais sejam interoperáveis, ou seja, que possuam fácil comunicação entre si tornando-os assim conectados, ênfase do estudo do mesmo é na qualidade e seguindo as melhores práticas de compartilhamento de dados na Web. O desafio abordado é a dificuldade em conectar diferentes fontes de dados devido a problemas de interoperabilidade e alinhamento semântico. O trabalho incluiu um estudo de caso sobre o problema da rotatividade docente, integrando diferentes fontes de dados, e um quase-experimento em que os sujeitos publicaram dados educacionais de maneira consistente, seguindo as melhores práticas em qualidade de dados conectados e compartilhamento de dados. A tese avança na teoria do design em publicação de dados educacionais, desenvolvendo uma infraestrutura de publicação, um modelo de referência semântico para dados educacionais e demonstrando sua aplicabilidade em problemas reais. Como pode ser observado o objetivo de pesquisa

fica claro que se trata de um modelo de infraestrutura para publicação de DAEC de qualidade, e um dos problemas apontados por este é que a dificuldade atual dos dados abertos educacionais está em grande parte na sua dificuldade em interoperabilidade e alinhamento semântico.

Até o presente momento são explorado trabalhos que perpassam pelo objetivo principal do presente trabalho como geração dos metadados, estratégias para avaliação da qualidade dos dados sem necessariamente serem educacionais ou não (mas sempre dados abertos governamentais) utilizando ferramentas da indústria, além de chegar ao estudo mais antigo entretanto divisor de águas no estado da arte onde fornece base teórica para partir na formulação de métricas transparentes que ajudaram na avaliação e finalmente ao analisar o estudo que mais se assemelha ao presente trabalho, que é ele o artigo de Junior and Dorneles (2021) onde aborda a avaliação de dimensões de qualidade de dados no contexto do agronegócio. A qualidade dos dados é crucial para melhorar a precisão das informações e, consequentemente, a assertividade das decisões. O estudo apresenta uma abordagem que utiliza dimensões para a verificação da qualidade dos dados, realizando validação em duas bases de dados reais: uma focada no meio ambiente e outra na agricultura familiar. O conceito de qualidade dos dados é explorado, considerando diferentes definições, incluindo a adequação ao uso e a multidimensionalidade, que se refere aos atributos que representam características específicas. Embora a qualidade dos esquemas também seja reconhecida, a maioria das definições e métricas de qualidade de dados se refere aos valores dos dados. O artigo destaca a falta de uma definição única e subjetividade no conceito de qualidade dos dados, ressaltando a importância de uma abordagem multidimensional. Trabalhos relacionados na literatura são revisados, com ênfase em análises específicas para o agronegócio, diferenciando-se de abordagens em outros domínios.

A abordagem apresentada no artigo identifica dimensões de qualidade de dados de forma abrangente, inicialmente definidas na literatura e refinadas por especialistas na área (que a priori não será o foco do presente trabalho) Na seção de trabalhos relacionados, são apresentadas pesquisas que abordam a qualidade de dados em diferentes contextos, incluindo estruturas completas de qualidade de dados e aplicações específicas na agricultura.

3. Qualidade em Dados Abertos Educacionais

Dessa maneira, a contextualização de trabalhos anteriores se fez necessária para localizar onde este está exatamente, entretanto é necessário se aprofundar nos temas que foram abordados anteriormente e são essenciais para o entendimento do que foi realizado nos experimentos, os próximos parágrafos respectivamente sobre qualidade de dados, dados abertos educacionais, metadados e relação entre metadados e critérios (dimensões) de qualidade vão tratar sobre o aprofundamento dos temas.

A qualidade de dados é um aspecto que pode ser crítico em contextos de dados educacionais. Esta seção aborda a definição e dimensões da qualidade de dados, evidenciando a importância desse conceito na tomada de decisões sejam educacionais ou não. Então discutisse a relação entre a qualidade de dados e ajuda na tomada de decisão. Conforme já mencionado anteriormente qualidade de dados pode ser definida como o dado que esteja apto para ser usado mediante as necessidades do consumidor desse dado Wang and Strong (1996), ou seja, qualidade é um conceito relativo e que vai depender do con-

texto do dado como já demonstrado também por Tayi and Ballou (1998). O conjunto de características destes dados podem ser chamados de dimensões ou ainda critérios (por vezes no trabalho os dois termos são usados como sinônimos).

"Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa – sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras."segundo a Open Knowledge Brasil (2024) e sabe-se que o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) utiliza o Open Knowledge Brasil (2024) como referencia para disponibilização destes dados no contexto educacional brasileiro assim citado em Ferreira et al. (2021) dados abertos educacionais abrem grandes possibilidades de análises e sua melhoria de qualidade pode garantir tomadas de decisões pedagógicas melhores na construção de políticas públicas educacionais.

Faz-se claro o sentido genérico quanto ao termo então para fins de pragmatismo foi usado a definição emprestada da ciência da informação para o termo quando refere-se a dados estruturados que descrevem características das entidades portadoras de informações (registros) Satija et al. (2020), por vezes também resumido como "dados sobre dados"ou "informações sobre informações". Existem ainda os tipos de metadados levantados também por Satija et al. (2020) entretanto não foi aprofundado neste trabalho, o tipo é autodeclarativo a natureza do experimento. Além disso destaca-se a importância dos metadados na interoperabilidade e reusabilidade de dados educacionais, juntamente com uma análise das normas e padrões relevantes para metadados, como exemplificado por Alves et al. (2010).

Explorando a interconexão entre os metadados gerados nesse estudo e a qualidade dos dados em si, discute como os metadados podem influenciar ou refletir os critérios de qualidade. Segundo Satija et al. (2020) os metadados estão cada vez mais sendo utilizados como fonte de ângulos diferentes na avaliação de dados conectados assim como no estudo de Penteado et al. (2021) em que a perspectiva de dados conectados é trazido à tona, No presente trabalho é feita então a junção de paradigmas onde se tem dados abertos educacionais conectados ou não sendo avaliados por critérios (dimensões) de qualidade baseados em metadados.

4. Ferramentas e Métodos

Antes de detalhar os métodos utilizados, é necessário contextualizar a ferramenta principal aplicada neste trabalho: o Apache Airflow [1] "é uma plataforma de código aberto para desenvolver, agendar e monitorar fluxos de trabalho orientados em lote. A estrutura Python extensível do Airflow permite criar fluxos de trabalho conectados a praticamente qualquer tecnologia. Uma interface web ajuda a gerenciar o estado dos seus fluxos de trabalho. O Airflow pode ser implantado de várias maneiras, variando desde um único processo em seu laptop até uma configuração distribuída para oferecer suporte até mesmo aos maiores fluxos de trabalho"

[1]Acesse a ferramenta em `https://airflow.apache.org/docs/apache-airflow/stable/index.html`.

Quadro 3: Comparativo entre Apache Airflow e outras ferramentas.

Ferramenta	**Vantagens**	**Desvantagens**
Apache Airflow	- Código aberto e altamente extensível - Suporte a múltiplas tecnologias (Python, Bash, SQL, etc.) - Interface web intuitiva para monitoramento de fluxos - Facilita a modularização de DAGs, permitindo expansão futura	- Curva de aprendizado para configurações avançadas - Desempenho pode ser afetado em pipelines muito complexas
Luigi	- Bom para fluxos de dados pesados e pipelines ETL - Fácil configuração inicial para tarefas simples - Permite dependências complexas	- Escalabilidade limitada comparada ao Airflow - Interface de monitoramento menos intuitiva
Prefect	- Interface web amigável para monitoramento - Suporte a fluxos de longa duração - Simplicidade no gerenciamento de tarefas falhas	- Comunidade e suporte menores comparados ao Airflow - Dependência da versão paga (Prefect Cloud) para recursos avançados
Dagster	- análise e observabilidade dos pipelines - Ferramentas nativas para monitorar fluxos - Boa modularização de pipelines	- Comunidade menor e menos madura - Integrações limitadas comparadas ao Airflow

Fonte: Autor (2024)

A ferramenta em questão além de ser de código aberto permite uma integração maior diante das possibilidades apresentadas nos trabalhos futuros. No trabalho de Oliveira et al. (2023) a escolha por uma ferramenta em especifico e não por um método deixou claro certas limitações quanto a exploração de possibilidades, com isto o presente trabalho tenta remover esses possíveis limites ao escolher o Airflow. Como citado na seções anteriores os metadados têm papel fundamental no desenvolvimento deste trabalho, com isso o primeiro passo necessário é o de geração desses arquivos de metadados a partir das tabelas escolhidas seguindo os seguintes passos:

1. Quais tabelas são similares (ou possuem um grau de similaridade alta) entre ambas instituições, a análise aqui foi feita de forma subjetiva considerando o escopo de macro tema, por exemplo, na Universidade Federal do Rio Grande do Norte (UFRN) não temos literalmente uma tabela que se chama Liquidações conforme a Universidade Federal Rural de Pernambuco (UFRPE), entretanto temos outra tabela com dados que servem a propósitos análogos na gestão e acompanhamento

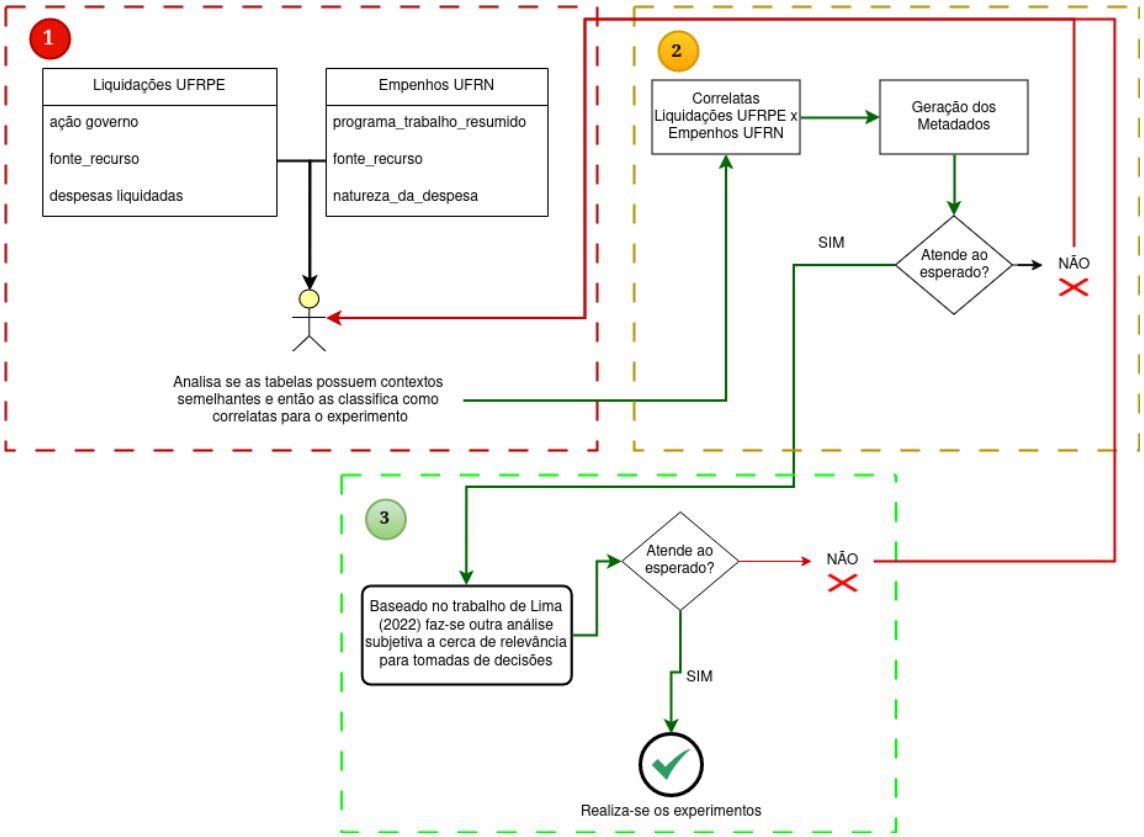
de recursos que se chama Empenhos;

2. Dado o grau de similaridade, essas tabelas são possíveis de metrificação (podem ser inseridas como entrada nas fórmulas propostas) com os metadados extraídos atualmente? Se sim, então essa tabela é forte candidata ao experimento;

3. E por fim considerando o grau de importância subjetiva para tomada de decisões de politicas publicas Lima (2022).

Estas etapas estão simplificadas na Figura 2 abaixo:

Figura 2. Fluxograma das etapas de escolha das tabelas candidatas ao experimento.



Fonte: autor (2024)

Estas então são as tabelas escolhidas por esse método, oriundas de um processo de padronização no qual pode-se visualizar no *Quadro 4* e o final do processo podem ser vistas no *Quadro 5*.

Quadro 4: Tabelas Originais como consta nos portais das IES, Tabelas padronizadas para o experimento e IES.

Tabelas Originais	Tabela Padronizada	IES
Liquidações	liquidacoes	UFRPE
Componentes por Currículo de Graduação	componentes_por_currículo	UFRPE
Quantitativo de alunos de graduação	qtd_alunos_graduacao	UFRPE
Matriculados nas turmas de graduação (2021.1)	matriculados_turma_graduacao	UFRPE
Microdados Censo - Cursos	censo_cursos	UFRPE
Cursos de Graduação	ensino_de_graduacao	UFRPE
Empenhos	liquidacoes	UFRN
Estrutura Curricular	componentes_por_currículo	UFRN
Ingressantes nas turmas de graduação (2023.1)	matriculados_turma_graduacao	UFRN
Cursos de Graduação	ensino_de_graduacao	UFRN

Fonte: autor (2024)

Considerado que ambas as IES não teriam necessariamente o mesmo nome para todas as tabelas, fez-se necessária essa correlação por escopo de atuação de cada tabela e criando assim uma padronização para efeitos do experimento.

Quadro 5: Tabelas escolhidas, portais eletrônicos acessados.

Tabelas Escolhidas	Portal Eletrônico Acessado
liquidacoes	UFRPE
componentes_por_currículo	UFRPE
qtd_alunos_graduacao	UFRPE
matriculados_turma_graduacao	UFRPE
censo_cursos	UFRPE
ensino_de_graduacao	UFRPE
liquidacoes	UFRN
componentes_por_currículo	UFRN
matriculados_turma_graduacao	UFRN
ensino_de_graduacao	UFRN

Fonte: autor (2024)

O acesso ao dados tanto da UFRPE quanto da UFRN foram realizados 16/12/2023 e estão disponíveis respectivamente em *https://dados.ufrpe.br/* e *https://dados.ufrn.br/*.

4.1. Da execução do trabalho

A execução do trabalho envolveu diversas etapas essenciais para a análise e criação de metadados a partir de dados coletados no formato *Comma separated values (CSV)*, conforme já explicitados na seção anterior, no qual essas etapas são elas: (i) coleta de dados, (ii) análise e criação dos metadados, (iii) estrutura dos metadados, (iv) análise exploratória e (v) escolha dos critérios de qualidade. Que estão respectivamente detalhados nos parágrafos seguintes.

(i) Os dados utilizados neste estudo foram coletados a partir de um arquivo CSV e neles contém as informações retiradas do portal de dados abertos da UFRPE e UFRN conforme explicitado na Figura 2 e nos Quadros 4 e 5.

(ii) A análise para criação dos metadados foi realizada por meio da implementação de um Grafo Aciclico Direcionado (GAD) ou por vezes comumente referenciado sua sigla em inglês de *Direct Acyclic Graph* (DAG) utilizando o Apache Airflow. A DAG, denominada 'metadados_dag', foi configurada para ser executada com os seguintes parâmetros:

```
# Dags params passiveis de automatização futura
default_args = {
    'owner': 'airflow',
    'start_date': datetime(2023, 1, 1),
    'retries': 1,
}
```

(iii) Esses parâmetros da DAG são mais relevantes quando se tem uma arquitetura inserida em um ecossistema já maduro, ou seja, que possuam várias DAGs e seus argumentos acabam por ter que ser orquestrado com todo o ecossistema de DAGs, a tangibilidade disso é citado nos trabalhos futuros. No contexto técnico do Airflow é utilizado o **Operador Python (PythonOperator)**, os operadores python em suma são executados para realizar chamadas python, então é executado a função gerar_metadados durante a execução da DAG. Esta função carrega o arquivo CSV, coleta metadados como tipos de dados, estatísticas descritivas e informações sobre valores nulos e distintos, e salva essas informações em um arquivo JavaScript Object Notation (JSON), boa parte disso é desempenhada pela biblioteca pandas que faz a manipulação desse CSV. O código completo pode ser visto nos Apêndices.

Os metadados coletados estão organizados na seguitente estrutura e incluem:

1. Nome da tabela - [Nome da tabela padronizada];
2. Número de linhas - [Número total de linhas na tabela];
3. Número de colunas - [Número total de colunas na tabela];
4. Tipagem das colunas - [Dicionário contendo os tipos de dados de cada coluna];
5. Contador Tipagem - [Dicionário contendo a contagem de cada tipo de dado na tabela];
6. Estatísticas descritivas - [Dicionário contendo estatísticas descritivas para as colunas numéricas];
7. Nome das colunas - [Lista dos nomes das colunas na tabela];
8. Valores nulos por coluna - [Dicionário contendo a contagem de valores nulos para cada coluna];
9. Valores distintos por coluna - [Dicionário contendo a contagem de valores distintos para cada coluna].

(iv) Na etapa de análise exploratória dos dados, utilizamos o Pandas para carregar os arquivos CSV e transformá-los em DataFrames, facilitando a visualização e o entendimento das informações. Observamos que os dados da UFRPE apresentam uma granularidade maior, possivelmente em decorrência de adequações à LGPD, que exige um tratamento mais detalhado das informações sensíveis. Em contraste, a UFRN exibe dados com granularidade menor, sugerindo uma possível não conformidade com a legislação.

Essa diferença impacta diretamente as porcentagens das dimensões de qualidade, podendo causar distorções nos resultados.

(v) A partir do estudo de Penteado et al. (2021) onde vai remontar quase que constantemente as principais dificuldades quando falado de operação de DAGC que é aplicado para o contexto do presente trabalho aos DAE ao tratar sobre interoperabilidade e alinhamento semântico como um dos principais problemas na disponibilização de dados abertos, dado essa premissa levantada pelo estudo citado é identificado dimensões (critérios) baseado nisso, para isso é referenciado aos critérios criados por Naumann and Rolker (2005) conforme pode ser visto no Quadro 1. Dessa maneira chega-se a:

1. Precisão: É a garantia que os dados possam ser compreendidos e utilizados de maneira consistente entre diferentes sistemas e plataformas. Além de certificar-se de que os significados dos termos e conceitos nos dados são consistentes, promovendo uma interpretação precisa e unívoca.
2. Completude: Assegurar que os dados transferidos entre sistemas abranjam todas as informações necessárias, evitando perdas de dados importantes durante a interoperabilidade. Garantindo assim que todos os conceitos relevantes estejam representados de maneira completa e que não haja lacunas na compreensão semântica. Exemplo.: Falta de linhas onde dificulte a compreensão se realmente tal coluna é um texto.
3. Consistência: Manter consistência nas representações e formatos de dados para facilitar a integração e a interpretação adequada em diferentes sistemas. Com fim de evitar contradições e inconsistências nas definições de termos e conceitos, promovendo uma visão coesa dos dados. Exemplo.: É esperado que um ID seguindo essa logica ou um int64 possua variabilidade alta dada que pode ser um identificador entretanto campos do tipo string que variam muito já que por vezes podem assumir qualquer tipagem de dado, exemplo '2024-01-01', onde representa uma data porém é um tipo string.
4. Credibilidade: Garantir que os dados possam ser confiavelmente trocados entre sistemas, sem perda de integridade ou distorção. Reforçando a confiança na interpretação dos dados, assegurando que diferentes partes possam concordar sobre o significado dos termos utilizados.

4.2. Métricas de Avaliação

Os metadados podem dizer muito além daquilo que por vezes foi considerado no estado da arte antes do estudo de Ryu et al. (2006), então a definição destas métricas considera cada um dos critérios (dimensões) e suas especificidades com a extração daquilo que é material, ou seja, as métricas são produzidas por aquilo que se tem disponível de metadados, assim segue-se as práticas também observadas em Junior and Dorneles (2021) no qual precisa-se por vezes definir objetivamente como será calculado cada dessas métricas.

Com isso pode-se definir no nosso estudo que credibilidade é igual ao número de colunas menos a discrepância de tipagem divididos pelo N número de colunas. Onde um tipo de sujeito e o seu grau é dado onde a estrutura da informação corresponde à própria então pode-se inferir que ela é avaliada pela correspondência entre os tipos de dados reais das colunas e os tipos esperados, conforme definido no dicionário de dados. A discrepância de tipagem ocorre quando o tipo de dado identificado nos metadados diverge do tipo esperado. Quanto maior a correspondência entre os tipos, maior a credibilidade

da base de dados. Esta métrica é expressa em termos percentuais, onde valores mais próximos de 100% indicam alta credibilidade.

Completude é igual ao total de valores não nulos divido pela multiplicação de número de linhas e número de colunas. Para Naumann and Rolker (2005) ao categorizar o tipo da dimensão em objetiva (ou de objeto) pode-se definir que o objeto (entidade) de uma coluna são os seus registros, ou seja, a completude de dados em uma coluna vezes a linha, portanto, pode ser interpretada como a proporção de registros que possuem valores não nulos em relação ao total de registros. Essa métrica reflete a integridade dos dados, assegurando que as colunas contêm as informações necessárias para análise. Valores próximos a 100% indicam que poucas ou nenhuma informação foi perdida, o que é essencial para garantir a utilidade dos dados.

Consistência é igual ao valor 1 subtraido do resultado da divisão das inconsistências pelo número de colunas. No qual inconsistências representa a soma das colunas com alta variabilidade e a consistência dos dados então é determinada pela uniformidade dos valores dentro das colunas, especialmente em colunas onde se espera baixa variabilidade, como IDs ou campos categóricos. A fórmula penaliza colunas onde há grande variabilidade ou onde os dados não correspondem ao tipo esperado. Isso assegura que a estrutura dos dados seja estável e previsível, o que é crucial para a interoperabilidade de sistemas apontada por Penteado et al. (2021), ainda sobre a fórmula o 1 é a representação matemática para enfatizar a consistência. Se não houver inconsistências (inconsistências=0), a fórmula retornará 100, indicando consistência completa. Se todas as colunas forem inconsistentes, a fórmula retornará 0.

Precisão é igual a divisão das inconsistências dos outliers pelo número de colunas. A precisão, no contexto deste trabalho, refere-se à medida em que os valores dentro das colunas correspondem aos padrões esperados, sem a presença de outliers ou anomalias significativas. A detecção de outliers é realizada para cada coluna, utilizando estatísticas descritivas como média e desvio padrão extraídas dos metadados. A soma das colunas que apresentam valores discrepantes é usada para calcular a precisão através da fórmula. Com isso para Naumann and Rolker (2005) o método de avaliação desse critério envolve técnicas de limpeza então ao remover inconsistências garantimos assim uma padronização então essa fórmula consiste da ideia de que quanto menos inconsistências tiver mais preciso está sendo.

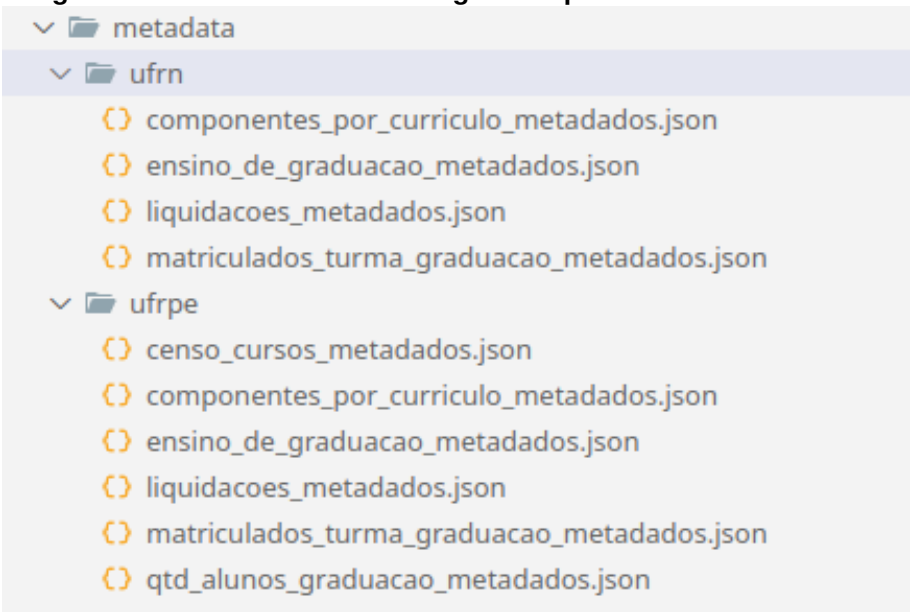
5. Resultados e Discussão

Nesta seção é discutido os resultados e suas interpretações no que concerne a contribuição deste trabalho, além dos gráficos utilizando a biblioteca do matplotlib que basicamente “é uma biblioteca abrangente para criar visualizações estáticas, animadas e interativas em Python”segundo Matplotlib (2024) e do Seaborn que “é uma biblioteca de visualização de dados Python baseada em matplotlib. Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos ”segundo Seaborn (2024).

O código está hospedado no GitHub (plataforma de hospedagem de código-fonte e arquivos com controle de versão usando o Git. Ele permite que programadores, utilitários ou qualquer usuário cadastrado na plataforma contribuam em projetos privados e/ou Open Source de qualquer lugar do mundo.) e disponivel em *https://github.com/-JuanMenezes/meta-qd*

Os metadados são gerados e armazenados conforme o diretório de origem, ou seja, o diretório com nome UFRPE gerará uma hierarquia de arquivos semelhante, o mesmo acontecerá com UFRN, conforme pode ser visto na Figura 3:

Figura 3. Diretório de metadados gerados pela DAG codificada no airflow.



Fonte: autor (2024)

Na seção anterior pôde ser visto um exemplo de estrutura de metadados gerados, o tipo de arquivo salvo escolhido foi o JSON pela sua boa capacidade de lidar com dados aninhados além de fácil interpretação sem necessitar de uma ferramenta de visualização de dados, um exemplo de um JSON do experimento gerado pode ser observado nos Apêndices. Além disso também é apresentado a Tabela 1 para facilitar compreensão dos resultados das IES do experimento, outro entendimento que também se faz necessário para entendimento da análise é a classificação daquilo que é considerado alta, média e baixa para o presente trabalho, isto pode ser visto na Tabela 2. Vale salientar que deve-se ter cautela ao interpretar cada faixa percentual a depender do critério e contexto.

Tabela 1: Tabelas padronizadas para o experimento com seus códigos.

Código Tabela	Tabela Padronizada	IES
1	componentes_por_currículo	UFRPE
2	qtd_alunos_graduacao	UFRPE
3	ensino_de_graduacao	UFRPE
4	liquidacoes	UFRPE
5	matriculados_turma_graduacao	UFRPE
6	censo_cursos	UFRPE
1	componentes_por_currículo	UFRN
2	ensino_de_graduacao	UFRN
3	liquidacoes	UFRN
4	matriculados_turma_graduacao	UFRN

Fonte: autor (2024)

Tabela 2: Tabela de classificação das faixas de percentuais obtidos no experimento.

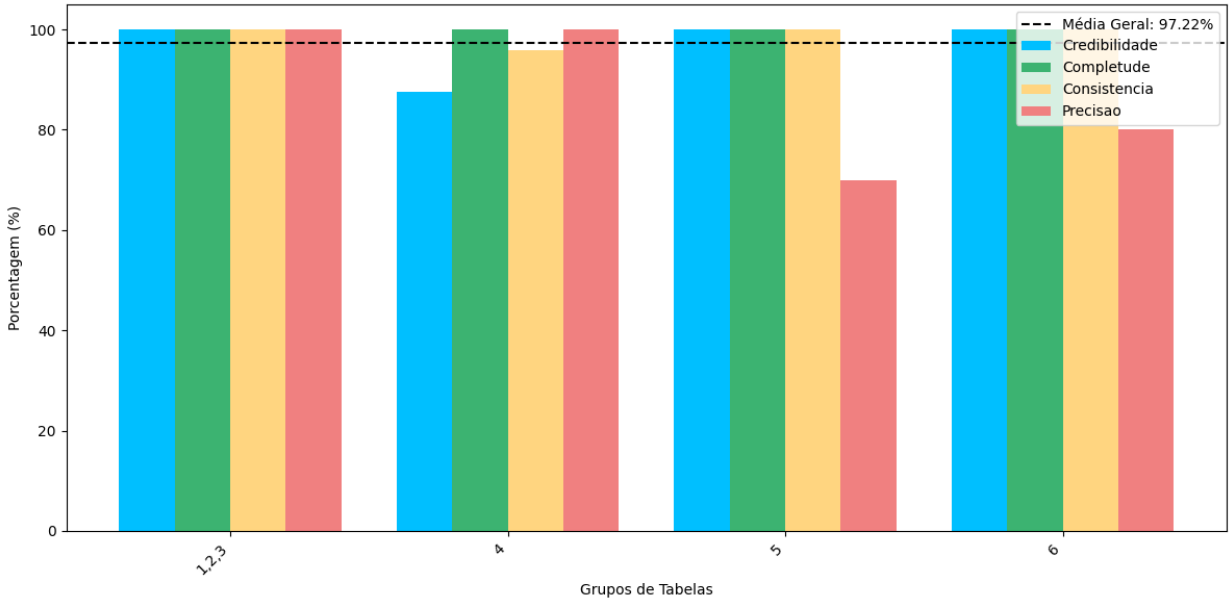
Classificação	Faixa Percentual
Muito alta	100%
Alta	80% - 99%
Média	61% - 79%
Baixa	21% - 60%
Muito baixa	0% - 20%

Fonte: autor (2024)

5.1. Cenário UFRPE

Diante disto exposto, chega-se a apresentação dos resultados por cenários e IES, o contexto da UFRPE para o Plano De Dados Abertos (PDA) é de uma IES que possui um PDA recente, como pode ser visto nos registros em Dados Abertos Gov BR (2024) grande parte dos conjuntos de dados foram apenas adicionados 1 ano atrás e ainda com espaços para melhorias e maturação principalmente no que discerne a tratamento de informações. Seguindo por cada tabela tem-se que:

Figura 4. Métricas agrupadas por tabela UFRPE



Fonte: autor (2024)

Conforme ilustrado na Figura 4, as tabelas de código 1, 2 e 3 da UFRPE (componentes por currículo, qtd alunos graduação e ensino de graduação) apresentaram um desempenho consistente, com todas as métricas de credibilidade, completude, consistência e precisão classificadas como Muito alta, atingindo 100% em cada uma delas. Este desempenho reflete tabelas robustas e confiáveis, sem inconsistências, valores nulos ou anomalias significativas, tornando-a altamente adequada para interoperabilidade de análises educacionais, como proposto por Penteado et al. (2021).

Quanto à tabela de código 4 (liquidacoes), observou-se uma ligeira redução na credibilidade, que foi classificada como Alta com 87,5%, e na consistência, também clas-

sificada como Alta com 95,83%. A completude permaneceu alta em 99,99%, enquanto a precisão manteve-se Muito alta em 100%. A pequena redução na credibilidade e consistência pode sugerir a presença de algumas colunas com tipagem ligeiramente divergentes do esperado e uma pequena variabilidade nos dados, mas ainda assim, os dados permanecem altamente confiáveis.

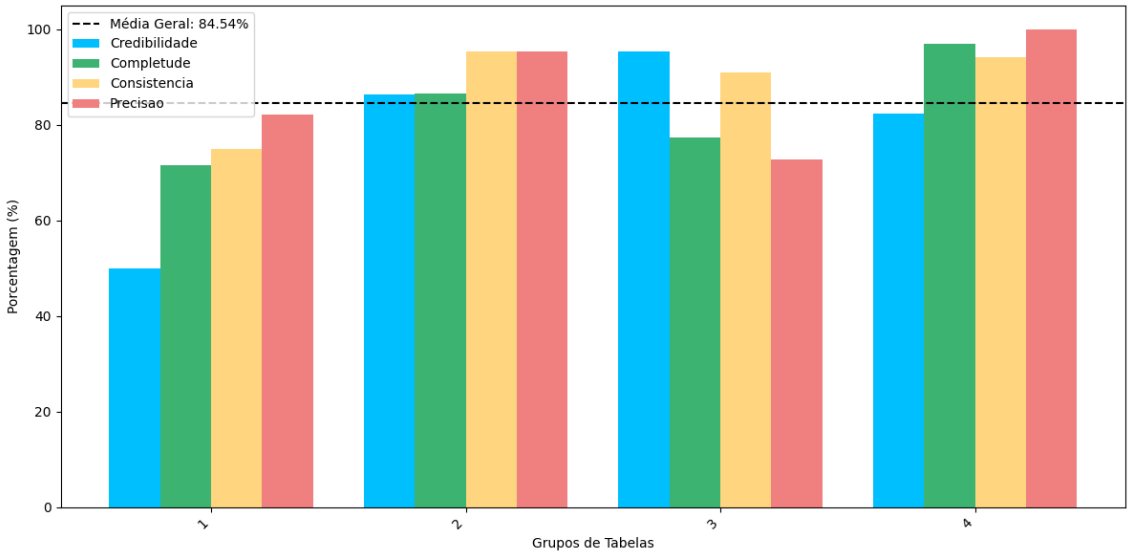
A tabela de código 5 (matriculados turma graduação) apresentou credibilidade, completude e consistência classificadas como Muito alta com 100%, enquanto a precisão foi classificada como Alta com 70%. A menor precisão pode indicar a presença de outliers ou valores atípicos que podem impactar a interpretação dos dados, mesmo que a estrutura e a tipagem estejam corretas.

Por fim, na tabela de código 6 (censo cursos), também mostrou credibilidade, completude e consistência classificadas como Muito alta com 100%, com uma precisão ligeiramente inferior, classificada como Alta com 80%. Embora a maioria dos dados esteja em conformidade com os padrões esperados, a presença de alguns outliers reduz a precisão, mas ainda assim, os dados são altamente utilizáveis e também nota-se que a quantidade de colunas pode alterar bastante essa métrica de precisão.

5.2. Cenário UFRN

O Contexto da UFRN já é diferente do apresentado anteriormente visto, dado que conforme os registros em Dados Abertos Gov BR (2024) os conjuntos de dados são adicionados e atualizados há mais de 4 anos o que pode indicar certa maturidade em relação aos dados da IES avaliada anteriormente. No cenário UFRN foram obtidos:

Figura 5. Métricas agrupadas por tabela UFRN



Fonte: autor (2024)

Começando pelo que pode ser visto na Figura 5, a tabela de código 1 (componentes por currículo), observa-se uma credibilidade média de 50%, o que sugere que metade das colunas está em conformidade com a tipagem esperada, mas ainda existem discrepâncias que podem impactar a interoperabilidade dos dados. A completude foi média

em 71,6%, indicando que há uma quantidade significativa de valores nulos que podem comprometer a integridade dos dados. A consistência foi média em 75%, refletindo uma boa uniformidade na tipagem e na estrutura dos dados. A precisão, por outro lado, foi alta em 82,14%, destacando que a maioria dos valores está dentro dos padrões esperados, apesar de algumas colunas apresentarem outliers.

Na tabela de código 2 (ensino de graduação), a credibilidade foi alta em 86,36%, indicando uma conformidade satisfatória com a tipagem esperada dos dados. A completude foi alta em 86,63%, mostrando que a maioria dos registros está completa, com poucos valores nulos. A consistência foi alta em 95,45%, sugerindo uma forte uniformidade nos dados, sem grandes variações na estrutura. A precisão também foi alta em 95,45%, indicando que os valores seguem de forma consistente os padrões esperados, com mínima presença de outliers.

Na tabela de código 3 (liquidações), a credibilidade foi alta em 95,45%, o que sugere que quase todas as colunas estão bem tipificadas e prontas para interoperabilidade. No entanto, a completude foi média em 77,3%, apontando a presença de valores nulos que podem impactar a análise. A consistência foi alta em 90,91%, refletindo pouca variabilidade indesejada nas colunas, enquanto a precisão foi média em 72,73%, o que mostra que mais da metade dos dados está dentro dos padrões esperados, embora ainda existam outliers que precisam ser considerados.

Por fim, na tabela de código 4 (matriculados turma graduação), a credibilidade foi alta em 82,35%, sugerindo que a maioria dos dados está bem tipificada. A completude foi alta em 97,05%, indicando que quase todos os registros estão completos. A consistência foi alta em 94,12%, mostrando uma uniformidade robusta na estrutura dos dados. Notavelmente, a precisão foi muito alta em 100%, o que significa que não foram detectados outliers ou valores fora dos padrões esperados, garantindo a máxima confiabilidade dos dados para análises detalhadas.

5.3. Cenário UFRPE x Cenário UFRN

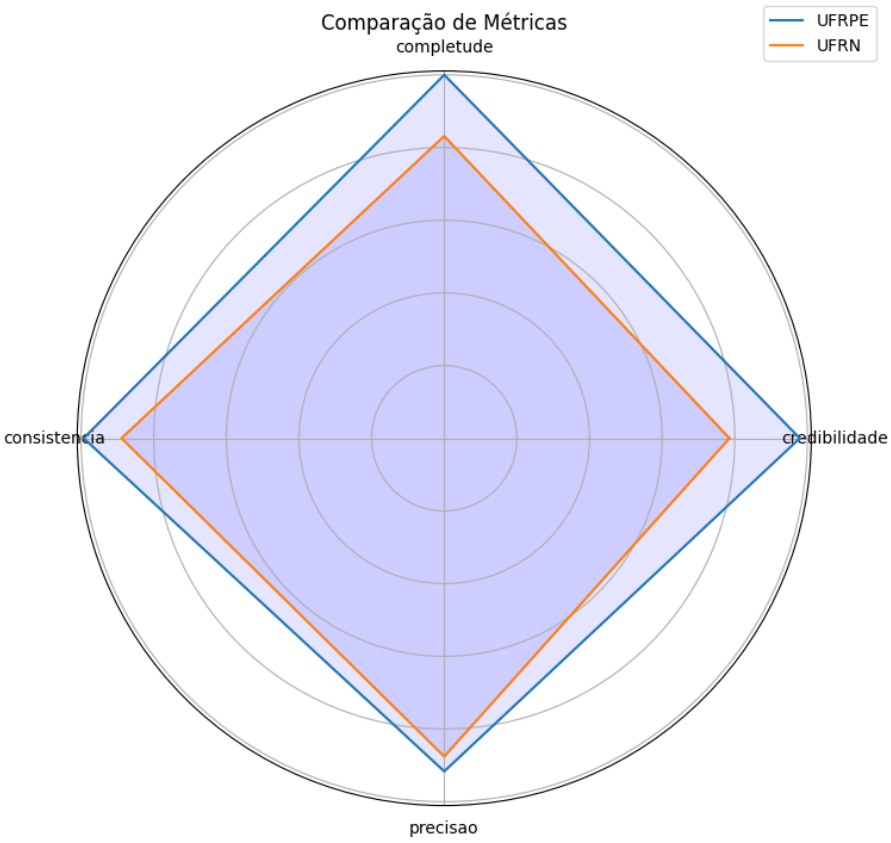
Dessa maneira, em termos comparativos entre UFRPE e UFRN umas de suas principais diferenciações é no campo de maturidade de disponibilização desses dados conforme citado informação do Dados Abertos Gov BR (2024), entretanto ficou evidenciado que isso não necessariamente implicará em uma maior qualidade pelo fato de que o entendimento da natureza do dado por quem o disponibiliza é crucial, visto o caso de tabelas da UFRN onde um agrupamento a tornaria mais completa. No caso especifico da UFRPE a dimensão com menor pontuação é a precisão por um erro de declaração da tipagem mas no geral todas dimensões tem níveis altos de qualidade, enquanto que a UFRN a dimensão com menor pontuação é a credibilidade pelo alto nivel de discrepancia de tipagem.

Trazendo luz média de cada dimensão por IES onde notou-se alguma ponderação, temos que Credibilidade na UFRPE apresentou uma média significativamente superior (97,92%) em comparação com a UFRN (78,54%). A diferença de quase 20 pontos percentuais pode indicar que a UFRN enfrenta desafios na conformidade com os padrões de dados estabelecidos, possivelmente devido a uma falta de padronização ou inconsistências na aplicação dos dicionários de dados, precisa-se investigar com mais bases de dados para chegar em tal afirmação entretanto é uma hipótese para tal disparidade. Enquanto que para Completude dos dados da UFRPE é quase perfeita (99,99%), enquanto a UFRN

apresentou uma média de 83,15% esta disparidade possivelmente pode estar acontecendo devido a lacunas nos processos de verificação de qualidade ou na falta de mecanismos automáticos de preenchimento ou tratamento de valores nulos.

Dessa maneira, vale salientar que na Figura 6 nota-se visualmente aquilo descrito no paragrafo anterior, nessa análise e discussão a de se fazer um pequeno adendo em relação as porcentagens que não necessariamente são simples indicadores de muito alto (100%) e muito baixo (0%) tem de se considerar o cenário da análise realizada, por exemplo, se para a formulação de politicas públicas como o bolsa permanência nas universidades a consistência muito alta é fundamental para garantir que os critérios de elegibilidade sejam aplicados de forma uniforme e justa para os estudantes daquela IES Lima (2022).

Figura 6. Gráfico de radar comparativo entre UFRPE e UFRN com média dentro de cada métrica (dimensão).



Fonte: autor (2024)

6. Considerações Finais

Dessa maneira, entendendo que os DAE disponibilizados pelas IES Públicas têm potencial de serem norteadores para políticas de desenvolvimento do pais Lima (2022). Analisar os metadados gerados a partir de tabelas disponíveis fez com que o trabalho obtivesse êxito ao responder se o gerenciamento de metadados (ou a falta deles) pode levar a uma melhora da qualidade dos dados, por exemplo, no cenário da UFRN onde o não agrupamento dos dados levou a uma queda na credibilidade percentualmente em comparação a UFRPE e também pôde-se verificar no experimento que mesmo ambas UFRN e UFRPE

possuindo um Plano de Dados Abertos (PDA) mas não possuindo esquemas de metadados disponíveis e de fácil acesso (data de tentativa de último acesso: 10 fev 2024) se tornou necessário então entender o que os dados dizem por si só gerando assim métricas a partir das tabelas disponíveis, que por si só foi o mais desafiador. Enquanto levanta-se a questão dos critérios escolhidos acabarem sendo quanto percentualmente melhor quando altera-se o tipo de tabela tem de se levar em consideração que os percentuais de cada critérios são indicativos e não taxativos, onde 100% não necessariamente equivale a pode usar irrestritamente e 0% não poder usar de forma alguma. Isso fica claro quando é levado em comparação com a tabela de código 6 do cenário da UFRPE cujo foi retirada do censo, a depender do cenário é tolerável que haja imprecisões nos dados dentro de um limite mas em contra-partida a ser sempre constante. Portanto, em programas de política pública onde a justiça e a equidade são fundamentais, a consistência na aplicação das regras e critérios pode ser mais crucial do que a precisão absoluta dos dados utilizados.

Ao levantar a questão sobre os critérios (dimensões) de qualidade na área de Dados Abertos Educacionais (DAE), é necessário considerar a importância de estudos em áreas estratégicas para o país, como o agronegócio, conforme discutido por Junior and Dorneles (2021). Um dos principais desafios enfrentados neste trabalho foi a falta de padronização entre as Instituições de Ensino Superior (IES) analisadas. Esse problema tornou necessário um entendimento mais profundo das tabelas, por exemplo, analise exploratória dos dados. De modo a permitir a comparação dos percentuais de critérios previamente definidos. Devido à limitação de tempo para a realização deste estudo, foi possível escolher e analisar quatro critérios (dimensões) e dez tabelas, processo ilustrado na Figura 2. Embora outras tabelas e critérios pudessem ser explorados, esta delimitação permitiu uma análise mais detalhada dentro do escopo do trabalho.

Conclui-se que a avaliação da qualidade dos Dados Abertos Educacionais (DAE) é complexa e multifacetada. Critérios como credibilidade e completude mostraram variações significativas entre as IES analisadas. A UFRPE, ao realizar o agrupamento adequado de dados para se adequar à LGPD, garantiu uma maior organização e qualidade das tabelas, enquanto a UFRN, sem esse agrupamento, apresentou menor credibilidade. A falta de padronização de metadados entre as IES foi um dos principais desafios enfrentados, destacando a importância de práticas adequadas de gerenciamento de dados e metadados. No entanto, os critérios de avaliação devem ser interpretados de forma indicativa, e não taxativa, já que uma pontuação de 100% não implica uso irrestrito dos dados, assim como 0% não significa inutilidade total. Para estudos futuros, recomenda-se a ampliação do número de critérios e tabelas analisados, além da adoção de Modelos de Linguagem de Grande Escala (LLMs), para automatizar a criação e verificação de metadados. Esses modelos podem ser usados para detectar inconsistências e melhorar a qualidade dos dados já nos metadados. Assim, a conformidade legal, o aprimoramento técnico e o uso de novas tecnologias são essenciais para garantir a confiabilidade e relevância dos DAE, com impacto direto em políticas públicas e projetos estratégicos para o país.

Referências

R. C. V. Alves et al. Metadados como elementos do processo de catalogação. 2010.

Dados Abertos Gov BR. Dados abertos gov br. Disponível em: `https://dados.gov.br/dados/organizacoes`, 2024. Acesso em Data de Acesso.

L. A. Ferreira, R. L. Rodrigues, and R. N. de Souza. Dados abertos educacionais brasileiros: Um mapeamento sistemático da literatura. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 1186–1195, 2021.

Great Expectation. Great expectation. Disponível em: `https://greatexpectations.io/`, 2024. Acesso em Data de Acesso.

C. S. Junior and C. F. Dorneles. Avaliaçao de dimensoes de qualidade de dados para o agronegócio. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 283–288. SBC, 2021.

B. K. Kahn, D. M. Strong, and R. Y. Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4):184–192, 2002.

M. S. Lima. Acesso aos dados abertos das universidades federais a partir dos indicadores de fluxo do ensino superior: análise e recomendações ao accountability. 2022.

Matplotlib. Matplotlib. Disponível em: `https://matplotlib.org/`, 2024. Acesso em Data de Acesso.

F. Naumann and C. Rolker. *Assessment methods for information quality criteria*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät ..., 2005.

G. P. Oliveira, B. M. Mendes, C. A. Bacha, L. L. Costa, L. D. Gomide, M. O. Silva, M. A. Brandão, A. Lacerda, and G. L. Pappa. Assessing data quality inconsistencies in brazilian governmental data. *Journal of Information and Data Management*, 14(1), 2023.

Open Knowledge Brasil. Título do site. Disponível em: `https://ok.org.br/dados-abertos/`, 2024. Acesso em Data de Acesso.

B. E. Penteado, J. C. Maldonado, and S. Isotani. Modelo de infraestrutura para publicação de dados abertos educacionais conectados de qualidade. In *Anais dos Workshops do X Congresso Brasileiro de Informática na Educação*, pages 01–10. SBC, 2021.

K.-S. Ryu, J.-S. Park, and J.-H. Park. A data quality management maturity model. *ETRI journal*, 28(2):191–204, 2006.

M. Satija, M. Bagchi, and D. Martínez-Ávila. Metadata management and application. *Library Herald*, 58(4):84–107, 2020.

Seaborn. Seaborn. Disponível em: `https://seaborn.pydata.org/`, 2024. Acesso em Data de Acesso.

G. K. Tayi and D. P. Ballou. Examining data quality. *Communications of the ACM*, 41 (2):54–57, 1998.

R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

A. Zuiderwijk, M. Janssen, S. Choenni, R. Meijer, and R. S. Alibaks. Socio-technical impediments of open data. *Electronic Journal of e-Government*, 10(2):pp156–172, 2012.

A. Apêndice

Códigos e utilitários