A armadilha da validade artificial: auditoria da perda semântica no crosswalk de metadados

Karolayne C. R. de Lima[1], **Marcos S. Sunye**[1]

[1]Departamento de Informática – Universidade Federal do Paraná (UFPR)
CEP: 81.530-090 – Curitiba – PR – Brasil

{kcrlima,sunye}@inf.ufpr.br

Abstract. *Open Science interoperability faces the challenge of informational degradation in metadata mappings. This work proposes a model based on structural constraints to quantify information loss (lossiness). The methodology employs the $\langle S, T, C, V \rangle$ tuple (Semantics, Type, Cardinality, and Value) in an auditing algorithm applied to 150 datasets from UFSCar, Embrapa, and GBIF repositories. Results indicate that mappings to flat standards (Dublin Core) lead to losses reaching up to 73.24% of the original richness, concentrated in Type and Semantic conflicts in structurally rich datasets. We conclude that syntactic validity masks reusability failures, making the proposed model an effective tool for data curators to audit semantic fidelity in data migrations. These findings demonstrate that structural interoperability cannot be presumed to preserve semantic usability.*

Resumo. *A interoperabilidade em Ciência Aberta enfrenta o desafio da degradação informacional em mapeamentos de metadados. Este trabalho propõe um framework para avaliação de interoperabilidade semântica em crosswalks, fundamentado na distinção formal entre validade estrutural e fidelidade informacional, operacionalizado pela tupla $\langle S, T, C, V \rangle$ e pela Métrica de Lossiness Ponderada (L_w). A metodologia aplica o algoritmo de auditoria a 150 conjuntos de dados dos repositórios UFSCar, Embrapa e GBIF. Os resultados indicam que mapeamentos para padrões planos (Dublin Core) resultam em perdas de até 73,24%, concentradas em conflitos de Tipo e Semântica. Conclui-se que a validade sintática mascara falhas de reusabilidade, sendo o modelo uma ferramenta eficaz para curadores auditarem a fidelidade semântica em migrações.*

1. Introdução

A interoperabilidade tornou-se um princípio estruturante das infraestruturas de ciência aberta. Repositórios digitais dependem de transformações entre padrões de metadados para possibilitar o intercâmbio em ambientes heterogêneos [Wilkinson et al. 2016, Borgman 2015]. Na prática, o sucesso dessas iniciativas costuma ser avaliado por critérios de conformidade estrutural: se o registro resultante satisfaz as regras de validação técnica, a transformação é considerada bem-sucedida.

Entretanto, validade estrutural não é sinônimo de preservação semântica. A literatura concentra-se em dimensões como completude, consistência e

[0]Os autores declaram que não utilizaram ferramentas de inteligência artificial generativa no processo de escrita deste manuscrito.

conformidade [Bruce and Hillmann 2004, Park 2009, Zeng and Chan 2006], com ênfase em compatibilidade sintática e alinhamentos formais [Haslhofer and Klas 2010, Rahm and Bernstein 2001]. Embora fundamentais para consistência técnica, essas abordagens pouco problematizam a retenção do contexto informacional durante a transformação. Como resultado, registros formalmente válidos podem apresentar comprometimento substancial de seu potencial de reuso científico.

O problema surge precisamente nos cruzamentos entre padrões quando *crosswalks* são tratados como simples tabelas de-para, ignorando diferenças estruturais e semânticas. Enquanto o campo `creator` no Dublin Core (DC) é preenchido como *string* livre, no Ecological Metadata Language (EML) é um objeto composto (nome, afiliação, papel). A conversão ingênua resulta em perda silenciosa de granularidade, a perda semântica [Doerr 2003]. Esse fenômeno configura *semantic drift*, equivalências estruturais que não correspondem a equivalências conceituais [Gradmann 2010].

As métricas consolidadas de qualidade [Bruce and Hillmann 2004, Park 2009, Zeng and Chan 2006] avaliam o registro como produto final, sem referência ao registro de origem. Similarmente, abordagens de *schema matching* [Rahm and Bernstein 2001, Haslhofer and Klas 2010] avaliam a qualidade do mapeamento como artefato, não a degradação informacional que sua aplicação produz em registros individuais. Essa lacuna é reconhecida pela EOSC [EOSC Association – FAIR Metrics and Data Quality Task Force 2022] e pelos *frameworks* FAIRness existentes, que operam sobre o estado estático do repositório, sem capturar o custo de transformações realizadas *upstream* [Candela et al. 2024]. A evidência empírica deste estudo ilustra essa condição. No cenário GBIF $\rightarrow$ Dublin Core, todos os 50 registros foram validados com sucesso e simultaneamente apresentaram L_w médio de 51,15%, fenômeno que denominamos Validade Artificial.

Para enfrentar essa lacuna, este estudo propõe um *framework* que formaliza a distinção entre validade estrutural e fidelidade informacional como propriedades independentes e mensuráveis. O modelo opera pela tupla $\langle S, T, C, V \rangle$ (Semântica, Tipo, Cardinalidade, Valor) associada à métrica de Perda Ponderada (L_w), que operacionaliza a degradação semântica como função da retenção diferencial de unidades informativas entre esquemas, ponderadas por sua utilidade para os Princípios FAIR [Wilkinson et al. 2016]. O estudo orienta-se pelas questões: **RQ1**: em que medida validade estrutural pode coexistir com perda semântica significativa? **RQ2**: como diferentes configurações de *crosswalk* impactam a preservação informacional? **RQ3**: é possível distinguir formalmente interoperabilidade estrutural de preservação semântica?

A validação apoia-se em 150 conjuntos de dados reais de UFSCar, Embrapa REDAE e GBIF, combinando padrões planos (Dublin Core), estruturados (DataCite) e hierárquicos ricos (Darwin Core/EML).

2. Fundamentação Teórica e Trabalhos Relacionados

Os Princípios FAIR [Wilkinson et al. 2016] posicionam a Interoperabilidade (I) e a Reusabilidade (R) como os pilares mais dependentes da qualidade dos mapeamentos. A interoperabilidade semântica constitui o nível mais complexo, exigindo que padrões representem e preservem o significado das informações [Ehrig 2007]. Euzenat e Shvaiko [Euzenat and Shvaiko 2013] distinguem correspondências de equivalência, subsunção e

disjunção entre conceitos. Já Obrst [Obrst 2003] situa a interoperabilidade em um espectro que vai do nível sintático ao pragmático, com a conformidade estrutural operando no nível sintático, enquanto preservação de fidelidade informacional exige o nível semântico.

A avaliação de qualidade de mapeamentos constitui subcampo consolidado. Rahm e Bernstein [Rahm and Bernstein 2001] sistematizaram abordagens de *schema matching* distinguindo técnicas baseadas em esquema, instância e linguagem. Lenzerini [Lenzerini 2002] formalizou integração de dados como satisfação de restrições sobre mapeamentos GAV/LAV, base formal sobre a qual o modelo proposto se apoia. Bernstein et al. [Bernstein and Melnik 2007] introduziram o conceito de *model management*, tratando mapeamentos como objetos de primeira classe sobre os quais operações algébricas se aplicam. O que essas abordagens não contemplam é a quantificação da degradação informacional que a aplicação de um mapeamento produz sobre registros individuais, a lacuna que este trabalho endereça.

Pipino et al. [Pipino et al. 2002] propuseram métricas subjetivas e objetivas para dimensões de qualidade de dados. Batini et al. [Batini et al. 2009] consolidaram *survey* das abordagens de melhoria, sem tratar a degradação introduzida por transformações como objeto autônomo. As métricas de Bruce e Hillmann [Bruce and Hillmann 2004], Park e Tosaka [Park 2009] e Zeng [Zeng and Chan 2006] operacionalizam completude, consistência e conformidade, mas tomam o registro resultante como unidade isolada, sem referência ao de origem. Essa assimetria é relevante, pois conformidade estrutural e retenção informacional podem divergir sistematicamente.

No contexto institucional, a EOSC reconheceu formalmente a heterogeneidade semântica entre repositórios como obstáculo central [EOSC Association – FAIR Metrics and Data Quality Task Force 2022]. O RDA FAIR Data Maturity Model [RDA FAIR Data Maturity Model Working Group 2020] e as métricas FAIRsFAIR [Devaraju et al. 2022] operam sobre o objeto de dados como um todo, sem mecanismos para quantificar degradação introduzida em *crosswalks*. Ferramentas como F-UJI realizam avaliações incapazes de diagnosticar perda por achatamento estrutural (*flattening*) de esquemas hierárquicos [Candela et al. 2024]. O projeto FAIRCORE4EOSC desenvolveu o MSCR para tornar mapeamentos objetos FAIR [EOSC Association – Semantic Interoperability Task Force 2023], mas não provê mecanismos para quantificar a perda de fidelidade ponderada após execução do mapeamento.

2.1. Distinções Conceituais Fundamentais

Validade *versus* Fidelidade Informacional. A Validade é propriedade binária, posto que um registro é válido quando satisfaz as restrições sintáticas do esquema de destino. Já a Fidelidade informacional é uma propriedade escalar que mede o quanto do conteúdo do registro de origem foi preservado no destino. Um registro pode ser simultaneamente válido e informacionalmente vazio, é o mecanismo central da Validade Artificial.

Interoperabilidade *versus* Reuso. A Interoperabilidade é propriedade da infraestrutura (capacidade técnica de trocar dados [Wilkinson et al. 2018]). O reuso é a propriedade do conteúdo (capacidade de ser utilizado além do contexto original [Wilkinson et al. 2016]). A ênfase exclusiva em interoperabilidade sintática não garante o objetivo de reuso que os Princípios FAIR visam promover.

Conformidade *versus* Preservação Informacional. A Conformidade é auditável por validadores automáticos (XSD, JSON Schema). A Preservação informacional requer comparação do estado antes e depois da transformação, o que exige $\langle S, T, C, V \rangle$ e L_w. A maioria das ferramentas existentes avalia a primeira, não a segunda [Candela et al. 2024, Devaraju et al. 2022].

Assim, o modelo proposto situa-se na interseção dessas literaturas, dado que complementa o ecossistema FAIRness ao introduzir uma camada de auditoria de transformação [Candela et al. 2024], responde à demanda da EOSC por qualidade orientada a máquinas [EOSC Association – FAIR Metrics and Data Quality Task Force 2022] e preenche a lacuna operacional do MSCR.

3. O Modelo de Mapeamento com Restrições

A tupla $\langle S, T, C, V \rangle$ proposta decompõe cada elemento de metadado em quatro dimensões críticas de restrição, derivadas da decomposição analítica das fontes de incompatibilidade identificadas na literatura [Lenzerini 2002, Rahm and Bernstein 2001, Batini et al. 2009]:

$$E = \langle S, T, C, V \rangle$$

Semântica (S): é a equivalência conceitual entre campos de diferentes esquemas. Verificada antes da execução algorítmica, durante a construção do Dicionário de Governança, por especialistas de domínio. Detecta o *semantic drift* que ocorre quando equivalências estruturais não correspondem a equivalências conceituais. Tipo (T): é a estrutura física da informação (dados simples *vs.* objetos compostos). Conflitos surgem ao encaixar hierarquias taxonômicas em campos de texto plano, destruindo organização interna e granularidade. Cardinalidade (C): são as regras de repetição e obrigatoriedade. Conflitos surgem ao transferir listas de valores para sistemas que aceitam apenas uma entrada. Valor (V): é o vocabulário permitido, diferenciando texto livre de vocabulários controlados. Quando o destino exige termos padronizados e a origem fornece texto livre, o registro torna-se semanticamente inválido.

A Tabela 1 resume os conflitos e suas consequências. O *crosswalk* passa a ser uma Função de Transformação $f(valor_{src}) \rightarrow valor_{tgt}$ sujeita a três invariantes: Compatibilidade de Tipo (T_{out} congruente com T_{tgt}), Resolução de Cardinalidade (regra de agregação quando $C_{src}.max > C_{tgt}.max$) e Validação de Domínio ($valor_{tgt} \in V_{tgt}$).

Tabela 1. Matriz de conflitos e impactos na interoperabilidade.

Dimensão	O que define	Exemplo de Conflito	Consequência Técnica
Semântica (S)	Conceito	Data coleta $\neq$ Data Pub.	Colisão semântica
Tipo (T)	Estrutura	Objeto $\rightarrow$ String plana	Achatamento (*Lossiness*)
Cardinalidade (C)	Quantidade	Lista $\rightarrow$ Valor único	Descarte de dados
Valores (V)	Domínio	Texto Livre $\rightarrow$ Vocab. controlado	Invalidez do registro

3.1. Métricas de Avaliação

Dois fenômenos diagnósticos emergem da aplicação conjunta das métricas de Validade (Val) e Perda Ponderada (L_w).

A Validade Artificial é o estado de um registro que satisfaz $Val = 1$ e simultaneamente apresenta $L_w > \theta_{trap}$ (adotado $\theta_{trap} = 0{,}30$). Ocorre por injeção de *defaults* (campos obrigatórios ausentes preenchidos com valores genéricos) ou por achatamento estrutural (objetos hierárquicos colapsados em strings planas).

A Anemia Semântica é o estado de baixa densidade informacional no esquema de origem, herdado da pobreza estrutural do padrão nativo, diagnosticado quando o esquema é plano e a proporção de campos com $w_i \geq 2{,}0$ é inferior a $\alpha_{anemia} = 0{,}40$. As métricas formais são:

1. **Validade (Val):**

$$Val_{struct} = \frac{\text{Registros em conformidade com } \langle T, C, V \rangle_{tgt}}{\text{Total de Registros}} \tag{1}$$

2. **Perda Ponderada (L_w):**

$$L_w = 1 - \frac{\sum(u_{tgt} \cdot w_i)}{\sum(u_{src} \cdot w_i)} \tag{2}$$

onde u_{src} e u_{tgt} são unidades informativas de origem e destino, e w_i é o peso por categoria: Identificadores (w=3,0), Proveniência (w=2,5), Taxonomia (w=2,0), Estrutural (w=1,5) e Descritivo (w=1,0). Esses pesos derivam da hierarquia funcional dos Princípios FAIR e do RDA FAIR Data Maturity Model [RDA FAIR Data Maturity Model Working Group 2020].

O Dicionário de Governança opera como mediador ontológico da componente Semântica (S), pois vincula campos técnicos de esquemas a conceitos funcionais, atribui pesos w_i e define restrições $\langle T, C, V \rangle$ por campo. A dimensão S é verificada pelo dicionário antes da execução; T, C e V são auditáveis algoritmicamente em tempo de execução. O fluxo de auditoria está ilustrado na Figura 1.

4. Metodologia

Este estudo é de natureza aplicada e abordagem mista. É qualitativo na modelagem conceitual das dimensões $\langle S, T, C, V \rangle$ e quantitativo na mensuração de Val e L_w sobre 150 registros. O algoritmo opera por simulação analítica de penalidades garantindo rastreabilidade, pois cada componente da perda é atribuível a uma dimensão específica.

4.1. Desenho Amostral e Coleta de Dados

O desenho adotou Amostragem Estratificada Intencional ($N = 150$, $n = 50$ por estrato), maximizando heterogeneidade estrutural e semântica. O tamanho amostral segue critérios de suficiência analítica [Batini et al. 2009] e saturação estrutural, a partir do perfilamento, os padrões dentro de cada repositório estabilizaram antes do quinquagésimo registro. Os três estratos representam arquétipos canônicos de expressividade (Tabela 2). A coleta foi realizada em Janeiro de 2026 via scripts Python (requests, xml.etree, json) sobre OAI-PMH e APIs REST, a partir do diretório Re3data. Os scripts de coleta e auditoria, o Dicionário de Governança e os 150 registros de metadados estão disponibilizados em repositório público[1] permitindo a reprodução integral dos experimentos.

[1] `http://dx.doi.org/10.5380/bdc/108`

Figura 1. Fluxo de auditoria.

O protocolo de inclusão exigiu serialização completa em JSON/XML, aderência a padrão reconhecido e presença de identificadores persistentes. Foram excluídos registros administrativos puros e links quebrados. Como limitação, o critério de curadoria mínima introduz viés de seleção positivo, podendo subestimar a degradação média em repositórios menos sistemáticos; os limiares diagnósticos absolutos requerem recalibração para populações distintas.

4.2. Execução do Algoritmo

O algoritmo opera em três etapas, cujo mapeamento para o fluxo da Figura 1 é: Etapa 1 corresponde à Caixa 1 (Mediação Semântica); Etapa 2 às Caixas 2, 3 e 4 (Validação Estrutural, Lossiness Semântico e Padrões); Etapa 3 à Caixa 5 (Critério de Anemia Semântica e Relatório de Integridade). **Etapa 1 (*Schema Profiling*):** iteração sobre os 150 JSONs para inferir tipo real, frequência de preenchimento e amostragem de valores, construindo o Esquema Efetivo de cada repositório. **Etapa 2 (Mediação Ontológica):** para cada par de esquemas, consulta ao Dicionário de Governança para equivalência semântica ponderada (S), auditoria de estrutura e valor ($\langle T, C, V \rangle$) e diagnóstico prescritivo de causa raiz. **Etapa 3 (Execução e Métricas):** cálculo de Val e L_w por registro, com identificação de *defaults* injetados e quantificação por dimensão da tupla.

Em termos de complexidade computacional, o algoritmo opera em $\mathcal{O}(n \cdot f \cdot d)$, onde n é o número de registros, f o número de campos por registro e d a profundidade

Tabela 2. Caracterização da amostra estratificada ($N = 150$, $n = 50$/repositório).

Repositório	Padrão Nativo	Características	Justificativa
RI UFSCar *Multidisciplinar*	Dublin Core (OAI-DC)	Estrutura plana, metadados bibliográficos generalistas.	**Baixa Complexidade**: testar Anemia Semântica.
Embrapa REDAE *Ciências Agrárias*	DataCite (Schema 4.x)	Estrutura híbrida com foco em DOI e vinculação autoral.	**Média Complexidade**: testar conflitos C e V.
GBIF *Biodiversidade Global*	Darwin Core / EML	Hierarquias complexas de taxonomia e cobertura espacial.	**Alta Complexidade**: testar resistência ao achatamento.

hierárquica máxima do esquema de origem. A etapa de maior custo é a Mediação Ontológica (Etapa 2), por envolver consulta ao Dicionário de Governança para cada campo. Em escala, essa etapa pode ser pré-computada por tipo de esquema, visto que o mapeamento S é invariante para campos do mesmo padrão. Para os 150 registros avaliados, o tempo total de execução mostrou-se compatível com uso interativo em hardware convencional, evidenciando viabilidade operacional para acervos de médio porte.

5. Resultados

5.1. Perfilamento dos Repositórios

A análise estrutural revelou diferenças marcantes na arquitetura da informação (Tabela 3).

A densidade bruta de preenchimento (u) não é sinônimo de riqueza científica. O repositório da UFSCar, com 100% de densidade em 12 campos Dublin Core, opera em Anemia Semântica latente. A estrutura plana (nível 1) e campos genéricos limitam a capacidade de carregar pesos w_i elevados. O GBIF, com menor densidade (70,3%) mas profundidade nível 5, concentra valor em objetos compostos como `taxonomicCoverages`, posto que qualquer achatamento destruirá vínculos semânticos. A Embrapa evidencia gargalos pontuais de cardinalidade (C) no aninhamento de autoria e proveniência.

5.2. Resultados dos *Crosswalks*

5.2.1. Cenário A: Dublin Core → DataCite (Enriquecimento)

O Cenário A apresenta menor taxa de perda ($L_w = 22{,}75\%$) com máxima estabilidade (Val=100%, $\sigma = 0$, Tabela 4). O L_w uniforme reflete a homogeneidade do Dublin Core, pois todos os 50 registros compartilham idêntica estrutura plana. A validade total exigiu injeção de *defaults* para campos obrigatórios do DataCite ausentes no DC original (`@dateType = 'Issued'`, `resourceTypeGeneral = 'Dataset'`), comprovando Validade Artificial. A perda distribui-se por `identificacao` (uso de Handle/URL em vez de DOI), `estrut_temp` (`dc:date` sem atributo `@dateType`) e `tax_especificidade` (ausente na origem).

Algoritmo 1: Auditoria de Crosswalk $\langle S, T, C, V \rangle$

Input: Registros M_{src}; $Schema_{tgt}$; Dicionário $\mathcal{D}$
Output: $R = (Val,\ L_w,\ D)$ por registro
// Etapa 1 - *Schema Profiling* (Caixa 1, Fig. 1)
foreach $r \in M_{src}$ **do**
  inferir tipo, frequência e valores de cada campo;
Construir $EsquemaEfetivo_{src}$;
// Etapa 2 - Mediação Ontológica (Caixas 2-4)
foreach *campo* $c \in EsquemaEfetivo_{src}$ **do**
  $S \leftarrow \mathcal{D}(c,\ Schema_{tgt})$; // equivalência conceitual
  auditar T, C, V contra as restrições de $Schema_{tgt}$;
// Etapa 3 - Execução e Métricas (Caixa 5)
foreach *registro* r **do**
  $Val(r) \leftarrow$ conformidade com $\langle T, C, V \rangle_{tgt}$;
  $L_w(r) \leftarrow 1 - \frac{\sum u_{tgt} \cdot w_i}{\sum u_{src} \cdot w_i}$;
  $D(r) \leftarrow$ taxa de *defaults* injetados;
return $(Val,\ L_w,\ D)$

Tabela 3. Indicadores de complexidade e risco de degradação informacional.

Indicador	UFSCar	Embrapa	GBIF
Densidade de Unidades (u)	100%	100%	70,3%
Profundidade Hierárquica (T)	1 (Plana)	3 (Híbrida)	5 (Complexa)
Risco de Cardinalidade (C)	Nulo	Médio	Crítico
Risco de Anemia Semântica (S)	Crítico	Baixo	Nulo

5.2.2. Cenário B: GBIF → Dublin Core (Achatamento)

Convém contextualizar que o Dublin Core foi projetado intencionalmente como padrão mínimo, plano e generalista, priorizando adoção ampla sobre expressividade estrutural. Sua limitação diante de hierarquias taxonômicas complexas é uma característica de *design* e não uma falha. O que o Cenário B demonstra é que a adoção do DC como único destino de *crosswalk* para dados de alta complexidade estrutural impõe um custo informacional mensurável e sistematicamente invisível aos validadores convencionais.

A Conformidade $Val = 1$ (100%) coexiste com perda média $L_w = 51{,}15\%$ ($\sigma = 14{,}06$ pp, min=41,54%, max=73,24%), evidenciando o fenômeno da validade artificial (Tabela 5). A distribuição é bimodal. Os 36 datasets sem `taxonomicCoverages` atingem $L_w = 41{,}54\%$; 14 com essa dimensão alcançam $L_w = 73{,}24\%$, a presença de hierarquias taxonômicas é o preditor determinístico da perda.

O GBIF opera com objetos complexos de alta granularidade. O padrão DC oferece apenas `dc:subject`, identificando conflito crítico de Tipo (T). A concatenação dos valores preserva conformidade ($Val = 1$), mas anula a inteligência semântica ao destruir a distinção entre níveis taxonômicos. A conformidade estrutural atua como falso positivo de qualidade.

Tabela 4. Extrato de Auditoria Ponderada (Amostra UFSCar – Cenário A)

ID	$\sum wu_{src}$	$\sum wu_{tgt}$	Unidades Perdidas	Val. L_w(%)
R01–R10	6,9	5,33	`identificacao`, `tax_espec.`	Sim (1) 22,75%
Média	6,90	5,33	*(achat. de tipo e downgrade PID)*	100% 22,75%

Tabela 5. Extrato de Auditoria Ponderada (Amostra GBIF – Cenário B).

ID	$\sum wu_{src}$	$\sum wu_{tgt}$	Unidades Perdidas	Val. L_w
R01, R05, R06	17,0	4,55	`tax_*`, `estrut_geo`	Sim 73,24%
R03	8,0	3,80	`ident.`, `estrut_geo`, `est_tipo`	Sim 52,50%
R02, R04, R07–R10	6,5	3,80	`ident.`, `estrut_tipo`	Sim 41,54%
Média	**9,80**	**4,02**	*(achat. domínio cient.)*	**100% 51,15%**

5.2.3. Cenário C: GBIF → DataCite (Preservação estrutural)

A adoção do DataCite reduz dramaticamente a entropia: $L_w = 7{,}93\%$ ($\sigma = 12{,}85$ pp), com $Val = 100\%$ (Tabela 6). Os 36 datasets sem taxonomia atingem $L_w = 0{,}00\%$; os 14 com DwC-A apresentam $L_w = 28{,}32\%$, com perda concentrada em `tax_lineana` e `tax_auditoria` ($w_i = 2{,}0$) — sem mapeamento formal em DataCite 4.x.

Tabela 6. Extrato de Auditoria Ponderada (Amostra GBIF – Cenário C).

ID	$\sum wu_{src}$	$\sum wu_{tgt}$	Unidades Perdidas	Val. L_w
R01, R05, R06	17,0	12,0	`tax_*`, `estrut_geo`	Sim 29,41%
R03	8,0	8,00	–	Sim 0,00%
R02, R04, R07–R10	6,5	6,50	–	Sim 0,00%
Média	**9,80**	**8,30**	*achat. domínio cient.*	**100% 7,93%**

A Tabela 7 decompõe o ganho ponderado (DataCite vs. DC) por categoria, evidenciando que os campos `identificacao`, `estrut_tipo` e `estrut_geo` concentram a diferença de 43,22 pontos percentuais entre os cenários B e C.

Tabela 7. Decomposição do ganho ponderado: DataCite vs. Dublin Core (50 registros GBIF).

Categoria	w_i	$\bar{u}_{src}$	$\bar{u}_{tgt}^{DC}$	$\bar{u}_{tgt}^{DataCite}$	Ganho (Δw)
`identificacao`	3,0	1,000	0,500	1,000	**+1,500**
`estrut_tipo`	1,5	1,000	0,200	1,000	**+1,200**
`estrut_geo`	1,5	0,340	0,000	0,340	+0,510
`estrut_data_attr`	1,5	0,300	0,000	0,300	+0,450
`tax_especificidade`	2,0	0,280	0,000	0,140	+0,280
`estrut_temp`	1,5	0,300	0,150	0,300	+0,225
`tax_lineana`	2,0	0,260	0,000	0,000	0,000
`tax_auditoria`	2,0	0,260	0,000	0,000	0,000
`desc_ling`	1,0	1,000	1,000	1,000	0,000
`desc_cont`	1,0	1,000	1,000	1,000	0,000
Total			**4,02**	**8,19**	**+4,165**

A Tabela 8 decompõe a perda por dimensão da tupla $\langle S, T, C, V \rangle$, revelando que o perfil de degradação difere estruturalmente entre cenários.

Tabela 8. Diagnóstico de causa raiz da perda informacional ponderada (L_w).

Dimensão	Cenário A (UFSCar → DC)	Cenário B (GBIF → DC)	Cenário C (GBIF → DataCite)
Semântica (S)	38%	13%	81%
Tipo de Dado (T)	53%	44%	19%
Cardinalidade (C)	0%	4%	0%
Valor/Acurácia (V)	9%	39%	0%

5.3. Análise Estatística

A Tabela 9 consolida as estatísticas descritivas de L_w por cenário. A ANOVA *one-way* rejeita a hipótese nula de igualdade entre médias ($F(2, 147) = 187{,}65$, $p < 0{,}001$). O teste *post-hoc* de Welch com correção de Bonferroni indica diferenças significativas entre todos os pares: A vs. B ($t = -13{,}61$), A vs. C ($t = 8{,}16$), B vs. C ($t = 15{,}58$), todos com $p < 0{,}001$.

Tabela 9. Estatísticas descritivas de L_w por cenário ($n = 50$; IC 95%, distribuição t de Student).

Cenário	Transformação	Média L_w	$\hat{\sigma}$ (pp)	Mín.	Máx.	IC 95%
A	UFSCar → DataCite	22,75%	0,00	22,75%	22,75%	[22,75; 22,75]
B	GBIF → Dublin Core	51,15%	14,06	41,54%	73,24%	[47,15; 55,15]
C	GBIF → DataCite	7,93%	12,85	0,00%	28,32%	[4,28; 11,58]

O Cenário A apresenta $\hat{\sigma} = 0$, uniformidade decorrente da homogeneidade do Dublin Core. O Cenário B exibe distribuição bimodal ($\hat{\sigma} = 14{,}06$ pp), pois a presença ou ausência de hierarquias taxonômicas é o preditor determinístico da perda. A separação entre B e C permanece superior a 41 pontos percentuais em todas as configurações de pesos testadas (análise de sensibilidade com esquemas *Flat* e *Taxonomy-Heavy*), confirmando que a degradação do Cenário B é estrutural e independente da ponderação adotada. A Figura 2 e a Figura 3 sintetizam visualmente o *trade-off* e as distribuições.

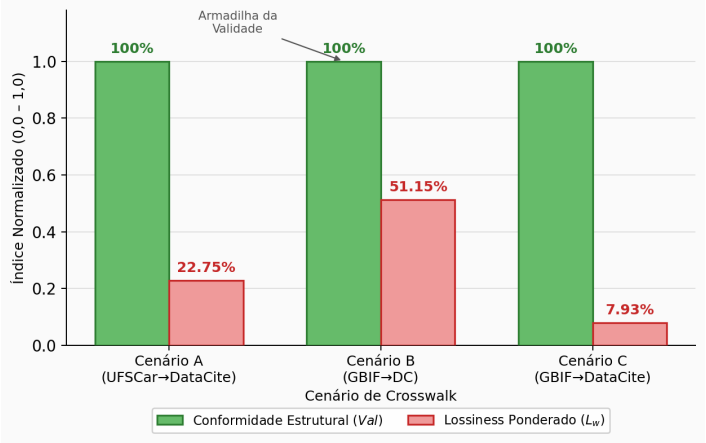

Figura 2. Visualização do *Trade-off*: validade *versus* perda de informação.

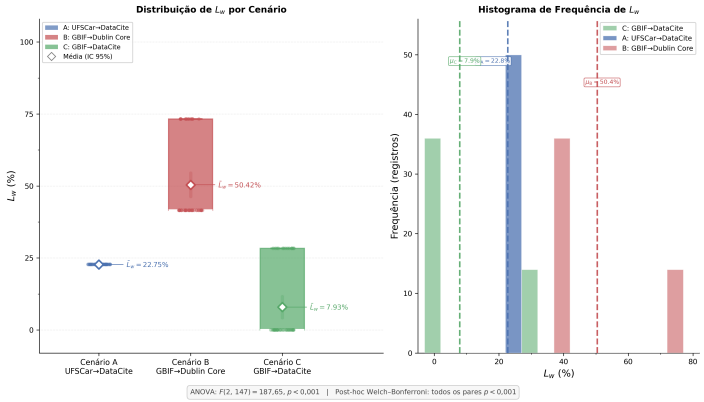

Figura 3. Distribuição de L_w nos três cenários ($n = 50$/cenário). ANOVA: $F(2, 147) = 187{,}65$, $p < 0{,}001$.

6. Discussão

Os três cenários convergem para um achado central, pois *conformidade estrutural e fidelidade informacional são propriedades independentes*. O Cenário B demonstra $Val = 100\%$ simultâneo a $L_w = 51{,}15\%$. Essa dissociação é uma propriedade dos próprios esquemas que o modelo torna mensurável sistematicamente.

O Conflito de Tipo (T) é a causa dominante no Cenário A ($T = 53\%$) e relevante no B ($T = 44\%$). Manifesta-se quando o destino requer tipo primitivo onde a origem armazena objeto composto. A conversão é tecnicamente viável (`taxonomicCoverages` → *"Animalia; Arthropoda"*), mas destrói a estrutura hierárquica. Este é o mecanismo central da validade artificial, quando o registro passa na validação XSD com informação degradada.

O Conflito de Valor (V) concentra-se no Cenário B ($V = 39\%$) e é nulo no C ($V = 0\%$). O DataCite, ao possuir vocabulários controlados formalizados para todos os campos mapeáveis do GBIF, elimina integralmente esse tipo de conflito.

O Conflito Semântico (S) domina o Cenário C ($S = 81\%$), a perda residual irredutível em DataCite. `tax_lineana` e `tax_auditoria` não possuem equivalente formal em DataCite 4.x. Esse conflito não pode ser resolvido pela escolha de padrão de destino, pois requer extensão do esquema ou adoção de DwC-A. A evidência empírica fornece argumento quantificado para que comunidades de biodiversidade negociem extensões de DataCite.

Os *frameworks* FAIRness existentes (F-UJI, FAIR-Checker, FAIRsFAIR [Devaraju et al. 2022], RDA FDMM [RDA FAIR Data Maturity Model Working Group 2020]) avaliam o estado estático de um repositório. Um registro pode chegar FAIR-compliant tendo perdido 73% de seu valor científico em um *crosswalk upstream*, nenhum *framework* existente detectaria isso, pois todos avaliam o estado final. O modelo proposto introduz uma camada de auditoria de transformação que se posiciona entre origem e destino, complementando os instrumentos existentes. Sua integração natural é como componente do MSCR, que registra *crosswalks* como artefatos FAIR. O modelo adicionaria a cada *crosswalk* um indicador L_w, tornando visível o custo informacional de cada mapeamento antes de sua adoção.

Políticas de interoperabilidade que definem conformidade com Dublin Core como requisito suficiente estão, inadvertidamente, sancionando perdas de até 73% para dados de alta complexidade. A limitação é intrínseca ao poder expressivo do DC para representar objetos hierárquicos [Haslhofer and Klas 2010]. Uma política orientada à preservação de valor científico precisaria estratificar requisitos por arquétipo de esquema e exigir um L_w mínimo aceitável.

7. Conclusão

Este trabalho propôs e validou um *framework* para avaliação de interoperabilidade semântica em *crosswalks*, formalizado pela tupla $\langle S, T, C, V \rangle$ e operacionalizado pela Métrica L_w. A evidência empírica de 150 registros demonstrou que conformidade estrutural e fidelidade informacional são propriedades independentes, visto que o Cenário B atingiu $Val = 100\%$ com $L_w = 51{,}15\%$ (pico 73,24%). A análise estatística confirmou diferenças altamente significativas ($F(2, 147) = 187{,}65$, $p < 0{,}001$), com separação de 43,22 pp entre cenários B e C para a mesma origem GBIF.

As três perguntas de pesquisa foram respondidas. Na RQ1, a validade e perda semântica coexistem integralmente (Cenário B). Na RQ2, a escolha do esquema de destino é o fator determinante ($L_w^B = 51{,}15\%$ vs. $L_w^C = 7{,}93\%$, diferença de 43,22 pp). E na RQ3, a tupla $\langle S, T, C, V \rangle$ e L_w formalizam essa distinção com estabilidade robusta à variação de pesos.

O *framework* tem três aplicações imediatas. A triagem pré-migração calcula L_w projetado para cada esquema candidato antes de migração irreversível. A política de repositório adota $L_w \leq \theta$ como pré-condição para aprovação de *crosswalks*. E a extensão do MSCR acrescenta indicador de custo informacional ausente nos validadores sintáticos atuais.

Como limitações, o escopo de três repositórios não cobre esquemas híbridos de ciências da saúde e sociais, e a dimensão S depende do Dicionário de Governança, cuja verificação de equivalência conceitual requer especialistas de domínio. Esse é o principal gargalo de escalabilidade, pois a curadoria manual torna-se impraticável em repositórios de grande porte ou ecossistemas multidisciplinares. Os caminhos para mitigá-lo incluem ontologias de domínio consolidadas (ENVO, MeSH, SNOMED-CT) para automação parcial de S, aprendizado de máquina para sugerir mapeamentos a partir de dicionários curados e governança colaborativa distribuída entre comunidades de prática. Note-se ainda que a responsabilidade pelo mapeamento correto recai sobre quem *consome* os metadados, e não apenas sobre quem os publica, o que motiva validação pré-migração pelo lado do consumidor. Os pesos w_i derivados do FAIR podem não refletir prioridades de domínios com lógica de reuso distinta, e a simulação analítica *versus* execução real requer validação comparativa. A agenda futura inclui expansão para novos domínios, validação com *pipelines* reais (OpenRefine, XSLT) e implementação como microsserviço em DSpace e Dataverse.

A sustentabilidade da Ciência Aberta depende de ecossistemas que tratem a fidelidade semântica como requisito mensurável e auditável. O *framework* $\langle S, T, C, V \rangle$ oferece os instrumentos conceituais e operacionais para isso.

Referências

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):1–52.

Bernstein, P. A. and Melnik, S. (2007). Model management 2.0: manipulating richer mappings. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, page 1–12, New York, NY, USA. Association for Computing Machinery.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, Cambridge, MA.

Bruce, T. R. and Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In *Metadata in Practice*. ALA Editions, Chicago.

Candela, L., Mangione, D., and Pavone, G. (2024). The FAIR assessment conundrum: Reflections on tools and metrics. *Data Science Journal*, 23:33.

Devaraju, A., Huber, R., Mokrane, M., et al. (2022). FAIRsFAIR data object assessment metrics (0.5). *Zenodo*.

Doerr, M. (2003). The cidoc conceptual reference model: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92.

Ehrig, M. (2007). *Ontology Alignment: Bridging the Semantic Gap*. Number 4 in Semantic Web and Beyond. Springer, New York.

EOSC Association – FAIR Metrics and Data Quality Task Force (2022). Towards a data quality framework for EOSC. Technical report, EOSC Association.

EOSC Association – Semantic Interoperability Task Force (2023). Converging on a semantic interoperability framework for EOSC. Technical report, EOSC Association.

Euzenat, J. and Shvaiko, P. (2013). *Ontology Matching*. Springer, 2nd edition.

Gradmann, S. (2010). Knowledge = information in context: on the importance of semantic contextualisation in europeana. White Paper 1, Europeana, The Hague.

Haslhofer, B. and Klas, W. (2010). A survey of techniques for managing metadata evolution. *ACM Computing Surveys*, 42(3):1–37.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246.

Obrst, L. (2003). Ontologies for semantically interoperable systems. In *Proceedings of the 12th ACM International Conference on Information and Knowledge Management (CIKM '03)*, pages 366–369, New Orleans, Louisiana, USA. ACM.

Park, J.-R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3-4):213–228.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211–218.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350.

RDA FAIR Data Maturity Model Working Group (2020). FAIR data maturity model: Specification and guidelines. Technical report, Research Data Alliance.

Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., and Dumontier, M. (2018). A design framework and exemplar metrics for fairness. *Scientific Data*, 5(1):180118.

Zeng, M. L. and Chan, L. M. (2006). Metadata interoperability and standardization: A study of methodologies and best practices at the system level. *D-Lib Magazine*, 12(6).