Uma Abordagem para a Gestão da Linhagem de Dados Heterogêneos

Hudson A. B. da Silva[1], José E. M. Jochem[1], João V. dos Santos[1], Eduardo F. R. de Sousa[1], Ronaldo dos S. Mello[1], Carina F. Dorneles[1], Renato Fileto[1]

[1]Departamento de Informática e Estatística - UFSC - Florianópolis - SC - Brasil

hudson.silva@ifpa.edu.br, joseeduardomj@gmail.com, santosjoao301@gmail.com, dudfuul@gmail.com, r.mello@ufsc.br, carina.dorneles@ufsc.br, r.fileto@ufsc.br

Abstract. *Audit and data governance in projects involving multiple sources require reliable traceability to ensure data quality and trust throughout their lifecycle. Data lineage is an essential tool to map the origin, transformations, and destinations of data. This work proposes an approach to define and implement the complete trajectory of data over time. It uses standardized metadata based on the Dublin Core, enriched to capture and persist the history of transformations. The approach involves modeling specific metadata to track ETL (Extraction, Transformation, and Load) operations and proves feasible in a project utilizing real-world data.*

Resumo. *A Auditoria e a governança de dados em projetos com múltiplas fontes exigem rastreabilidade confiável, garantindo a qualidade e confiabilidade dos dados ao longo de seu ciclo de vida. A linhagem de dados surge como uma ferramenta essencial para mapear a origem, transformações e destinos dos dados. Este trabalho propõe uma abordagem para definir e implementar o percurso completo dos dados ao longo do tempo. Utilizou-se metadados padronizados pelo Dublin Core, enriquecidos para capturar e persistir o histórico de transformações. A abordagem inclui a modelagem de metadados específicos para rastrear operações de ETL (Extração, Transformação e Carga) e demonstrou ser viável em um projeto com dados reais.*

1. Introdução

Em um cenário de crescente complexidade na integração de dados provenientes de múltiplas fontes como bancos de dados (BDs) transacionais e APIs heterogêneas, a *Governança de Dados* emerge como pilar essencial para assegurar a consistência, a conformidade e a confiabilidade das informações. Projetos que envolvem a harmonização de dados heterogêneos exigem não apenas processos robustos de ingestão e transformação, mas também rastreabilidade precisa em todo o ciclo de vida dos dados [International 2017, Inmon 1992, Batini et al. 2016].

Nesse contexto, a *Linhagem de Dados* assume papel estratégico, permitindo auditorias transparentes e documentação detalhada de origens, transformações e fluxos [Simmhan et al. 2005]. A combinação de metadados enriquecidos, alinhados a padrões

como o *Dublin Core (DC)* [DCMI 2024], amplia essa capacidade, oferecendo descrições padronizadas que facilitam a interoperabilidade e a fiscalização da qualidade.

Assim, em ambientes com múltiplas fontes de dados, a governança aliada à linhagem não apenas mitiga riscos de inconsistências e violações regulatórias, mas também transforma dados brutos em ativos confiáveis, prontos para impulsionar decisões estratégicas e inovações orientadas por evidências. Além disso, em projetos de integração de diversas fontes de dados, uma abordagem para linhagem de dados também assegura conformidade com regulamentações, como a Lei Geral de Proteção de Dados pessoais (LGPD) [Brasil 2018].

Diante disso, este artigo propõe uma abordagem para garantir qualidade e linhagem de dados com enfoque no domínio de dados governamentais, e uma prova de conceito no escopo do projeto *Projeto Céos*[1], uma iniciativa de pesquisa e desenvolvimento executada pela *Universidade Federal de Santa Catarina (UFSC)* com a colaboração e financiamento do *Ministério Público de Santa Catarina (MPSC)*. Este projeto inclui o estudo, projeto e implementação de fluxos de trabalho (*workflows*) para coleta, integração, organização, processamento e análise de dados voltados à extração de conhecimento de forma automatizada ou semiautomatizada, visando apoiar a tomada de decisão em processos do *MPSC*.

As principais contribuições deste trabalho são: *(i)* a proposta de uma abordagem materializada na forma de um *framework* para gestão da linhagem de dados, que inclui uma extensão do padrão DC com diversos elementos de metadados relevantes no contexto de linhagem de dados, uma modelagem conceitual dos metadados necessários para esta gestão, e uma arquitetura que indica os processos a serem realizados por este arcabouço; *(ii)* uma prova de conceito aplicada a um projeto de pesquisa que lida com dados governamentais reais; *(iii)* uma abordagem que serve como um guia para futuras soluções voltadas à gestão de linhagem de dados.

Este artigo apresenta outras seis seções. A Seção 2 traz uma fundamentação sobre Linhagem de Dados. A Seção 3 apresenta os trabalhos relacionados, a Seção 4 detalha a abordagem proposta e a Seção 5 apresenta uma aplicação da abordagem em um projeto de pesquisa com dados reais. Por fim, a Seção 6 é destinada à conclusão.

2. Linhagem de Dados

Linhagem de dados pode ser entendida como o processo que captura o histórico de derivação de um conjunto de dados, incluindo suas fontes, processos intermediários e dependências, sendo essencial para auditoria, reprodutibilidade e governança [Simmhan et al. 2005]. Em termos de literatura, a obra de [Kimball and Ross 2013], apresenta bases metodológicas precursoras para a linhagem de dados. O autor destaca a necessidade de rastrear a origem e o fluxo de informações em ambientes de data warehousing, enfatizando processos de *ETL (Extract, Transform and Load)* como mecanismos críticos para movimentação e transformação estruturada. Embora o termo *"linhagem"* não seja explicitamente adotado, sua abordagem para mapeamento de fontes e dependências estabeleceu um arcabouço conceitual para demandas modernas de governança e transparência.

[1] *https://https://ceos.ufsc.br/*

Ainda, [Inmon 1992] destaca a necessidade de documentação rigorosa da origem e do contexto dos dados em ambientes corporativos, antecipando princípios de rastreabilidade. Na década seguinte, [Olson 2003] formalizou a linhagem de dados como componente estruturante da governança de dados, vinculando-a diretamente à qualidade e à confiabilidade das informações. Essa evolução reflete a transição de uma abordagem operacional (centrada em *data warehousing)* para uma visão estratégica, alinhada a demandas regulatórias e de integridade analítica.

A institucionalização da linhagem de dados gerou *frameworks* como o DAMA-DMBOK [International 2017], que a posiciona como elemento crítico da gestão de metadados e governança. Paralelamente, padrões técnicos como o ISO/IEC 11179 [ISO/IEC 2003] estabeleceram diretrizes para metadados interoperáveis. Essas iniciativas não apenas validaram a linhagem como requisito transversal a diferentes domínios, como finanças e saúde, mas também reforçaram sua função na mitigação de riscos e na otimização de fluxos analíticos em ecossistemas multicamadas.

Neste contexto, a LGPD reforça a importância da linhagem de dados ao vincular princípios como *accountability* ou responsabilização (Art. 6º, VIII) e transparência (Art. 9º) à obrigatoriedade de rastreabilidade. O Art. 37, por exemplo, exige registros detalhados das operações de tratamento, incluindo origem, finalidade e compartilhamento de dados, elementos que só podem ser documentados integralmente por meio de métodos para controle de linhagem de dados. Ainda, o princípio da prevenção (Art. 6º, VI) obriga organizações a adotarem medidas proativas, como a monitoração contínua da linhagem, para evitar violações ou uso inadequado. Esses requisitos evidenciam que, mesmo sem citar explicitamente o termo, a LGPD institucionaliza a linhagem de dados como ferramenta indispensável para conformidade, auditorias e manutenção da confiança dos titulares em cenários de alta complexidade operacional.

Um processo típico de gestão de linhagem de dados possui quatro componentes essenciais: *(i) origem* (fontes primárias como BDs, APIs ou sensores); *(ii) transformações* (processos de limpeza, enriquecimento ou agregação); *(iii) fluxo* (movimentação entre sistemas); e *(iv) consumo* (uso em relatórios, modelos analíticos ou processos decisórios) [International 2017]. Através de metadados estruturados pelo *DC*, como *dc:source (origem)*, *dc:description (processamento)* e *dc:relation (fluxo)*, este trabalho aplica essas etapas, garantindo rastreabilidade e interoperabilidade.

3. Trabalhos Relacionados

Esta seção apresenta dois grupos de trabalhos relacionados: *(i)* estudos gerais que enfatizam a urgência de métodos para governança de dados focados em linhagem e qualidade; e *(ii)* estudos que aplicam modelos ou *frameworks* para mitigar lacunas relacionadas à linhagem de dados. Eles são detalhados a seguir.

3.1. Estudos Gerais

Em uma revisão sistemática da literatura recente foram identificados 66 estudos relevantes publicados entre 2016 e 2022 [Gierend et al. 2024]. Destes, apenas 19 trabalhos concentram-se em desafios relacionados à anotação, metadados e modelagem de proveniência, evidenciando lacunas na padronização de descrições contextuais. Além disso, destaca que apenas 7% das ferramentas de gestão de dados oferecem módulos integrados

para captura, representação e visualização de proveniência, e também reforça a necessidade de abordagens mais robustas para rastreabilidade.

A pesquisa de [Mendoza et al. 2023] identifica desafios críticos que comprometem a usabilidade e confiabilidade dos dados públicos, como: *(i)* metadados pobres ou ausentes, que dificultam a interpretação contextual; *(ii)* falta de padronização em formatos, unidades e estruturas de dados [Gurstein 2011, Vetrò et al. 2016]; *(iii)* dados incompletos ou incongruentes, que limitam análises críticas; e *(iv)* dificuldades de acesso, como *links* quebrados ou portais pouco intuitivos. Além disso, a falta de rastreabilidade da fonte e a gestão inadequada do ciclo de vida dos dados [Vetrò et al. 2016] comprometem a transparência e a atualização dos conjuntos públicos.

[Faria et al. 2018] apresenta uma análise comparativa de *frameworks* de governança de dados e definição de uma estrutura hibrida que atenda suas especificidades, o *GovDadosMB*. O estudo revela que a qualidade na governança e na linhagem de dados depende da integração estruturada de componentes críticos.

Um questionamento realizado em [dos Santos et al. 2011] revela lacunas críticas em processos de ETL que comprometem a qualidade de dados em ambientes analíticos. Segundo a pesquisa, 28% dos respondentes não tratam conflitos de domínio ou conversões de dados durante o ETL, enquanto 41% enfrentam dificuldades para detectar atributos inconsistentes. Além disso, 73% dos participantes indicaram ausência de um repositório centralizado de metadados para documentar origens, transformações e destinos dos dados, resultando em riscos elevados de ambiguidade semântica e perda de rastreabilidade.

[Barata and Prado 2015] analisa a implementação de processos de governança em dois setores estratégicos: automotivo (Empresa A) e portuário-logístico (Empresa B). O trabalho relata que, enquanto políticas de dados e qualidade são priorizadas (65%-78% de adoção na Empresa A; 69%-75% na B), processos como gerenciamento de metadados (0% em ambas) e *compliance* (65% na A; 0% na B) são negligenciados. A disparidade é evidente em áreas como *data warehousing* (15% na A; 0% na B) e documentação (0% na A; 50% na B), revelando uma abordagem fragmentada. O estudo atribui essas lacunas a fatores culturais, como falta de priorização estratégica, e estruturais, como carência de frameworks adaptáveis, destacando que mesmo setores com alta criticidade operacional, como o portuário, subestimam a rastreabilidade e a padronização semântica.

3.2. Estudos Aplicados

O estudo de [Reis Jr et al. 2019] propõe um *framework* para rastrear a criação e publicação de dados públicos, destacando a necessidade de registrar quando, como e por que os dados são gerados. Essa abordagem, validada em um caso de uso real, alinha-se à definição clássica de proveniência de [Buneman et al. 2001], que enfatiza a documentação de origens e processos de transformação.

Complementarmente, [da Silva et al. 2016] apresenta uma arquitetura baseada no modelo PROV-DM[2] e *DC* para monitoramento ambiental, integrando um repositório de grafos RDF e uma API para coleta e consulta de metadados. O estudo demonstra que a padronização semântica (via PROV-DM) e a interoperabilidade (via *DC*) permitem não apenas a reconstituição histórica de dados, mas também a reutilização crítica de processos

[2] *https://www.w3.org/TR/prov-dm/*

e ativos em ecossistemas complexos.

Outro estudo propõe um *framework* para monitorar a proveniência de dados e detectar erros de processamento em fluxos de ETL [Johns et al. 2025]. A solução combina técnicas de rastreabilidade baseadas no PROV-DM com algoritmos de validação em tempo real, identificando inconsistências como domínios inválidos, conversões mal formadas ou violações de integridade semântica. Em uma avaliação experimental, o *framework* demonstrou redução de 40% em erros críticos durante a ingestão de dados, além de melhorar a documentação automática de metadados por meio de *logs* estruturados.

Embora esses estudos prévios, como o de [da Silva et al. 2016], tenham avançado na modelagem de grafos com PROV-DM e DC, e [Johns et al. 2025] e [Reis Jr et al. 2019] abordem métricas de qualidade em ETL e mapeamento de entidades em grafos, respectivamente, eles apresentam lacunas em termos de granularidade de metadados e integração, que são contextualizadas em [da Silva et al. 2024]. Este trabalho supera tais limitações ao propor: *(i)* uma extensão do DC com 15 elementos personalizados que capturam metadados de qualidade e regras de transformação em ETL; *(ii)* uma abordagem centrada em um *framework* que combina ferramentas maduras para gestão de metadados com repositórios próprios de metadados, garantindo rastreabilidade *end-to-end* (origem, transformações, destino); e *(iii)* métricas de qualidade integradas que associam indicadores como completude e consistência a *logs* estruturados. A abordagem proposta é detalhada a seguir.

4. Abordagem Proposta

A abordagem para gestão de linhagem de dados aqui proposta foi desenvolvida para resolver lacunas críticas identificadas em ambientes de dados heterogêneos, como metadados inconsistentes, rastreabilidade incompleta e falhas de qualidade em processos de ETL. A proposta inicialmente enriquece o padrão *DC* com novas propriedades (elementos) relevantes no contexto da linhagem de dados. Estes elementos são responsáveis pela manutenção de metadados referentes à *descrição, origem, transformações* e *qualidade* dos dados. Este enriquecimento é resultado de um trabalho anterior dos autores deste artigo [da Silva et al. 2024].

Na sequência, uma modelagem conceitual foi elaborada para a representação unificada de metadados considerados relevantes no contexto de linhagem de dados, integrando métricas de qualidade, como *completude* e *confiabilidade*, na forma de atributos auditáveis. Esta modelagem é apresentada na Figura 1 e considera que os metadados referentes à linhagem de dados estão persistidos em um BD relacional, o que justifica as entidades *Tabela* e *Colunas*.

A modelagem conceitual considera metadados em dois níveis de granularidade: *Conjunto de Dados* (granularidade geral) e *Tabela* (granularidade específica). Ela também estrutura a linhagem de dados ao incorporar a entidade *Transformação* e seus relacionamentos, e permite rastrear como os dados foram transformados ao longo do tempo, quem realizou essas transformações, data e qual código foi utilizado em cada processo. Com este histórico de alterações realizadas garante-se a transparência, confiabilidade e reprodutibilidade das análises realizadas sobre os dados.

Ainda, a modelagem conceitual representa o enriquecimento do padrão *DC* mencionado anteriormente. Esta representação considera atributos relacionados à qualidade

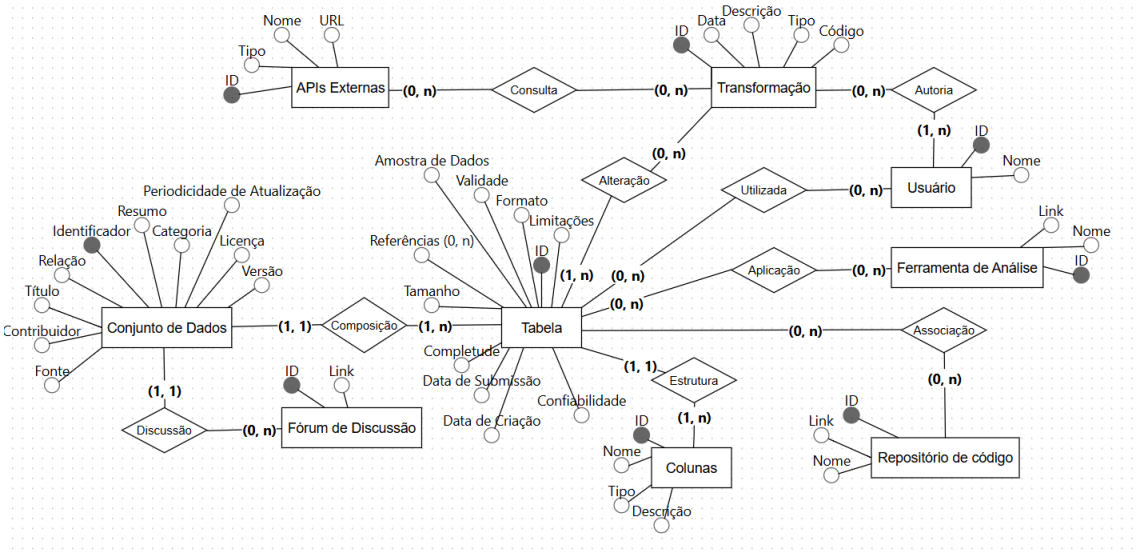

Figura 1. Modelagem conceitual proposta

dos dados, como *completude, validade* e *confiabilidade*, ao contexto de uso, incluindo *licença, limitações* e *usuários envolvidos* e à curadoria técnica com destaque para *ferramentas de análise, fóruns de discussão* e *repositório de código*. Dessa forma, os metadados descrevem não apenas informações sobre os recursos, mas também sua estrutura técnica, condições de utilização e histórico de modificações, tornando-se mais eficazes para diferentes perfis de usuários e finalidades analíticas.

Nesta abordagem é proposta também a arquitetura de um *framework* de alto nível para gestão da linhagem de dados, conforme ilustra a Figura 2. Esta arquitetura abrange coleta, processamento e armazenamento de metadados. Um dos seus componentes principais é o *Message Broker*, que é um sistema de mensageria responsável por processos de ETL de fontes de dados heterogêneas. Seu objetivo é permitir o projeto e a implementação de *workflows* (fluxos) distribuídos com comunicação assíncrona baseada em filas de mensagens. A intenção é não apenas o suporte a processos customizados de ETL, mas também a garantia da rastreabilidade dos dados ao longo destes processos através do desenvolvimento de algoritmos especializados para mapear origens e dependências dos dados. Estes algoritmos realizam a extração dos metadados definidos na modelagem conceitual e permitem a reconstituição histórica detalhada da linhagem dos dados. Conforme mostra a Figura 2, a abordagem proposta sugere o *RabbitMQ*[3] como sistema de mensageria devido à sua maturidade e confiabilidade. Ele adota o modelo *Publish/Subscribe*[4], que organiza mensagens em filas específicas.

Na sequência, os dados e metadados processados pelo *Message Broker* são encaminhados ao componente *Data Transformer* que consome, processa e carrega dados em repositórios específicos. Um outro componente específico, denominado *MExtractor*, extrai metadados importantes e procede a ingestão dos mesmos no BD PostgreSQL e na *OpenMetadata (OMD)*[5], uma plataforma aberta que centraliza metadados para gover-

[3] *https://www.rabbitmq.com/*
[4] *https://cloud.google.com/pubsub/docs/pubsub-basics*
[5] *https://open-metadata.org/*

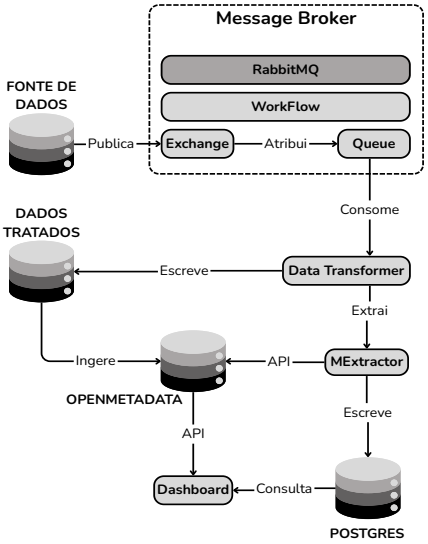

Figura 2. Arquitetura de um framework para gestão da linhagem de dados

nança, descoberta e observabilidade de dados. A escolha pelo OMD foi o resultado de uma análise de soluções para gestão de metadados. Esta análise é fruto de um artigo anterior dos autores deste trabalho [da Silva et al. 2025].

Por fim, uma interface visual (*dashboard*) permite consultas sobre os dados operacionais extraídos dos processos ETL, assim como a obtenção de informações sobre a linhagem dos dados obtidas pelo MExtractor. No que tange à linhagem de dados, espera-se que o *dashboard* viabilize a auditoria contínua das transformações ocorridas nos dados de fontes brutas, além da geração de relatórios de conformidade para atender, por exemplo, requisitos da LGPD, e monitoramento em tempo real.

5. Aplicação da Abordagem no Domínio de Dados Governamentais

Esta seção descreve um estudo de caso que aplica a abordagem no domínio de dados governamentais no âmbito do projeto *Projeto Céos*. A Figura 3 ilustra como o *framework* proposto pela abordagem foi instanciado no contexto do projeto. O fluxo exibido na parte superior representa o componente *Message Broker* implementado no *RabbitMQ*, enquanto que a parte inferior, centrada no componente *MExtractor*, representa as demais partes do *framework*.

O fluxo implementado no *RabbitMQ* lida com diversos repositórios e processos. O processo *Collector* é responsável pela coleta dos dados de fontes externas e potencialmente heterogêneas, mantendo esses dados brutos persistidos em um formato de entrada (por exemplo, JSON), enquanto o processo *Verifier* verifica a integridade dessas fontes. O processo *Splitter* divide os dados em partições para que o processo *Processor* possa realizar o processamento e transformação dos mesmos. Por fim, o processo *Quality Checker* realiza a verificação da qualidade desses dados e os persiste no repositório *Licitação Nova ou Atualizada* que mantém dados novos coletados e transformados, ou dados já existentes que foram atualizados com novas informações advindas de uma outra fonte.

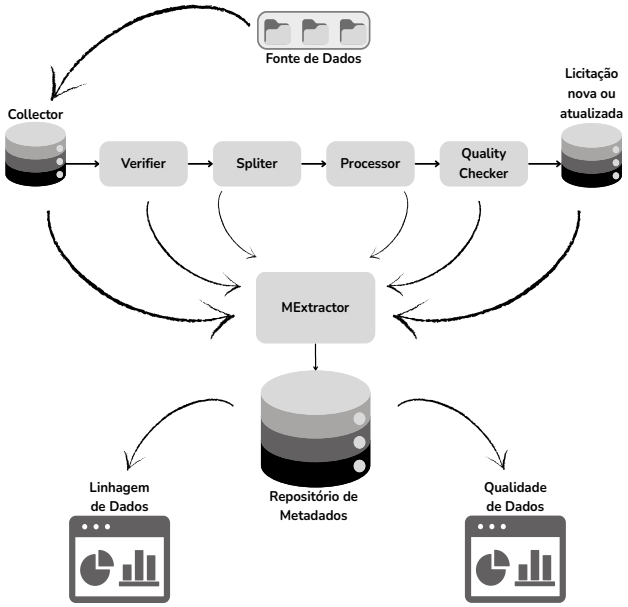

Figura 3. Instanciação do *framework* proposto no *Projeto Céos*.

Como pode ser visto na Figura 3, o *MExtractor* se comunica com cada um dos processos citados anteriormente, realizando a extração de metadados de cada um deles. Por fim, ele procede o armazenamento dos metadados extraídos tanto no OMD quanto no BD PostgreSQL, como ilustrado na Figura 2. Representa-se ambos os repositórios como um *Repositório de Metadados* que é utilizado como fonte de dados para a elaboração de *dashboards* que apresentam informações como linhagem e qualidade dos dados.

Cabe ressaltar que todos os dados de entrada coletados pelo *Collector* são mantidos no formato JSON bruto. Em seguida, o *MExtractor* intercepta eventos em três fases do fluxo (ingestão, transformação e verificação) para extrair metadados críticos. Para ilustrar esta extração, a Tabela 1 apresenta três exemplos de metadados catalogados nestas fases.

Cada metadado exemplificado apresenta um propósito distinto. A *Data de Publicação* indica quando o dado foi concebido. Os *Tratamentos Aplicados* documentam as transformações aplicadas sobre o dado, ou seja, por quais processos e alterações o dado passou desde que entrou no fluxo. Já a *Amostra de Dados* fornece um fragmento representativo do resultado final, usualmente seguindo um padrão pré-estabelecido para ser inserido no BD PostgreSQL, fragmento esse que pode ser apresentado ao usuário em *dashboards* para uma melhor compreensão dos dados coletados.

Uma especificação completa de todos os metadados definidos, bem como os respectivos pontos de captura destes metadados ao longo do fluxo de linhagem de dados encontra-se disponível no Github[6].

[6]*https://gist.github.com/medeirosJose/c3831d0901a9ab173d9f8b98194e2faf.*

Tabela 1. Exemplos de metadados extraídos em fases específicas do fluxo

Fase / Componente	Metadado	Mecanismo de Captura
Ingestão / Collector	Data de Publicação	Parsing do campo `date` no JSON recebido via `ColetorDom.process_message`
Transformação / Processor	Tratamentos Aplicados	`ProcessorDom` + `MExtractor`: interceptação de *logs* de limpeza e enriquecimento (remoção de duplicatas, preenchimento de vazios, normalização de tipos)
Verificação / Quality Checker	Amostra de Dados	`MExtractor`: extração e persistência das primeiras N entradas do JSON final gerado por `QualityChecker`, para confirmação rápida do resultado

Tabela 2. Exemplo de metadado no nível de conjunto de dados

Metadado	Estágio / Componente	Mecanismo de Captura
Título	`ColetorDom.process_message`	Extração do campo `title` do JSON de resposta da API.

5.1. Diretrizes para Captação de Metadados

Para a implementação do serviço de mensageria no *RabbitMQ*, foi definido um conjunto de diretrizes para captação de metadados em três níveis de granularidade: *conjunto de dados, recurso/tabela* e *metadados transversais ao fluxo*.

O nível de *conjunto de dados* se refere a metadados de âmbito global, isto é, atributos que descrevem o conjunto de dados em sua totalidade, sendo cruciais para rastreabilidade e governança. Um exemplo é mostrado na Tabela 2 para um registro do metadado *Título*. Outros elementos de metadados (*Versão, Direitos, Fonte, Identificador*, etc.) seguem o mesmo padrão e estão detalhados no Github[7].

No nível de *recurso/tabela*, extrai-se atributos relativos a cada arquivo ZIP ou JSON coletado de uma fonte bruta. A Tabela 3 mostra o exemplo do metadado *Formato*. A lista completa, incluindo metadados como *Tamanho, Referências, Colunas* e *Amostra*, está disponível no Github[8].

[7] *https://gist.github.com/medeirosJose/c3831d0901a9ab173d9f8b98194e2faf* .
[8] *https://gist.github.com/medeirosJose/c3831d0901a9ab173d9f8b98194e2faf* .

Tabela 3. Exemplo de metadado no nível de recurso/tabela

Metadado	Estágio/Componente	Mecanismo de Captura
Formato	`filtrar_registros_para_data`	Extrair o formato no objeto JSON do recurso

Tabela 4. Exemplo de metadado no nível transversal ao fluxo

Metadado	Estágio / Componente	Mecanismo de Captura
Ferramentas de Análise	Após `load_extractors()` e `load_verifiers()`	Listagem de bibliotecas utilizadas no fluxo (por exemplo, `pandas` e `numpy`).

Por fim, os *metadados transversais ao fluxo* estão presentes em todo o fluxo de linhagem de dados. A Tabela 4 apresenta o metadado *Ferramentas de Análise*. Os demais metadados, como *Confiabilidade, Repositório de Código* e *Service-Account*, podem ser consultados com exemplos de uso no Github[9].

Cabe ressaltar que diversos metadados são sistematicamente capturados e reutilizados ao longo de cada etapa do fluxo do *Message Broker* com o objetivo de fortalecer a linhagem e o monitoramento dos dados. Alguns metadados, como aqueles relacionados às datas e ferramentas utilizadas são repetidos em todas as etapas.

5.2. Integração com OpenMetadata e Proposta de *Dashboard*

Após a conclusão do fluxo de ETL e a carga no BD PostgreSQL, o *framework* instanciado ativa um conector REST nativo do OMD. Essa integração importa automaticamente esquemas e perfis de colunas, executa um *profiler* interno para gerar métricas de qualidade, como *completude, validade* e *anomalias*, e registra a linhagem de cada atributo. O OMD então produz um catálogo enriquecido de metadados disponível para consulta, como exemplificado na Figura 4.

Para suportar auditorias e monitoramento, foi proposto e desenvolvido um *dashboard* que exibe, de forma integrada, tanto a linhagem completa dos dados (dado bruto, transformações e dado final) quanto os principais indicadores de qualidade, conforme ilustrado na Figura 5.

A interface gráfica do *dashboard* divide-se em três áreas principais. Na área de *Linhagem de Dados*, cada processo do fluxo é representada por um *card*. As cores e ícones reforçam a distinção semântica de cada processo. A área *Metadata Repository* exibe o repositório central que consolida todos os metadados extraídos, funcionando como ponto único de consulta histórica.

Por fim, a área lateral *Metadados* está organizada em quatro abas distintas, cada uma enfatizando um nível de granularidade (ver Seção 5.1). Esta área detalha os metadados gerados em cada processo selecionado pelo usuário na área de *Linhagem de Dados*.

[9] *https://gist.github.com/medeirosJose/c3831d0901a9ab173d9f8b98194e2faf* .

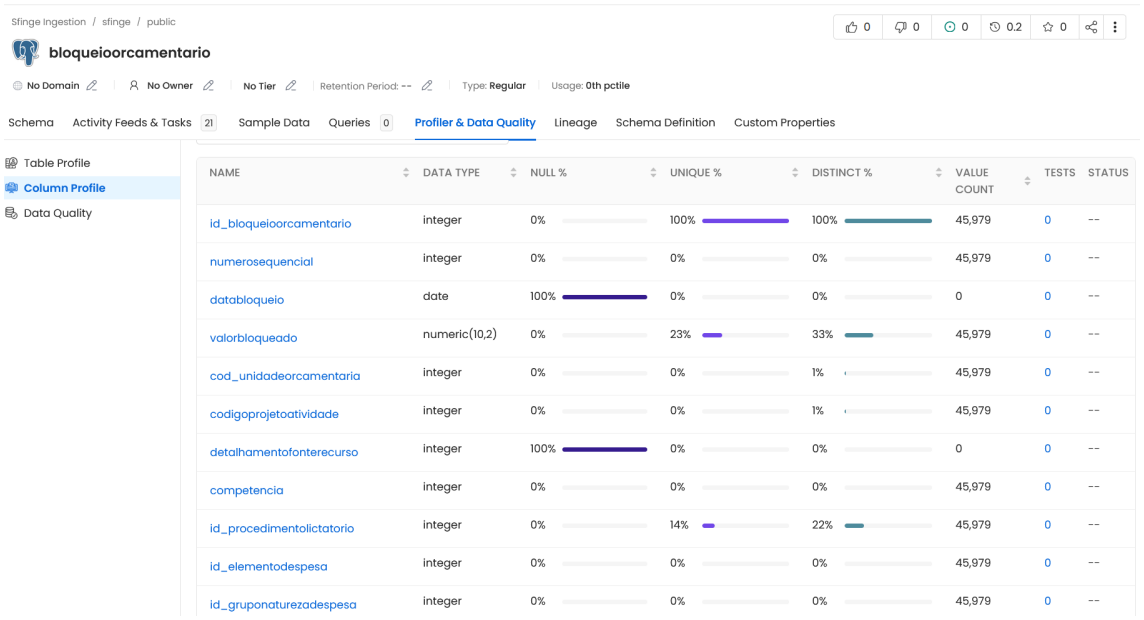

Figura 4. Informações sobre qualidade de Dados no OMD

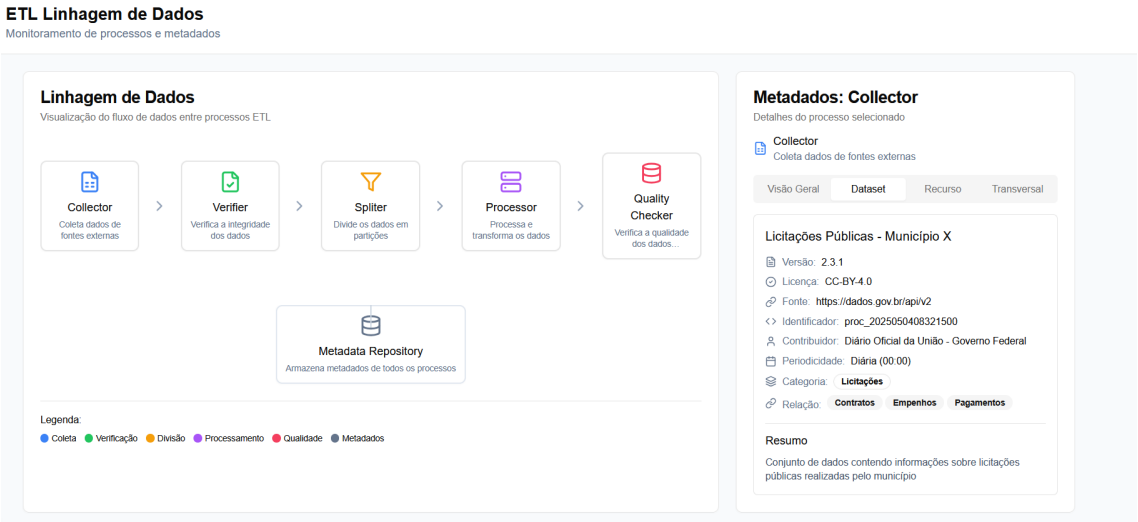

Figura 5. *Dashboard* de linhagem de dados e metadados - dados anonimizados

Na aba *Visão Geral*, o usuário encontra o nome do módulo (por exemplo, *Collector*) acompanhado de uma breve descrição de sua função no fluxo, além de um resumo sintético das principais métricas, como *tamanho* e *quantidade* de arquivos processados. A aba *Dataset* agrupa os metadados de alto nível, como *título do conjunto, versão, licença* e *organização proprietária*, oferecendo uma visão completa do contexto da fonte de dados.

Na aba *Recurso* são exibidos os atributos específicos do arquivo ou tabela processada (*formato do dado, tamanho do arquivo, lista de colunas* e *data de submissão*), permitindo avaliar imediatamente a granularidade e a composição da carga de dados. Por fim, a aba *Transversal* reúne informações sobre as *bibliotecas* e *ferramentas de análise* utilizadas (por exemplo, *pandas*, *numpy* e *scikit-learn*), a *métrica de confiabilidade do serviço* (percentual de execuções bem sucedidas), *links* diretos para o repositório de código e *referências* a outras fontes externas, como portais oficiais de publicação.

6. Conclusão

Este trabalho propõe uma abordagem para gestão da linhagem de dados com foco em processos ETL que devem gerar dados mais adequados ao consumo por pessoas e aplicações específicas, como aquelas que realizam análises sobre esses dados. A abordagem visa garantir a rastreabilidade e a governança eficiente de dados, sugerindo o uso de padrões robustos como *DC* e ferramentas consolidadas como *RabbitMQ* e *OMD* materializados como um *framework* instanciável da abordagem em diversos domínios de aplicação.

O *framework* sugerido pela abordagem foi instanciado no contexto de dados governamentais no âmbito do *Projeto Céos*. Ele foi considerado bastante útil pelos especialistas do domínio para um conjunto preliminar de fontes cujos dados foram coletados. Segundo relato de alguns especialistas, havia uma demanda para conhecer, por exemplo, a origem de um processo licitatório pl_x, ou seja, qual a sua fonte, pois isso auxilia na investigação de eventuais fraudes no caso desta mesma fonte já ter processos com caráter fraudulento, aumentando a probabilidade de pl_x também apresentar alguma irregularidade. Por fim, outros fluxos de ETL estão sendo desenvolvidos para outras fontes de dados relevantes para o projeto.

A fim de sanar limitações deste trabalho, pretende-se desenvolver como trabalhos futuros: considerar a linhagem de dados em níveis de granularidade mais específicos, como colunas e tuplas. Hoje trabalha-se em niveis mais gerais, como conjunto de dados e tabelas. Além disso, propõe-se desenvolver um mecanismo automatizado para registrar e rastrear versões específicas de *scripts* de transformação utilizados nos fluxos, permitindo a identificação precisa da versão do código pela qual cada conjunto de dados passou. Essa funcionalidade visa reforçar a auditabilidade do processo como um todo, possibilitando uma reconstituição detalhada e confiável das alterações aplicadas. Ainda, uma análise de desempenho de acesso é pretendida, visando otimizar buscas por informações de linhagem de dados.

7. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

[Barata and Prado 2015] Barata, A. and Prado, E. (2015). Governança de dados em organizaçoes brasileiras. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 267–274. SBC.

[Batini et al. 2016] Batini, C., Scannapieco, M., et al. (2016). Data and information quality. *Cham, Switzerland: Springer International Publishing*, 63.

[Brasil 2018] Brasil (2018). Lei geral de proteção de dados pessoais (lgpd). Lei nº 13.709, de 14 de agosto de 2018. Diário Oficial da União, Brasília, 14 ago. 2018.

[Buneman et al. 2001] Buneman, P., Khanna, S., and Tan, W.-C. (2001). Why and where: A characterization of data provenance. *Lecture Notes in Computer Science*, 2237:316–330.

[da Silva et al. 2016] da Silva, D. L., Batista, A., and Correa, P. L. (2016). Data provenance in environmental monitoring. In *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pages 337–342. IEEE.

[da Silva et al. 2024] da Silva, H. A. B., Santos, J. V. d., Jochem, J. E. M., Fleck, A., Mello, R. d. S., Dorneles, C. F., and Fileto, R. (2024). Uma Proposta Baseada no Dublin Core para Catalogação de Metadados de Fontes de Dados Governamentais. In *XXXIX Simpósio Brasileiro de Banco de Dados (SBBD)*. Sociedade Brasileira de Computação.

[da Silva et al. 2025] da Silva, H. A. B., Santos, J. V. d., Souza, E. F. R. d., Jochem, J. E. M., Mello, R. d. S., Dorneles, C. F., and Fileto, R. (2025). Análise de Ferramentas de Código Aberto para Gestão de Metadados: OpenMetadata e Amundsen. In *XX Escola Regional de Banco de Dados (ERBD)*. Sociedade Brasileira de Computação.

[DCMI 2024] DCMI (2024). Dcmi metadata terms. *https://www.dublincore.org/specifications/dublin-core/dcmi-terms/section-1*. Último acesso: 12 de março de 2025.

[dos Santos et al. 2011] dos Santos, I. M. F., Andreatta, A. A., and Siqueira, S. W. (2011). Qualidade dos dados nas organizações sob o enfoque de apoio a decisão: Um estudo exploratório. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 166–177. SBC.

[Faria et al. 2018] Faria, M. R., Lopes, M., de Faria Cordeiro, K., et al. (2018). Govdadosmb: um framework de governança de dados corporativos para a marinha do brasil. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 241–246. SBC.

[Gierend et al. 2024] Gierend, K., Krüger, F., Genehr, S., Hartmann, F., Siegel, F., Waltemath, D., Ganslandt, T., and Zeleke, A. A. (2024). Provenance information for biomedical data and workflows: Scoping review. *Journal of medical Internet research*, 26:e51297.

[Gurstein 2011] Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*. Disponível em: <https://firstmonday.org/ojs/index.php/fm/article/view/3316/2764>. Acesso em: 30 mar. 2025.

[Inmon 1992] Inmon, W. H. (1992). *Building the Data Warehouse*. Wiley.

[International 2017] International, D. (2017). *DAMA-DMBOK: Data Management Body of Knowledge*. Technics Publications, Nova Jersey, 2. ed. edition.

[ISO/IEC 2003] ISO/IEC (2003). Iso/iec 11179: Information technology — metadata registries (mdr). International Organization for Standardization.

[Johns et al. 2025] Johns, M., Baum, L., and Prasser, F. (2025). Tracking provenance in clinical data warehouses for quality management. *International Journal of Medical Informatics*, 193:105690.

[Kimball and Ross 2013] Kimball, R. and Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling*. John Wiley & Sons.

[Mendoza et al. 2023] Mendoza, I., Corrêa, R., and Bernardini, F. (2023). Como a governança de dados pode auxiliar na mitigação de barreiras de uso de portais de dados governamentais abertos? uma análise da literatura. In *Workshop de Computação Aplicada em Governo Eletrônico (WCGE)*, pages 212–223. SBC.

[Olson 2003] Olson, J. E. (2003). *Data quality: the accuracy dimension*. Elsevier.

[Reis Jr et al. 2019] Reis Jr, C., Martins, L., Victorino, M., and Holanda, M. (2019). Modelo de dados de proveniência para uma arquitetura de dados abertos governamentais. In *Workshop de Transparência em Sistemas (WTranS)*, pages 11–20. SBC.

[Simmhan et al. 2005] Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *ACM Sigmod Record*, 34(3):31–36.

[Vetrò et al. 2016] Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., and Morando, F. (2016). Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly*, 33(2):325–337.