UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Marília

Ananda Fernanda de Jesus

Qualidade de dados *Linked Data* para *seleção de fontes* e *criação de links*: estudo teórico, terminológico e processual

Marília
2025

Ananda Fernanda de Jesus

Qualidade de dados *Linked Data* para *seleção de fontes* e *criação de links*: estudo teórico, terminológico e processual

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação, como requisito para a obtenção do título de Doutor em Ciência da Informação pela Faculdade de Filosofia e Ciências, Universidade Estadual Paulista (UNESP), Campus de Marília.

Área de Concentração: Informação, Tecnologia e Conhecimento.

Linha de Pesquisa: Informação e Tecnologia

Orientador: Dr. José Eduardo Santarem Segundo

Marília
2025

J58q

Jesus, Ananda Fernanda de
Qualidade de dados Linked Data para seleção de fontes e criação de links : estudo teórico, terminológico e processual / Ananda Fernanda de Jesus. -- Marília, 2025
277 p.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Filosofia e Ciências, Marília
Orientador: José Eduardo Santarem Segundo

1. qualidade de dados. 2. linked data. 3. processo de avaliação de qualidade. 4. revisão sistemática da literatura. 5. design science research. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Impacto potencial desta pesquisa

Entende-se que ao trabalhar o processo de seleção de dados *Linked Data*, levando em consideração diversos aspectos de qualidade, o potencial impacto social da presente tese perpassa de maneira indireta a distintos Objetivos do Desenvolvimento Sustentável (ODS), promovendo o acesso à informação para tomada de decisão e auxiliando no monitoramento dos esforços relacionados a esses ODS. Destaca-se a contribuição para o ODS 16, no que tange a busca por instituições eficazes, responsáveis e transparentes. A seleção e aplicação de dados *Linked Data* contribui para fortalecer a transparência em ambientes governamentais e melhorar a gestão e a elaboração de políticas públicas. Também tem o potencial de auxiliar no processo eficiente de tomada de decisão "responsiva, inclusiva, participativa e representativa", como destacado no item 16.7**,** assim como contribuir para viabilizar o acesso público a informação, indicado no item 16.10.

Potential impact of this research

Working on the Linked Data selection process, considering various aspects of quality, the potential social impact of this thesis indirectly permeates different Sustainable Development Goals (SDGs), promoting access to information for decision-making and assisting in the monitoring of efforts related to these SDGs. The contribution to SDG 16 stands out, regarding the search for effective, accountable and transparent institutions. The selection and application of Linked Data have the potential to contribute to strengthening transparency in government environments and improving the management and development of public policies. It also has the potential to assist in the efficient process of "responsive, inclusive, participatory and representative" decision-making, as highlighted in item 16.7, and contribute to enabling public access to information, indicated in item 16.10.

Ananda Fernanda de Jesus

Qualidade de dados *Linked Data* para *seleção de fontes e criação de links*: estudo teórico, terminológico e processual

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Universidade Estadual Paulista – Campus de Marília, como requisito para a obtenção do título de doutor em Ciência da Informação.

Área de Concentração: Informação, Tecnologia e Conhecimento.

Linha de Pesquisa: Informação e Tecnologia

Banca Examinadora

Prof. Dr. José Eduardo Santarem Segundo
UNESP – Câmpus de Marília
Orientador

Prof. Dr. Mario Guido Barité Roqueta
Universidad de la República, Uruguay

Prof. Dr. Caio Saraiva Coneglian
UNESP – Câmpus de Marília
UNIMAR – Universidade de Marília

Prof. Dr. Felipe Arakaki
UnB – Universidade de Brasília

Profa. Dra. Natália Marinho do Nascimento
UNESP – Câmpus de Marília

Marília, 01 de dezembro de 2025.

AGRADECIMENTOS

Ao meu orientador, por apoiar as minhas ideias, sempre confiar no meu trabalho e julgamento, me dar o espaço que precisei para me desenvolver enquanto pesquisadora, mas também pelas conversas, trocas e pelo companheirismo que tivemos ao longo de todos esses anos.

Ao grupo de pesquisa GTERM, na figura do professor Dr. Mario Barité, por terem me acolhido e compartilhado seus conhecimentos, que foram extremamente enriquecedores para a presente tese, mas também para meu desenvolvimento pessoal e profissional.

Aos amigos e amigas que fiz no Uruguai, nas pessoas de Eugenia, Lucia, Mariana, Melanie e Leandro, que me acolheram amavelmente e compartilharam comigo suas experiencias em todos os lugares incríveis que conhecemos juntos, mas também nos momentos pequenos de convivência do dia a dia, agradeço por tornarem a difícil experiência de estar longe de casa muito mais tranquila.

Ao Wes, com quem compartilhei os momentos incríveis, assustadores e desafiadores que estiveram por trás da escrita da presente tese, agradeço por ter me escutado quando estava empolgada, desanimada ou desesperada, por ter lido tantos dos meus textos e me escutado praticando as minhas apresentações. Agradeço profundamente por ter tanta certeza de que tudo daria certo, por acreditar em mim mesmo quando eu não acreditava.

A Carla, a outra pessoa que também sempre teve muito mais certeza que eu de que tudo daria certo, agradeço as muitas horas de conversas, pelo apoio e suporte ao longo de todo esse processo.

A Karen e Kazumi, que junto com a Carla, sempre se fizeram presentes, mesmo com a constante distância geográfica, agradeço pelas risadas e surtos compartilhados.

A Ligia por ser minha dupla acadêmica, por compartilhar comigo toda essa experiência e torná-la muito menos solitária.

A minha família, na figura dos meus pais, Ivan e Rosangela, dos avós, Aparecida, Waldemar, Dona Dú e seu Zé, dos muitos tios e tias, nas pessoas de Nilva e João, por me tornarem a pessoa que sou, sem eles nada disso seria possível.

A universidade pública que me abriu tantas portas e me permitiu realizar sonhos que antes nem mesmo ousava sonhar, sou eternamente grata.

Aos membros da banca pelas enriquecedoras contribuições para o desenvolvimento da presente tese.

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil. Processo nº 2021/03349-0.

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

RESUMO

Os dados têm se destacado como insumo para o desenvolvimento e melhoria de produtos, processos e serviços em distintos setores da sociedade, levando a busca por formas de garantir a sua recuperação, reuso e qualidade. Nesse contexto, uma das possibilidades é o *Linked Data*, conjunto de princípios elaborado para a publicação estruturada e conectada de dados na *Web*. A adoção do *Linked Data* não garante a qualidade desses dados, que estão sujeitos a diferentes níveis de curadoria, problemas estruturais, relacionados à veracidade e precisão de seu conteúdo. Um dos desafios para a adoção do *Linked Data* é o processo de seleção de dados, levando a questão norteadora da pesquisa: como selecionar dados *Linked Data* para criação de *links* com fontes externas? A presente tese parte da hipótese de que as características dos dados *Linked Data* influenciam na terminologia e no processo de avaliação de qualidade de dados, e de que esse processo pode ser mais bem compreendido por meio de um estudo teórico exaustivo, da análise terminológica e do estudo do processo de avaliação de qualidade de dados *Linked Data.* O objetivo foi fomentar maior clareza teórica e terminológica e maior compreensão a respeito da seleção de dados *Linked Data*, por meio de sua relação com o processo de avaliação de qualidade, abordando definições, ferramentas e produtos relacionados, mapeando os agentes, etapas e atividades da seleção de dados *Linked Data* para criação de links com fontes externas. Como resultados, propôs-se a elaboração dos seguintes produtos: glossário da qualidade de dados *Linked Data*, fluxo do processo de seleção de dados *Linked Data, checklist* e modelo de protocolo para auxiliar na seleção de fontes. Para a composição desses produtos foi adotada uma abordagem multimetodológica, dividida nos seguintes aspectos: 1) Criação de corpus teórico exaustivo – pautado em análise exploratória, Revisão Sistemática da Literatura e estudo documental; 2) Elaboração do glossário – pautado em estudo terminológico com aplicação do método da Grade e elaboração de Árvore de Domínio; 3) Elaboração do fluxo, modelo de protocolo e *checklist* e para seleção de dados *Linked Data para* interligaçã*o* – pautado em *Design Science Research*. Os resultados foram divididos em estudo teórico, terminológico e processual. Nos resultados do estudo teórico foram discutidos os principais aspectos da qualidade de dados e sua relação com a Ciência da Informação. Levantou-se ainda o estado da arte da qualidade de dados *Linked Data*, identificando e analisando os instrumentos e ferramentas disponíveis para auxiliar nesse processo. Na etapa de estudo terminológico foi estabelecida a árvore de domínio da qualidade de dados *Linked Data*, bem como os termos que compõe o glossário. Na etapa de estudo processual foram discutidos os processos, etapas, atividades e instrumentos que perpassam a seleção de dados *Linked Data* para interligação. Com base nas três etapas foi composto o fluxo para seleção de dados *Linked Data*, dividido em duas abordagens: pautada em análise exploratória e pautada em explicitação de necessidade informacional previamente estabelecida. Foram apresentados ainda protocolos para auxiliar na condução de ambas as abordagens e uma *checklist* composta por critérios de exclusão que auxiliam na seleção de dados. Conclui-se que a seleção de dados para interligação é um processo complexo e amplamente contextual, que depende de uma etapa de planejamento prévio, do estabelecimento de um modelo de qualidade, da abordagem a ser adotada e dos objetivos a serem alcançados e que pode ser facilitado pela adoção de diferentes ferramentas.

Palavras–chave: qualidade de dados; linked data; processo de avaliação de qualidade; revisão sistemática da literatura; design science research.

ABSTRACT

Data has emerged as an input for the development and improvement of products, processes, and services in different sectors of society, leading to search for ways to ensure its recovery, reuse, and quality. In this context, one possibility is Linked Data, a set of principles developed for the structured and connected publication of data on the web. Adopting Linked Data does not guarantee the quality of this data, which is subject to varying levels of curation, structural issues, and issues related to the veracity and accuracy of its content. One of the challenges in adopting Linked Data is the data selection process, leading to the guiding question of this research: how to select Linked Data for creating links to external sources? This thesis is based on the hypothesis that the characteristics of Linked Data influence the terminology and the data quality assessment process. This process can be better understood through exhaustive theoretical study, terminological analysis, and the study of the Linked Data quality assessment process. The objective was to conduct a theoretical, terminological, and procedural study of Linked Data quality, addressing definitions, processes, instruments, and products, aiming to foster greater terminological clarity and a deeper understanding of Linked Data selection, based on the quality assessment process. As a result, we propose the development of two products: a Linked Data quality glossary and a Linked Data selection process flowchart. A multi-methodological approach was adopted, divided into the following aspects: 1) Creation of an exhaustive theoretical corpus—based on exploratory analysis, a systematic literature review, and documentary study; 2) Development of the glossary—based on a terminological study applying the Grid method and developing a Domain Tree; 3) Development of the Linked Data selection flowchart—based on Design Science Research. The results were divided into theoretical, terminological, and procedural study results. The theoretical study stage discussed the main aspects of data quality and its relationship to Information Science. The state-of-the-art in Linked Data quality was also assessed, identifying and analyzing the instruments and tools available to assist in this process. During the terminology study stage, a Linked Data quality domain tree was established, based on which the terms that comprise the glossary were established. The definitions that comprise the glossary were also constructed and presented. During the procedural study stage, the processes, steps, activities, and instruments involved in Linked Data selection for creating links with external sources were discussed. Based on these three steps, a Linked Data selection flow was developed, divided into two approaches: one based on exploratory analysis and the other based on the explanation of previously established information needs. Protocols to assist in conducting both approaches were also presented, as well as a checklist consisting solely of exclusion criteria to aid in data selection. It is concluded that the selection of data for interconnection is a complex and largely contextual process, which depends on a prior planning stage, the establishment of a quality model, the approach to be adopted and the objectives to be achieved and which can be facilitated by the adoption of different tools

Keywords: data quality; linked data; quality assessment process; systematic literature review; design science research.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, os dados passaram a ser um objeto de estudo de diversas áreas do conhecimento, ganhando destaque em discussões que atingiram, inclusive, o público não especializado.

Na ciência, destaca-se a preocupação com a qualidade e o compartilhamento dos dados de pesquisa. No âmbito das organizações, destaca-se a preocupação com o uso estratégico dos dados para a ampliação da vantagem competitiva, diminuição dos gastos e aumento dos lucros. No âmbito social, ressalta-se a preocupação com o uso ético, seguro e responsável desses dados, especialmente tratando-se de dados pessoais e/ou sensíveis.

Esse aumento nas discussões a respeito de dados pode ser associado a dois fenômenos principais: o crescimento exponencial da geração de dados e os avanços relacionados a técnicas e tecnologias para o processamento desses dados.

> O mundo está repleto de dados. Dados são criados toda vez que alguém utiliza seu celular, onde quer que esteja, que produtos com seus códigos de barras são fabricados, despachados, armazenados e vendidos, e que veículos com GPS vão e vêm pelas estradas, circulam pela rede e são passíveis de serem analisados, processados e transformados em informações de valor (Isotani; Bittencourt, 2015, p. 12).

Esse aumento na produção de dados teve início com a popularização dos computadores portáteis, atraindo a atenção de áreas de estudo como a computação e de diversas organizações, públicas e privadas, levando a preocupações sobre como avaliar e garantir a qualidade desses dados.

> A questão do impacto da qualidade dos dados em processos computacionais e de informação são bastante reconhecidos desde meados do século XX, visto que, a partir do desenvolvimento e popularização dos primeiros computadores digitais com uma maior capacidade de processamento (*mainframes*) a quantidade de geração e processamento de dados vem crescendo, literalmente explodindo, a partir dos anos 1990 com a massificação dos microcomputadores e uma pletora de outros dispositivos [...] (Dias *et al.*, 2023, p. 4).

Esse cenário foi amplificado pelo barateamento dos sensores que permitem coletar dados em tempo real a respeito de diversos fenômenos, pela popularização

da *Web*, pelo estabelecimento das redes móveis e pelo uso constante de *smartphones* por uma parte significativa da população.

Nesse contexto, a geração de dados muitas vezes ocorre de maneira involuntária, dados são gerados quando uma pessoa se movimenta com seu celular no bolso, quando dorme com relógios que monitoram o sono e até quando estaciona em uma vaga dotada de sensores, em uma simples ida ao supermercado.

Além do aumento da quantidade de dados, o aumento da capacidade computacional para o processamento de dados e os avanços das técnicas e tecnologias para seu processamento e visualização, ampliaram o interesse por dados. Entretanto, esses avanços não eliminaram a preocupação com a qualidade dos dados.

> Mesmo como os avanços nos produtos de *software* que possibilitam a consistência nos processos de tratamento de dados, os desafios persistem. O impacto da possibilidade de entrada de 'lixo' afeta todos os tipos de sistema, desde os sistemas tradicionais de folha de pagamento até os mais sofisticados sistemas de Inteligência Artificial (IA), baseados em modelos de aprendizado profundo, que são absolutamente dependentes de vastos volumes de dados (Dias *et al.*, 2023, p. 4).

Uma das possibilidades para melhorar a recuperação e o reuso de conjuntos de dados são os princípios do *Linked Data,* propostos em 2006 por Tim Berners-Lee, para orientar a publicação e conexão de dados na *Web.* Esses princípios foram pensados para facilitar o uso de agentes computacionais na recuperação da informação, promover resultados de busca mais significativos para os usuários e facilitar a aplicação e o reuso de dados para a geração de informações.

Inicialmente foram propostos quatro princípios, sendo disponibilizados ao longo dos anos uma série de guias e melhores práticas para auxiliar na publicação de dados como *Linked Data.* De acordo com Berners-Lee (2006) os princípios são: 1) Use *uniform resource identifier* (URIs)*;* 2) Use *Hypertext Transfer Protocol* (HTTP) URIs; 3) Forneça informações utilizando o *Resource Description Framework* (RDF) e o *Protocol and Resource Description Framework Query Language* (SPARQL); e 4) Inclua *links* para outros URIs.

Entre os desafios para a publicação e uso de dados *Linked Data*, destaca-se o processo de seleção de conjuntos de dados. Embora publicar dados de acordo com

esses princípios forneça certo nível de estruturação, esses dados também estão sujeitos a más formações e a problemas com a precisão, atualidade e a veracidade de seus conteúdos.

A seleção de conjuntos de dados com base em aspectos de qualidade é um resultado do processo de avaliação de qualidade, discutido e aplicado no âmbito da qualidade de dados, que permite quantificar e qualificar os níveis de qualidade desses dados. A qualidade de dados, enquanto um domínio, teve sua estruturação entre o final da década de 1960 e o início da década de 1970. (Batini; Scannapieco, 2006; Langer *et al.*, 2018).

Desde o princípio, a qualidade de dados se estabelece como um domínio de caráter interdisciplinar, atraindo pesquisadores e profissionais de diversas áreas, como a engenharia, ciência da computação e estatística (Batini; Scannapieco, 2006). Soma-se a essa lista os profissionais da Ciência da Informação (CI), que de maneira mais recente, tem se dedicado a discutir diversos aspectos relacionados a qualidade de dados.

Com os dados se estabelecendo como insumo para o funcionamento de organizações do setor público e privado, bem como para melhoria de diversas atividades cotidianas, amplifica-se a preocupação sobre como selecionar conjuntos de dados adequados, tendo em vista que a qualidade dos conjuntos de dados está diretamente relacionada com a qualidade dos produtos, processos e serviços derivados da aplicação dos dados.

Wang e Strong (1996) realizaram uma sistematização do processo de avaliação de qualidade, sendo ainda hoje base para os estudos e discussões em qualidade de dados. Os autores dividem o processo de avaliação de qualidade em categorias, dimensões, critérios e métricas de qualidade. As categorias, também denominadas como classes ou perspectivas, afetam a forma como se estuda, discute e avalia a qualidade de dados. As principais categorias de qualidade são: intrínsecas, contextual, representacional e acessibilidade.

O processo de avaliação pode levar em consideração uma ou todas essas categorias de qualidade. Entretanto, quando pensadas como perspectivas, elas refletem um processo de decisão, que irá estabelecer os parâmetros para definir se um conjunto de dados possui ou não os níveis necessários de qualidade.

Entre as categorias de qualidade, destaca-se na literatura a dimensão contextual, sendo relacionada ao conceito de "*fitness for use*", ou, adequado ao uso

(Zaveri *et al.*, 2012; Nooghabi; Dastgerdi, 2016; Ahmed, 2017). O aspecto contextual se destaca, pois, conjuntos de dados perfeitamente adequados para determinada tarefa podem não atuar satisfatoriamente em outro contexto. Além disso, mesmo em processos de avaliação focados em outras perspectivas, não é possível se desvincular totalmente de algumas questões contextuais.

Quando se trata de aspectos intrínsecos, para identificar se um conjunto de dados está livre de erros, é necessário entender seu contexto de criação, identificando como foram gerados, se tiveram como base modelos, formatos, princípios etc. Tanto em aspectos representacionais, como nos de acessibilidade, muitas vezes torna-se necessário considerar o público-alvo dos dados.

Nesse mesmo sentido, avaliar a qualidade de dados *Linked Data* impõe uma série de desafios próprios, que perpassam todas as perspectivas de qualidade de dados. Esses desafios são relacionados ao contexto de criação, adequação as diretrizes e melhores práticas existentes e ainda à estrutura desses dados, especialmente pautados no *RDF*.

Com base nesse contexto, a presente pesquisa partiu do questionamento: como selecionar dados *Linked Data* para criação de *links* com fontes externas? As próximas subseções apresentam a tese, a hipótese e a proposição da presente pesquisa, bem como seus objetivos, justificativa e estrutura.

1.1 Tese, hipótese e proposição da pesquisa

Com base no contexto apresentado, a tese defendida consiste em que a seleção de dados *Linked Data* depende da avaliação de qualidade desses dados, demandando conhecimento a respeito da sua terminologia, a escolha da perspectiva a ser adotada, a seleção de dimensões, critérios e métricas, bem como o conhecimento das diferentes ferramentas e metodologias que podem auxiliar na sua realização.

Parte-se da hipótese de que as características dos dados *Linked Data* influenciam na terminologia e no processo de avaliação de qualidade de dados, e de que esse processo pode ser mais bem compreendido por meio de um estudo teórico exaustivo, da análise terminológica e do estudo do processo de avaliação de qualidade de dados *Linked Data*.

Como resultados, foram propostos dois produtos: glossário para a qualidade de dados *Linked Data* e fluxo do processo de seleção de dados *Linked Data*. A elaboração desses produtos foi pautada nos seguintes procedimentos: 1) **Criação de um *corpus* teórico exaustivo –** baseado em análise exploratória, Revisão Sistemática da Literatura (RSL) e estudo documental; 2) **Elaboração do glossário –** pautado em estudo terminológico com aplicação do método da Grade e elaboração de Árvore de Domínio; 3) **Elaboração do fluxo para seleção de dados *Linked Data* *para interligação* –** pautado no *Design Science Research*.

1.2 Justificativa

O processo de seleção de dados *Linked Data* é um dos principais desafios para a adoção desses princípios. É no âmbito da qualidade de dados que se desenvolvem processos, produtos e instrumentos que permitem selecionar dados apropriados para diferentes aplicações. A qualidade de dados torna-se, portanto, fundamental para o processo de seleção de dados *Linked Data*, objeto de estudo da presente tese.

O estudo da qualidade de dados é complexo, por seu caráter multidimensional, interdisciplinar e, em muitos casos, contextual. Avaliar a qualidade de dados depende da escolha da perspectiva desse estudo, da seleção de critérios e métricas, do conhecimento acerca das diferentes ferramentas e metodologias que podem auxiliar na realização desse processo.

Soma-se a essa complexidade o necessário conhecimento não só dos conceitos basilares da qualidade de dados, mas também a complexidade intrínseca aos dados publicados como *Linked Data*, suscetíveis a diferentes problemas de qualidade, devido a sua estrutura, a forma como são criados e disponibilizados.

Nesse sentido, a qualidade de dados *Linked Data* possui características e termos próprios, que fazem com que os termos gerais da qualidade de dados assumam facetas diferentes, que precisam ser discutidas e compreendidas, e que levam a necessidade de adaptação desses conceitos para que representem adequadamente o processo de avaliação de dados *Linked Data.*

Em seu aspecto científico, a presente pesquisa se justifica pela discussão da qualidade de dados enquanto um objeto de estudo da Ciência da Informação e pela apresentação dos conceitos necessários para o seu entendimento e aplicação.

A qualidade de dados ainda é um objeto de pesquisa recente da área, especialmente em nível nacional, entretanto, a preocupação com a qualidade dos dados perpassa diversos objetos tradicionais da Ciência da Informação, destacando-se as relações com as áreas de representação da informação e organização do conhecimento, que tem entre seus produtos dados e metadados, que precisam ser confiáveis e consistentes.

Os princípios *Linked Data* são um objeto de estudo consolidado na Ciência da Informação, inclusive em nível nacional, tendo uma comunidade de pesquisa própria, o que torna relevante o desenvolvimento da presente pesquisa.

Do ponto de vista dos profissionais da informação, justifica-se pela sistematização do processo de seleção de dados *Linked Data*, que pode ser utilizado por esses profissionais na escolha de fontes de dados, visando reduzir o retrabalho e agilizar os processos de representação e organização de novos recursos informacionais.

Nesse sentido, tanto do ponto de vista dos pesquisadores como dos profissionais da Ciência da Informação, a presente pesquisa se justifica pelos ineditismos dos produtos propostos, considerando que que não existem glossários terminológicos especializados na avaliação de qualidade de dados *Linked Data* e nem fluxos processuais claros que mapeiem a seleção desses dados.

O glossário proposto busca atuar como uma ponte entre as áreas da Ciência da Informação e da Qualidade de Dados, proporcionado um entendimento comum para os termos, processo e produtos relacionados com a avaliação de qualidade e seleção de fontes de dados.

O fluxo, complementado pela proposta de protocolos e *checklist*, mapeia o processo de seleção de fontes, identificando as abordagens da seleção, seus principais agentes, etapas e atividades. Esses aspectos, ainda não explorado pela literatura cientifica, são amplamente necessários tanto para o desenvolvimento de pesquisas como para as práticas profissionais.

No que tange à abordagem social, a relevância desse estudo está relacionada a ampliação da adoção do *Linked Data*, entre os quais destacam-se: melhores resultados de busca, serendipidade e enriquecimento semântico.

Justifica-se também pela relevância da qualidade de dados no contexto tecnológico vigente, marcado por um crescimento exponencial da quantidade de dados e pela multiplicação das ferramentas, técnicas e da capacidade de

processamento desses dados. Nesse contexto, a qualidade dos dados tem impacto direto na qualidade das informações, tornando ainda mais relevantes os processos de avaliação de qualidade.

1.3 Objetivos

Fomentar maior clareza teórica e terminológica e maior compreensão a respeito da seleção de dados *Linked Data*, por meio de sua relação com o processo de avaliação de qualidade, abordando definições, ferramentas e produtos relacionados, mapeando os agentes, etapas e atividades da seleção de dados *Linked Data* para criação de *links* com fontes externas.

Para isso, foram estabelecidos os seguintes objetivos específicos:

- Discutir a qualidade de dados enquanto objeto de estudo da Ciência da Informação;
- Apresentar o estado da arte da qualidade de dados *Linked Data*;
- Estabelecer a Árvore de Domínio da qualidade de dados *Linked Data*, identificando os principais conceitos e traçando as relações hierárquicas entre eles;
- Elaborar um glossário de qualidade de dados *Linked Data*;
- Apresentar um fluxo para seleção de dados *Linked Data* para interligação com fontes externas.

1.4 Estrutura da tese

A presente tese está dividida em 8 seções, sendo elas:

1 Introdução – que apresenta a contextualização necessária para o entendimento da pesquisa, sua pergunta norteadora, a hipótese construída, a justificativa, os objetivos gerais e específicos da pesquisa.

2 Procedimentos metodológicos – onde são apresentados os procedimentos adotados para atingir os objetivos e permitir a construção dos produtos propostos. A pesquisa foi dividida em 5 etapas, sendo apresentada uma subseção para cada uma dessas etapas.

3 A qualidade de dados e a sua relação com a Ciência da Informação – que apresenta os resultados da análise exploratória, com a finalidade de promover familiarização com os termos necessários para a compreensão da qualidade de dados, bem como discutir a sua relação com a Ciência da Informação.

4 A qualidade de dados *Linked Data* – que apresenta os resultados do estudo teórico da qualidade de dados *Linked Data*, focando especialmente na estrutura desses dados. Apresenta ainda os resultados da Revisão Sistemática da Literatura, composto pelo estado da arte da qualidade de dados *Linked Data*, os principais problemas de qualidade que afetam esses dados e as ferramentas que podem auxiliar no seu processo de avaliação. Ao final apresenta ainda os resultados parciais do estudo terminológico conduzido.

5 Estudo terminológico da qualidade de dados *Linked Data* – que apresenta os resultados do estudo terminológico, composto pela apresentação da árvore de domínio e do processo de construção das definições que compõe o glossário.

6 Análise processual da qualidade de dados *Linked Data* – *que apresenta os* resultados da discussão processual da qualidade de dados *Linked Data*, abordando atividades, instrumentos e ferramentas que permeiam a avaliação de qualidade, com foco na sua aplicação na seleção de fontes de dados *Linked Data* para a criação de *links* entre fontes distintas.

7 Fluxo da seleção de fontes e da interligação de dados *Linked Data* – que apresenta os produtos da presente tese, sendo eles os fluxos da seleção de dados *Linked Data*, modelos de protocolo e *checklist* composta por critérios de exclusão auxiliar no processo de seleção.

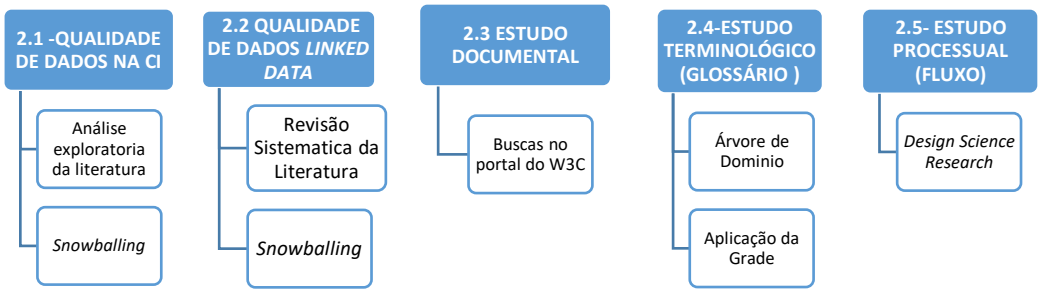
8 Considerações finais – que apresenta uma síntese dos resultados obtidos e as considerações finais derivadas desses resultados, indicando como os objetivos específicos foram atingidos e respondendo à pergunta de pesquisa proposta.

2 PROCEDIMENTOS METODOLÓGICOS

Quanto aos seus objetivos, a pesquisa se caracteriza como exploratória, descritiva e aplicada. No que se refere a análise de seus resultados, a pesquisa pode ser caracterizada como qualitativa, partindo da proposta de elaboração de dois produtos: um glossário de qualidade de dados *Linked Data* e um fluxo para a seleção desses dados.

Para embasar a elaboração de ambos os produtos, e atingir os objetivos propostos, tornou-se necessária uma abordagem multimetodológica. Para uma maior compreensão, as metodologias aplicadas foram distribuídas em 5 etapas. A figura 1 apresenta uma síntese dessas etapas e dos procedimentos/metodologias utilizados para a sua condução, numeradas de acordo com as subseções nas quais serão detalhados os procedimentos metodológicos adotados:

Figura 1 - Síntese das etapas e dos procedimentos empregados na construção da pesquisa.



Fonte: autora (2025)

As primeiras etapas a serem conduzidas foram as relacionadas a construção do *corpus* teórico e documental, já que atuaram como insumo para as demais etapas. Em seguida foram realizados os estudos terminológicos, visando garantir maior aprofundamento e clareza acerca da qualidade de dados *Linked Data*, a identificação dos termos relacionados e a construção das definições para esses termos. O fluxo proposto foi baseado em uma sistematização dos resultados obtidos através da

condução das etapas anteriores. As próximas seções apresentam um maior detalhamento das metodologias e procedimentos adotados em cada uma das etapas.

2.1 A qualidade de dados na CI

Com o intuito de identificar como a Ciência da informação tem discutido a questão da qualidade de dados enquanto objeto de estudo, e de construir uma base para a discussão dos conceitos relacionados a qualidade de dados, foi realizada uma análise exploratória das publicações de qualidade de dados na Ciência da Informação.

Estabeleceu-se o recorte de artigos, trabalhos de evento, teses e dissertações produzidos pela comunidade nacional, sendo observada também as referências adotadas pelos autores através da técnica de *Snowballing,* que “consiste em avaliar a lista de referências e a lista de citações de estudos já conhecidos para identificar novos estudos relevantes para a pesquisa (Silva, 2017, p. 29)”.

Foi utilizada a palavra-chave “qualidade de dados”, com busca nas bases: Base de Dados em Ciência da Informação (BRAPCI), Catálogo de Teses e Dissertações, Anais do Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (ENANCIB) e do Workshop de Informação, Dados e Tecnologia (WIDAT), sendo selecionados artigos, teses e dissertações focados em qualidade de dados.

A análise exploratória buscou identificar como tem sido conduzidas as discussões a respeito da qualidade de dados, como o termo é definido pelos autores, quais os principais enfoques temáticos e como é abordada pela literatura a relação entre Ciência da Informação e qualidade de dados. Para isso, foi coletado nos estudos selecionados: 1) enfoque dos documentos; 2) definição de qualidade de dados; 3) relação entre qualidade de dados e Ciência da Informação; e 4) termos correlatos.

Os enfoques foram organizados com base em categorias estabelecidas a *priori*, sendo elas: 1) criação de artefato de avaliação de qualidade; 2) condução de processo de avaliação de qualidade; 3) abordagem teórica de qualidade de dados.

Os documentos aceitos foram classificados quanto ao seu enfoque no momento da seleção. Para o processo de categorização, foram considerados os objetivos, procedimentos metodológicos e resultados obtidos em cada estudo. Observou-se ainda os principais termos correlatos.

Para analisar como a qualidade de dados é abordada enquanto um objeto de estudo da Ciência da Informação, foram coletados nos estudos aceitos os termos utilizados pelos autores para abordar a qualidade de dados.

Para identificação, foram considerados: 1) termos utilizados nas definições de qualidade de dados apresentadas pelos autores; 2) termos utilizados para explicar a relação entre qualidade de dados e ciência da informação; 3) termos que descrevem os objetos de estudo dos processos de avaliação conduzidos ou das ferramentas de avaliação propostas pelos autores.

2.2 Revisão Sistemática da Literatura

A Revisão Sistemática da Literatura é uma metodologia adotada com a finalidade de garantir a identificação, seleção, avaliação e sumarização de um *corpus* teórico relevante a respeito de determinado fenômeno ou tema de pesquisa (Kitchenham *et al*., 2004). A RSL pode ser definida como:

> Uma revisão bibliográfica acrescida de critérios e de etapas que visam garantir a consistência e a representatividade dos documentos analisados e que preza pelo registro das tomadas de decisão do pesquisador em cada etapa, de forma a permitir que a pesquisa possa ser auditada, replicada e continuada do ponto em que foi interrompida (Jesus, 2021, p. 63).

As Revisões Sistemáticas da Literatura são conduzidas com base em um fluxo de trabalho pré-estabelecido, de maneira **processual**, utilizando i**nstrumentos e seguindo etapas** que visam diminuir o retrabalho, aumentando a confiabilidade, representatividade e reprodutibilidade dos resultados.

De acordo com Felizardo *et al.* (2015) o processo da condução da RSL se realiza entorno de três etapas principais: planejamento, execução e sumarização. Essas etapas são compostas das seguintes tarefas:

- **Planejamento**: Análise exploratória, preenchimento do protocolo de busca que irá orientar a pesquisa e busca piloto;
- **Execução**: Busca nas bases de dados, seleção dos documentos e extração de informações relevantes;
- **Sumarização**: Agrupamento dos documentos por semelhança, criação de categorias para classificação dos resultados, sistematização das informações

de interesse em imagens e quadros-resumo com resultados quantitativos e qualitativos.

É possível afirmar que a Revisão Sistemática da Literatura possui, portanto, três princípios basilares, sendo eles: planejamento, registro e compartilhamento. Tanto o planejamento como o registro se materializam no preenchimento do protocolo de pesquisa, que irá orientar todas as etapas apresentadas.

Também existe a preocupação com o registro das decisões tomadas pelo pesquisador ao longo do processo de seleção e extração das informações. O compartilhamento do protocolo e dos dados de pesquisa é o que permite que todas as decisões possam ser auditadas e que a pesquisa possa ser atualizada/continuada, ou que outros recortes possam ser feitos com base no mesmo *corpus* teórico. Nesse sentido, o plano de gestão de dados de pesquisa da presente tese prevê, ao final, o compartilhamento dos dados de pesquisa, onde serão disponibilizados os dados do processo de seleção dos documentos, da extração das informações pertinentes e da sumarização dos resultados.

Com base nos protocolos propostos por Kitchenham *et al* (2004), felizardo *et al.* (2015), Felizardo *et al.* (2015) e no preenchimento do protocolo proposto por Jesus, (2021), elaborou-se um protocolo de pesquisa para orientar a condução da RSL. O quadro 1 apresenta o preenchimento do protocolo de pesquisa que orientou a condução da RSL.

Quadro 1 - Protocolo de pesquisa da Revisão Sistemática - Qualidade de dados *Linked Data*

Protocolo de pesquisa	
Pergunta de pesquisa (principal)	Como tem sido abordada a qualidade de dados no contexto do *Linked Data*?
Objetivos	Identificar as principais abordagens da qualidade de dados Linked Data, identificando problemas, ferramentas e metodologias que possam ser utilizadas no processo de avaliação de qualidade desses dados.
Estratégia de busca	*("Linked Data" OR "Linked Open Data") AND ("Data Quality")*
Bases de dados consultadas	*1ª rodada Web of Science 2º rodada LISTA; 3° rodada BRAPCI. Atualização: Wos*
Período abrangido	Sem restrição temporal
Idiomas	Português, inglês e espanhol.
Tipos de documentos	Artigos publicados em periódicos científicos, artigos de revisão e artigos apresentados em conferência, desde que revisados por pares e indexados nas bases de dados consultadas.
Critérios de Inclusão	(I) Foco principal na discussão de qualidade de dados publicados de acordo com os princípios do *Linked*

	Data (I) Foco em apresentar um artefato para avaliação de qualidade de dados
Critérios de exclusão	(E) Não está nos idiomas estabelecidos para a pesquisa; (E) Apenas menciona a temática de interesse; (E) Não aborda a temática de interesse; (E) Não foi possível obter acesso ao documento completo;
Estratégia de seleção dos estudos	Na primeira rodada de seleção foram considerados os títulos, resumos e palavras-chave dos documentos. Na segunda rodada, realizada durante a fase de extração, foi considerada a leitura completa e crítica dos documentos.
Formulário de extração	1) Enfoque do documento 2) Desafios e problemáticas 3) Procedimentos e etapas 4) Definições relacionadas a qualidade de dados *Linked Data* 5) Ferramentas de avaliação e Informações a respeito das aplicações realizadas
Critérios para avaliação de qualidade	Foram considerados apenas artigos de periódicos, sendo considerados como critérios o processo de seleção por pares e os critérios de qualidade dos próprios periódicos. Também foram avaliados, em caráter de discussão e não de exclusão, a profundidade dos artigos em relação a temática abordada e a aderência e completude dos procedimentos.
Forma de análise dos resultados	A análise dos resultados foi realizada de acordo com cada um dos itens do formulário de coleta: 1) Agrupamento em quadros resumo, categorias estabelecidas a *posteriori*, visando compreender como a temática é explorada; 2) Agrupamento em quadro resumo, com a criação de três categorias, elaboradas a *posteriori* (1. desafios relacionados às características das fontes de dados 2. desafios relacionados com a estrutura dos dados *Linked Data* e 3.) desafios relacionados ao processo de avaliação de qualidade; 3) Sumarização das principais dimensões relacionadas a qualidade de dados *Linked Data*, visando à aplicação de estudo terminológico e a composição do fluxo de seleção de dados *Linked Data;* 4) Os conceitos identificados foram objeto do estudo terminológico, sendo organizados em árvore de domínio. As definições para os conceitos foram abordadas através do método da Grade, visando a elaboração de definições para o glossário de qualidade de dados *Linked Data*; 5) Identificada a complexidade das ferramentas/metodologias disponíveis foi elaborado um formulário para a coleta de informações relevantes composto pelos campos: Nome do artefato; Contexto de sua criação (com histórico, responsáveis e motivação que embasam a sua necessidade de criação); Tipo de artefato; Atividade que realiza; Forma como desempenha essa atividade; Domínio ao qual é direcionado; Público a que se destina; Categorias, dimensões e critérios englobados pelo artefato; Status; Prova de

	conceito; Possibilidade de customização (considerando inclusive se o artefato é aberto, gratuito ou livre). Os artefatos foram então classificados com base em seus objetivos, nas atividades que realizam e na forma como realizam essas atividades.

Fonte: Autora (2025)

As buscas na BRAPCI foram realizadas por meio de busca manual, pelo termo “qualidade de dados”, seguida da aplicação dos critérios de exclusão. Os documentos recuperados foram sistematizados em formato de planilha, onde foram identificadas as duplicatas e aplicados os critérios de exclusão.

Após a seleção, realizou-se a leitura dos documentos aceitos, agrupando-os em categorias temáticas construídas a *posteriori*, de acordo com a identificação de padrões nas temáticas dos documentos. A coleta dos enfoques temáticos dos documentos foi realizada através de uma análise de seus objetivos, metodologias e resultados obtidos.

A coleta e a sistematização de informações sobre os artefatos não tiveram o intuito de puramente listar os artefatos existentes, ou de promover uma análise individual exaustiva desses artefatos. Buscou-se em um primeiro momento, a identificação dos tipos de artefatos existentes, das suas possibilidades e limitações para auxiliar no processo de seleção de dados *Linked Data*.

Também foram coletados e analisados os problemas de qualidade específicos de dados publicados como *Linked Data*. Ao longo da leitura flutuante dos artigos aceitos, foram coletados os problemas de qualidade mencionados pelos autores, em seguida foram construídas categorias a *posteriori* para a sistematização dessas problemáticas. As buscas nas bases de dados foram realizadas entre dezembro de 2021 e maio de 2022, com atualização em janeiro de 2025, utilizando filtro temporal.

Apresentados os procedimentos adotados para a condução da RSL, a próxima subseção apresenta os procedimentos do estudo documental.

2.3 Estudo documental

O estudo documental foi realizado com base nos documentos publicados pelo W3C, com o objetivo de promover a identificação e análise de documentos que abordem a questão da qualidade de dados publicados como *Linked Data*.

As buscas foram realizadas no portal do W3C em inglês, utilizando como palavras-chave relacionadas à qualidade de dados como “data quality”, à estrutura dos dados *Linked Data*, como “Resource Description Framework” e “Uniform Resource Identifier”.

Também foram considerados termos relacionados a instrumentos, recomendações e melhores práticas que influenciam na qualidade de dados *Linked Data,* como “*Web Best Practices*”, “*Linked Data Best Practices,*” e “*Data Quality Vocabulary*”.

Foram considerados documentos como melhores práticas, recomendações, documentos oficiais de formatos e vocabulários. Também foram consideradas as referências mencionadas pelos documentos recuperados através da estratégia de busca, utilizando a técnica de *Snowballing*.

Além dos documentos que tratam de qualidade de dados, também fizeram parte do estudo a documentação a respeito do *Linked Data*, que abordam sua estrutura e processo de adoção. O *corpus* documental foi sistematizado em um quadro-resumo, apresentando uma breve síntese de cada documento.

2.4 Estudo terminológico

O estudo terminológico da presente tese foi relacionado aos domínios de qualidade de dados e qualidade de dados *Linked Data*. O estudo possui uma abordagem sistemática, concentrando-se em um domínio ou área específica do conhecimento (Catalá; Barité, 2016).

Esse aspecto da pesquisa foi conduzido com base nas seguintes etapas: 1) análise do *corpus* teórico-documental selecionado com base nas etapas anteriormente descritas; 2) seleção dos termos; 3) organização dos termos em árvore de domínio; 4) elaboração das definições; 5) estruturação do glossário.

As Árvores de Domínio são ferramentas que, a partir da estruturação hierárquica da relação entre termos “[...] delimitam o território e as fronteiras do domínio ou campo do conhecimento, pois estabelecem e fixam o espaço temático que será estudado” (Catalá; Barité, 2016, p. 95, tradução nossa), e permitem ainda, “[...] enquadrar cada termo em algumas de suas filiais ou subáreas e assim garantem tanto a existência do termo como sua adesão ao domínio (Catalá; Barité, 2017, p. 95, tradução nossa).

A árvore foi elaborada na forma de diagrama hierárquico, os termos foram identificados a partir do *corpus* criado com base nos procedimentos apresentados. Após o estabelecimento da Árvore de Domínio da qualidade de dados *Linked Data*, a próxima etapa é o estabelecimento das definições de cada termo, utilizando-se o método da Grade.

A estrutura para a condução do método da Grade foi apresentada por Barité (2011). Atualmente, o método é amplamente utilizado nas pesquisas desenvolvidas pelo Grupo *Terminología y Organización del Conocimiento (GTERM)*.[1]. O método consiste na eleição de fontes que apresentem definições para o termo em estudo, com base a uma ou mais garantias.

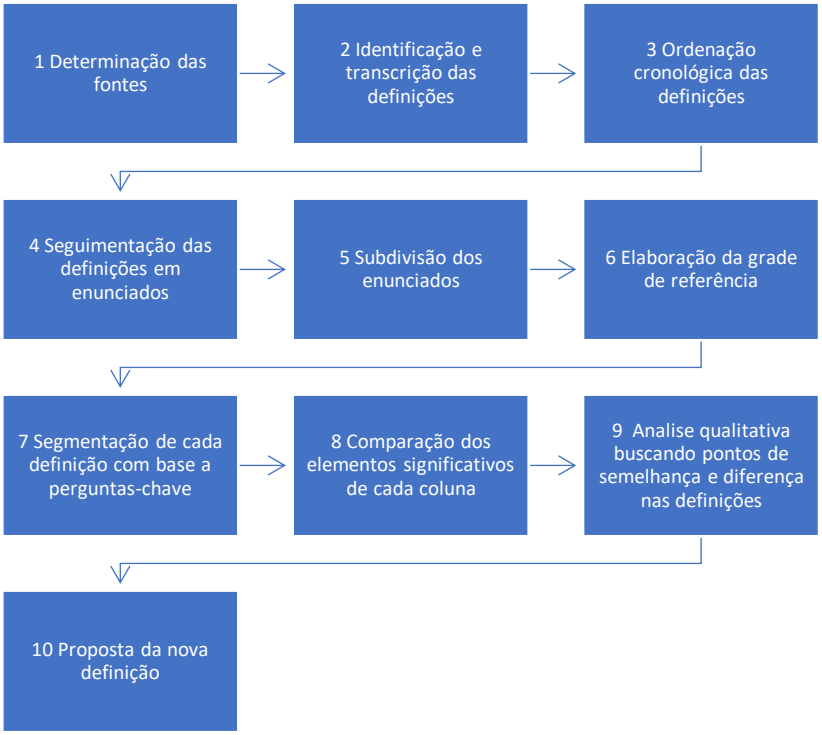
> Estas definições serão segmentadas em cinco tipos de enunciados predicativos (essenciais, acidentais, informativos, históricos e relacionais). Os segmentos dos enunciados essenciais e acidentais, bem como alguns enunciados de outra ordem, serão localizados em uma grade, para a realização de uma análise comparativa dos mesmos. Da destilação da informação colocada na grade, será elaborada uma definição que combina os aspectos teóricos, pragmáticos e metodológicos da garantia literária (Barité, 2011, p. 27-28, tradução nossa).

O objetivo é a elaboração de uma definição completa, clara e representativa, seguindo uma estrutura baseada em enunciados essenciais, acidentais, informativos e históricos que atuem como uma definição para um termo em um determinado domínio, com base na análise do discurso da própria comunidade desse domínio. “Ao recuperar o conjunto de declarações relevantes das definições, o resultado final é necessariamente mais abrangente e exaustivo do que qualquer uma das definições anteriores, e por isso deve superar todas elas.” (Barité, 2011, p. 238, tradução nossa).

A figura 2 apresenta as diferentes etapas da aplicação do método da Grade na elaboração de definições:

[1] “Com uma história de mais de duas décadas, o GTERM trata de pesquisas em dois temas centrais: Terminologia e Organização do Conhecimento. Nestas duas áreas, o GTERM é o único grupo uruguaio de diálogo com pesquisadores do exterior, especialmente da América e da Europa.Participa regularmente nos campos acadêmicos da International Society for Knowledge Organization-ISKO e da Rede Ibero-Americana de Terminologia; bem como nas Conferências de Pesquisa da FIC e em diversos eventos acadêmicos em Udelar e na região.” Disponível em: https://fic.edu.uy/grupo/terminologia-y-organizacion-del-conocimiento-gterm

Figura 2 - Processo da criação e definições usando o método da Grade



Fonte: adaptado de Barité (2011)

A definição das fontes foi feita com base no *corpus* construído, enquanto as definições para cada um dos termos identificados na Árvore de Domínio foram organizadas em quadros-resumo, ordenadas cronologicamente.

A segmentação a que se refere a etapa 4 segue a classificação de Barité e Rauch (2006), sendo elas:

- **Essenciais** – atributos que necessariamente participam da definição de um objeto;
- **Acidentais** – atributos que somente estão presentes em algumas instancias de um objeto, mas que podem ser significativos;
- **Informativos** – aqueles que não se referem aos atributos de um objeto definido, mas sim a dados ou informações que pretendem ilustrar aos leitores sobre aspectos relacionados com o objeto em seu contexto ou na realidade;
- **Históricos** – enunciados que permitem uma análise retrospectiva do objeto, destacando marcos temporais e as alterações sofridas pelo objeto ao longo do tempo.

Além disso, também foi observado nas definições como os autores caracterizam o termo (gênero) e a existência de uma menção explícita do contexto no qual foi cunhada a definição.

Os quadros foram organizados com base em dois processos: 1) coleta de termos utilizados e definições relevantes; 2) aplicação do método da Grade.

2.5 Construção do Fluxo

A construção do fluxo para a seleção de dados *Linked Data* foi baseada na aplicação do *Design Science Research* (DSR), método voltado para a resolução de problemas por meio da elaboração de artefatos e da geração e compartilhamento de conhecimento através da experiência adquirida durante a elaboração do artefato.

O método tem o objetivo de promover um maior conhecimento/compreensão do problema em questão, assim como da solução para esse problema, por meio da elaboração de um artefato (Hevner *et al.*, 2004). De acordo com Bax (2015, p. 201):

> A DSR envolve construir, investigar, validar e avaliar artefatos, tais como construtos, arcabouços, modelos, métodos e instâncias de sistema de informação a fim de resolver novos problemas práticos. Além disso, o estudo de métodos, comportamentos e melhores práticas relacionadas com a análise do problema e com o processo de desenvolvimento de artefato são abrangidos.

A elaboração do artefato promove, portanto, maior clareza sobre o domínio do problema em questão, ao mesmo tempo que permitirá testar e confirmar a viabilidade de potenciais soluções.

> O artefato projetado, que deve ser criado de maneira sistemática, por meio da explicitação do problema, do registro dos procedimentos e de constante refinamento até que seja considerado adequado para solução do problema proposto. O processo de elaboração deve ser registrado e os resultados precisam ser formalmente comunicados (Jesus *et al*, 2023, p. 2).

Peffers *et al.* (2007) dividem a condução do DSR em 6 etapas, sendo elas: 1) Identificação do problema e descrição da motivação; 2) Estabelecimento dos objetivos para a solução do problema; 3) Desenvolvimento do artefato; 4) Demonstração, ou aplicação do artefato; 5) Avaliação da eficiência e efetividade do artefato; e 6) Comunicação dos resultados tanto para a sociedade de maneira geral como para o público especializado. O quadro 2 apresenta as informações relevantes para cada uma dessas etapas:

Quadro 2 - Protocolo da condução do DSR

Protocolo de condução do DSR	
Identificação do problema e descrição da motivação	A seleção de conjuntos de dados *Linked Data* é uma atividade complexa, que depende da avaliação de qualidade desses dados. O processo de avaliação de qualidade depende da escolha da perspectiva, da seleção de critérios e métricas, do conhecimento acerca dos diferentes artefatos e metodologias que podem auxiliar na realização desse processo. Soma-se a essa complexidade o necessário conhecimento das características intrínsecas aos dados publicados como *Linked Data*. Embora a literatura científica apresente diversas soluções para a avaliação de qualidade de dados *Linked Data*, a construção de um fluxo para seleção de fontes de dados *Linked Data* justifica-se pela identificação da necessidade de sistematização acerca de todos esses conceitos, visando tornar o processo mais claro e aplicável, tanto para os pesquisadores do tema quanto para a comunidade prática.
Estabelecimento dos objetivos para o artefato	Criação de um fluxo que auxilie no processo de seleção de dados *Linked Data*, identificando os processos, as atividades e etapas para a realização da seleção. Também espera-se promover a diferenciação entre tipos de artefatos que podem auxiliar na condução do processo em questão.
Desenvolvimento do artefato	A elaboração do fluxo foi baseada em: 1) Estado da arte da qualidade de dados *Linked Data* - A análise dos documentos visando identificar como ocorre o processo de avaliação, quais as etapas necessárias e quais os possíveis artefatos a serem empregados nesse processo. 2) Análise da documentação do W3C, que permite identificar os aspectos necessários para a análise intrínseca da qualidade de dados *Linked Data* e ainda a identificação de artefatos para auxiliar nesse processo. 3) Estudo terminológico, que permite a identificação dos termos-chave para a construção do fluxo e a apresentação das definições necessárias para a compreensão do mesmo.
Demonstração, ou aplicação do artefato	A avaliação foi descritiva e argumentativa com base na descrição dos potenciais cenários de uso do artefato e argumentar sua utilidade com base na literatura do *corpus*
Avaliação da eficiência e efetividade do artefato	Foi discutida a relevância do fluxo proposto com base na discussão de cada uma de suas etapas e da aplicabilidade dessas etapas na seleção de dados *Linked Data*.
Comunicação dos resultados tanto para a sociedade	Os resultados foram comunicados por meio da tese e de artigos científicos/trabalhos em eventos derivados.

de maneira geral como para o público especializado.	

Fonte: Baseado em Peffers *et al*. (2007)

Apresentados os procedimentos metodológicos adotados na pesquisa, a próxima seção irá apresentar a discussão a respeito do termo qualidade de dados e sua relação com a Ciência da Informação.

3 A QUALIDADE DE DADOS E A SUA RELAÇÃO COM A CIÊNCIA DA INFORMAÇÃO

Essa seção foi construída com o intuito de apresentar uma breve contextualização a respeito da qualidade de dados e de sua relação com a Ciência da Informação.

A subseção 3.1 parte de uma breve contextualização a respeito da importância dos dados para sociedade e de seu papel enquanto objeto de estudo na Ciência da Informação. Em seguida, serão apresentados termos importantes para a compreensão da qualidade de dados, como dimensões, critérios e métricas.

A subseção 3.2 apresenta a qualidade de dados enquanto objeto de estudo da Ciência da Informação, tendo como base uma análise exploratória das abordagens da temática em âmbito nacional.

3.1 Dados e qualidade de dados: breve contextualização

Os dados vêm obtendo destaque em diversos setores da sociedade, eles "[...]desempenham um papel crucial na sociedade de tecnologia da informação e comunicação (TIC): eles são gerenciados por aplicativos empresariais e governamentais e são fundamentais em todos os relacionamentos entre governos, empresas e cidadãos." (Batini; Scannapieco, 2016, p. 6, tradução nossa).

Os dados também se tornaram relevantes para a comunidade cientifica, que vêm reconhecendo a importância dos conjuntos de dados para o desenvolvimento e o avanço da ciência. Um exemplo desse reconhecimento no Brasil é a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) que:

> [...] reconhece a importância da gestão adequada dos dados de pesquisa como parte essencial das boas práticas de pesquisa. Para tanto, considera necessário que os dados resultantes de projetos financiados pela Fundação sejam gerenciados e compartilhados de forma a garantir o maior benefício possível para o avanço científico, tecnológico, socioeconômico e cultural (FAPESP, 2024, não paginado).

Com os avanços tecnológicos recentes, especialmente em campos como análise de dados, inteligência artificial e *business intelligence*, o conhecimento do potencial valor dos dados, antes concentrado no público especializado, nas

universidades e nos ambientes empresariais, passou a ser de conhecimento mais amplo.

Um dos exemplos do reconhecimento social da importância e do valor de conjuntos de dados no Brasil é o estabelecimento da Lei 13.709/2018, Lei Geral de Proteção de Dados Pessoais (LGPD) que dispõe sobre o:

> [...] tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural (Brasil, 2018, não paginado).

Nesse cenário, onde os dados têm impacto no desenvolvimento das pessoas, dos governos, das empresas e da ciência, a qualidade dos dados tem o potencial de influenciar no desenvolvimento de diversos setores da sociedade.

As discussões a respeito dos dados enquanto objetos de estudo da Ciência da Informação tem se ampliado nos últimos anos. Essa relação é discutida por Santos e Sant'ana (2013, p. 208):

> O conceito de dados pode gerar avanços significativos para a área de ciência da informação, pois representa um problema de grande impacto e abrangência para a atividade de pesquisa, reforçando os pressupostos da ciência da informação. Nessa perspectiva, o tratamento da informação, a representação de recursos, a recuperação e a disseminação de informação se tornam áreas vinculadas à descrição, ao armazenamento, à preservação, ao acesso e à gestão de dados.

Um exemplo da relevância dos dados na Ciência da Informação nacional é a alteração do nome e da ementa de um dos Grupos de Trabalho (GT) da Associação de Pesquisa e Pós-graduação em Ciência da Informação (ANCIB). O GT8, antes denominado Informação e Tecnologia, passou a ser denominado Dados, Informação e Tecnologia, e teve a sua ementa alterada para a inclusão dos dados:

> Estudos e pesquisas teórico-práticos sobre e para o desenvolvimento de tecnologias de informação e comunicação que envolvam os processos de coleta, geração, representação, armazenamento, recuperação, disseminação, uso, reuso, gestão, análise, processamento, tratamento, governança, visualização, segurança e preservação **de dados e informação** em ambientes informacionais. (ANCIB, 2024, não paginado, grifo nosso).

A qualidade de dados ainda é um objeto de pesquisa recente para a Ciência da Informação, mas a preocupação em desenvolver métodos e ferramentas para avaliar e melhorar a qualidade dos dados em diversos contextos começa a ser discutida na década de 1960, se estabelecendo na década de 1970.

Nesse período começa a se formar um novo domínio, com uma comunidade interdisciplinar que se reúne em torno da busca por formas de identificar e corrigir problemas de qualidade em conjuntos de dados. (Batini; Scannapieco, 2006).

Uma das abordagens mais citadas a respeito da qualidade de dados, influenciando a estrutura e o desenvolvimento de pesquisas atuais, é a sistematização do processo de avaliação de qualidade proposta por Wang e Strong (1996).

A importância dessa sistematização é reafirmada pela própria literatura da qualidade de dados. Sadiq (2013) partindo de uma ampla análise das publicações relacionadas a qualidade realizou uma análise dos autores mais citados nesses trabalhos. Dessa análise destaca-se o trabalho de Wang, que contava em 2013 com 4,364 citações e de Strong com 1986 citações.

Os autores atribuem esse número de citações aos aspectos estruturais da obra de Wang e Strong para a área da qualidade de dados:

> Algumas das primeiras contribuições vieram de Wang, R. Y., Strong, D. e associados na identificação de Dimensões de Qualidade de Dados e Avaliação de Qualidade de Dados. Essas contribuições foram amplamente utilizadas por pesquisadores posteriores, como é evidente pela alta contagem de citações acima" (Sadiq, 2013, p. 7, tradução nossa).

Batini e Scannapieco (2016, p. 9) também ressaltam a importância das contribuições de Wang: "Wang *et al* fornecem uma perspectiva geral sobre a qualidade dos dados, fornecendo uma coleção heterogênea de contribuições de diferentes projetos e grupos de pesquisa."

Wang e Strong (1996) apontam que no final dos anos 1990 a qualidade de dados já era uma preocupação das empresas, que buscavam formas de melhorar a qualidade dos seus dados. Segundo os autores, os investimentos de qualidade concentravam-se em apenas um dos muitos aspectos da qualidade, a acurácia. Entretanto, para os consumidores de dados, qualidade assumia uma definição muito mais ampla, que contempla uma série de outros aspectos.

Partindo desse entendimento os autores realizaram um estudo abrangente, visando compreender e estruturar a qualidade de dados a partir do seu significado para os consumidores de dados.

> Embora as empresas estejam melhorando a qualidade dos dados com abordagens e ferramentas práticas, seus esforços de melhoria tendem a se concentrar estritamente na acurácia. Acreditamos que os consumidores de dados têm uma conceituação de qualidade de dados muito mais ampla do que os profissionais de Sistemas de Informação percebem. O objetivo deste artigo é desenvolver uma estrutura que capture os aspectos da qualidade dos dados que são importantes para os consumidores de dados. (Wang; Strong, 1996, p. 5, tradução nossa).

A primeira etapa do estudo partiu de uma análise da literatura de qualidade de dados, visando a criação de categorias que permitissem agrupar por semelhança os diferentes aspectos da qualidade de dados. Essas categorias foram elaboradas para auxiliar na condução do estudo, mas depois passaram por um processo de refinamento “Como resultado desse reexame, renomeamos duas das quatro categorias. As categorias resultantes, portanto, são: DQ intrínseca, DQ contextual, DQ representacional, DQ acessibilidade”. (Wang; Strong, 1996, p. 19, tradução nossa).

Wang e Strog (1996, p. 22, tradução nossa) explicam ainda cada uma dessas categorias:

> DQ intrínseco denota que os dados têm qualidade por si só. DQ contextual destaca o requisito de que a qualidade dos dados deve ser considerada dentro do contexto da tarefa em questão. DQ representacional e DQ de acessibilidade enfatizam a importância do papel dos sistemas.

A qualidade, em uma perspectiva intrínseca, estaria relacionada a características inerentes a esses dados, com uma abordagem mais genérica, de uma maneira independente do domínio e do contexto de futura aplicação. Nessa perspectiva um conjunto de dados pode ser considerado de qualidade quando está livre de erros estruturais e de más formações.

A perspectiva contextual deriva da acepção de qualidade relacionada ao conceito de “*fitness for use*”, onde a qualidade dos dados está relacionada com os objetivos e com as tarefas que o usuário pretende executar com esses dados. Nessa perspectiva, um conjunto de dados pode ser perfeitamente adequado para

determinada tarefa e não atuarem satisfatoriamente em outro contexto (Juran, 1988; Wang; Strong, 1996; Zaveri *et al.*, 2012).

Em uma perspectiva contextual, discutir, avaliar, mensurar e melhorar a qualidade de dados são ações que dependem das características do domínio e das necessidades do contexto em que serão empregados esses dados. Um conjunto de dados é considerado de qualidade quando atende satisfatoriamente às necessidades do domínio ou de seu contexto de aplicação

Nelson, Todd e Wixom (2005) apontam que em uma perspectiva representacional, busca-se avaliar se os dados estão bem estruturados, estão livres de redundância e são de fácil compreensão.

A perspectiva de acessibilidade está relacionada à capacidade dos usuários de obterem acesso aos dados de maneira eficaz e eficiente. Nessa perspectiva um dado é considerado de qualidade quando pode ser localizado e obtido quando necessário.

As categorias são compostas por dimensões, que agrupam problemas semelhantes de qualidade que serão observados nos dados para permitir mensurar a sua qualidade. As dimensões são compostas por critérios que descrevem o atributo específico de qualidade que será avaliado (Wang; Strong, 1996).

A escolha das dimensões e dos critérios que serão considerados depende tanto da perspectiva adotada como do domínio ao qual os dados pertencem. Para avaliar se um conjunto de dados possuem ou não boa qualidade, são necessários indicadores, nesse contexto os indicadores são denominados como métricas. Cada critério pode ser mensurado a partir de mais de uma métrica, que são elaboradas com base nas perspectivas adotadas (Assaf; Senart; Troncy, 2016; Melo, 2017).

A organização em categorias e dimensões faz com que a qualidade possa ser abordada por diferentes perspectivas, que afetam a forma como se realiza o processo de avaliação de qualidade e como se identifica se um conjunto de dados é ou não de boa qualidade.

A preocupação com a qualidade e confiabilidade dos dados torna-se relevante para a Ciência da Informação, seja por meio da aplicação de técnicas que garantam a boa qualidade dos dados derivados de seus processos, seja por meio do estudo dos impactos de problemas de qualidade de dados nos processos informativos. A próxima seção discute a qualidade de dados como um objeto de estudo da Ciência da Informação.

3.2 A qualidade de dados como objeto de estudo da CI

As discussões formais a respeito de qualidade de dados na Ciência da Informação ainda são recentes, especialmente nacionalmente. Entretanto, a busca por garantir a qualidade dos conjuntos de dados perpassam diversos aspectos da área:

> [...] embora ainda não haja um consenso sobre o conceito de qualidade, observa-se na Ciência da Informação diversas proposições de critérios ou atributos para qualificar os dados e as informações. Tais critérios são usados para definir, medir e gerenciar a qualidade e variam de acordo com as abordagens e vertentes sob as quais os estudos são realizados (Fagundes; Macedo; Freund, 2017, p. 197).

A Ciência da Informação possui, desde sua estruturação, a expertise no desenvolvimento de instrumentos que visam padronizar e estruturar conjuntos de dados, tais como vocabulários, códigos de catalogação e padrões de metadados. Além disso, dados são o produto de processos importantes da área, especialmente no âmbito da representação da informação e da organização do conhecimento.

Em outro aspecto, se consolidam como objetos da área o estudo de fenômenos derivados de mudanças tecnológicas, e amplamente relacionados a produção massiva de dados, tais como *Big Data*, *Linked Data*, Inteligência Artificial, *Machine Learning*, Internet das Coisas entre outros temas.

Buscou-se então compreender como tem sido abordada a qualidade de dados na Ciência da Informação, por meio da análise da literatura nacional e da aplicação de técnicas de *Snowballing*.

Os documentos foram analisados de acordo com o seu tipo documental, sendo 12 artigos, 8 trabalhos apresentados em eventos e 3 dissertações[2]. Nenhuma tese com o objetivo de discutir qualidade de dados foi recuperada na base consultada com o filtro para Ciência da Informação.

[2] Os textos "A descrição formal da qualidade de dados publicados na *Web*: análise do *Data Quality Vocabulary* (DQV)"(Jesus; Santarem Segundo, 2023); "Qualidade de dados *Linked Data*: análise da temática sob a perspectiva da Ciência da Informação (Jesus; Santarem Segundo, 2023); e "A questão da qualidade em dados publicados como *Linked Data*: um mapeamento sistemático da literatura (Jesus; Santarem Segundo, 2022) não foram considerados por serem de autoria da autora e de seu orientador, construídos no âmbito da presente tese.

Buscou-se então compreender como tem sido abordada a qualidade de dados na Ciência da Informação, por meio da análise da literatura nacional e da aplicação de técnicas de *Snowballing*.

Com o objetivo de compreender como a qualidade de dados tem sido abordada na literatura enquanto um objeto de estudo, buscou-se analisar os objetivos dos estudos recuperados. O quadro 3 apresenta os documentos selecionados para compor o *corpus* e uma breve descrição dos seus objetivos.

Quadro 3 - Documentos selecionados para compor o *corpus*

AUTORES	TÍTULO	DESCRIÇÃO
Bentancourt e Rocha (2012)	Metadados de qualidade e visibilidade na comunicação científica.	Realiza o processo de avaliação de qualidade nos dados de uma revista de acesso aberto. O procedimento de escolha das dimensões a serem observadas foi embasado no modelo de Qualidade de Dados ISO/IEC 25012 e nas especificações, recomendações e melhores práticas do *Dublin Core*.
Almeida *et al.* (2016)	Melhoria na qualidade de dados com a aplicação de "*data cleaning*" na base de dados de acidentes aeronáuticos da aviação civil brasileira	Realiza o processo de avaliação de qualidade com o objetivo de mensurar a melhoria na qualidade dos dados após aplicação de um processo de limpeza de dados, em uma base de dados de acidentes aeronáuticos. Foi elaborada uma fórmula para comparar a qualidade antes e depois do processo, com foco na identificação de anomalias.
Fagundes; Macedo; Freund (2017)	A produção científica sobre qualidade de dados em *Big Data*: um estudo na base de dados *Web of Science*	Discute a reação entre *Big Data* e qualidade de dados. apontando que o tema da qualidade de dados em ambientes *Big Data* ainda é recente e pouco explorado, sendo as soluções encontradas genéricas e sem abordagem contextual profunda.
Melo (2017)	Metodologia de avaliação de qualidade de dados no contexto do *Linked Data*.	Discutem a criação de um modelo e de metodologia para a avaliação de qualidade de dados publicados como *Linked Data.*
Melo; Botega; Santarem Segundo (2017a)	Metodologia de Avaliação de Qualidade para Dados Conectados	Discutem a criação de um modelo e de metodologia para a avaliação de qualidade de dados publicados como *Linked Data,* o modelo apresenta as dimensões e métricas a serem utilizados. O modelo e a metodologia propostos foram baseados tanto nos princípios para

		a publicação de dados do W3C como em uma revisão de literatura sobre qualidade em dados *Linked Data*.
Melo; Botega; Santarem Segundo (2017b)	Metodologia de Avaliação de Qualidade para Dados Conectados	É uma publicação expandida do trabalho apresentado no evento (Melo; Botega; Santarem Segundo, 2017a).
Silva *et al.* (2017)	Desenvolvimento de Ontologia Ciente de Qualidade de Informações para a Melhoria de Consciência Situacional no Domínio de Gerenciamento de Emergências	Propõe uma ontologia de domínio para o gerenciamento de emergências, mais especificamente sobre incêndios florestais, visando contribuir para a representação de informações e com processos de avaliação de situações de fogo, considerando requisitos de qualidade necessários para esse tipo de dados.
Espíndola *et al.* (2018)	Governança de dados aplicada à Ciência da Informação: análise de um sistema de dados científicos para a área da saúde.	Busca identificar possibilidades de melhorias de qualidade em um conjunto de dados de um sistema de avaliação motora para crianças e idosos. O procedimento de avaliação conduzido teve como base ciclo de vida de dados e ferramentas de governança de dados.
Moura Junior e Aragão (2018)	Metas, ações e indicadores como subsídios para análise da qualidade de dados	Realiza a análise da qualidade dos dados contidos nos repositórios institucionais da Universidade Federal da Paraíba (UFPB), com o objetivo de aplicarem esses dados nos indicadores de desempenho institucionais. Como base para a realização do processo de avaliação foi utilizado o Plano de Desenvolvimento Institucional (PDI) vigente da UFPB, depois foram identificados os dados necessários ao monitoramento de cada meta e indicador de desempenho e verificada a qualidade desses dados.
Piccolo (2018)	Qualidade de dados dos sistemas de informação do DATASUS: análise crítica da literatura	Realiza um levantamento bibliográfico sobre a questões de qualidade em sistemas de informação do DATASUS. Conclui que ainda são necessários mais estudos que tenham por objetivo avaliar a qualidade dos dados desse sistema.
Juliani *et al.* (2019)	Governança de dados aplicada no processo de catalogação. Revista	Busca identificar possibilidades de melhorias no acervo do catálogo de uma instituição de ensino pública.

	Brasileira de Biblioteconomia e Documentação	Realiza uma análise qualitativa dos dados, com amostragem gerada aleatoriamente, utilizando como base ferramentas de governança de dados, avaliando a existência de duplicação em dados de autoridade de pessoas e cabeçalhos de títulos.
Moreira *et al.* (2020)	A qualidade na recuperação de dados governamentais: um estudo sobre dados de políticas públicas na internet.	Realiza o processo de avaliação de qualidade dos dados do Programa Nacional de Fortalecimento da Agricultura Familiar, disponíveis no portal do Banco Central. Para estabelecer as dimensões a serem usadas os autores partiram de um levantamento bibliográfico e utilizaram o método de Análise de Conteúdo.
Macedo (2021)	Dados abertos governamentais: modelo de governança voltado a qualidade de dados para publicação em rede	Propõe um modelo de governança para a qualidade de dados governamentais publicados em ambiente *Web*, focado em abrangência, consistência, eficácia e transparência.
Martins *et al.* (2021a)	Requisitos de qualidade para dados de agregação em museus: o caso IBRAM	Discute a criação de requisitos de qualidade de dados para viabilizar o processo de avaliação dos dados providos ao serviço de agregação de objetos digitais de museus do Instituto Brasileiro de Museus - IBRAM. A elaboração do modelo foi baseada em pesquisa bibliográfica e documental, utilizando a documentação de renomadas instituições agregadoras do patrimônio cultural.
Coelho Júnior e Lemos (2023a)	Qualidade de dados em acervos museais: uma avaliação semiautomática para os acervos sob gestão do IBRAM	Abordam a realização de um processo de avaliação de qualidade dos dados das bibliotecas vinculadas ao IBRAM. A avaliação realizada teve um caráter semiautomático e foi realizada com base em um alinhamento entre as normativas e o guia de catalogação adotado, foram identificados os elementos presentes no conjunto de dados e as regras específicas de uso e preenchimento desses elementos, permitindo a identificação de inconformidades.

Coelho Júnior e Lemos (2023b)	Tratamento da informação em acervos culturais: avaliação do uso de vocabulários controlados em coleções museológicas sob gestão do instituto brasileiro de museus	Abordam a realização de um processo de avaliação de qualidade dos dados das bibliotecas vinculadas ao IBRAM, tendo como a conformidade com os vocabulários controlados.
Martins *et al.* (2021b)	Requisitos de qualidade para dados de agregação em museus	Artigo expandido do trabalho publicado em evento (Martins *et al.*, 2021a).
Piccolo *et al.* (2021)	Qualidade de dados em gestão de dados de pesquisa	Realiza um estudo bibliométrico sobre qualidade de dados na gestão de dados de pesquisa.
Turi e Comarela (2022)	Impacto da adequação à Lei Geral de Proteção de Dados Pessoais na metrificação da qualidade de dados	Busca verificar o impacto das técnicas de anonimização propostas na Lei Geral de Proteção de Dados em uma amostra de diferentes plataformas de dados. O processo de avaliação de qualidade levou em consideração as dimensões e métricas estabelecidas na ISO/IEC 25012:2008.
Coelho Junior (2023)	Qualidade de dados em acervos do patrimônio cultural: uma proposta diagnóstica semiautomática para objetos culturais sob gestão do instituto brasileiro de museus	Desenvolve uma aplicação de avaliação diagnóstica semiautomática que permite a otimização da qualidade dos dados em acervos culturais
Dias et al. (2023)	*Garbage in*, *garbage out* (GIGO): enfrentando esta máxima nos conjuntos de dados associados ao programa dinheiro direto na escola (PDDE)	Discute a relevância da qualidade dos dados na Ciência de Dados, descrevendo o processo de extração, transformação e carga de dados para a geração de painéis de informação.
Lemos e Coelho Junior (2023)	Qualidade de dados em acervos do patrimônio cultural: uma avaliação diagnóstica semiautomática nos objetos culturais sob gestão do Instituto Brasileiro de Museus.	Discutem a elaboração de uma aplicação de avaliação diagnóstica semiautomática que permita a otimização da qualidade dos dados em acervos culturais. Esse é o único artefato identificado no levantamento que é voltado para a aplicação semiautomática.

Fonte: Autora (2025)

Analisando os objetivos e os resultados dos estudos selecionados, apresentado no quadro 3, foi possível categorizá-los em três categorias principais: documentos

focados na condução de processo de avaliação de qualidade; documentos que propõe/discutem um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data* e documentos que realizam uma abordagem teórica da qualidade de dados *Linked Data.*

Observa-se que a maioria dos estudos tem como foco a condução do processo de avaliação de conjuntos de dados. Em relação a esses documentos observa-se que os domínios aos quais pertencem os dados avaliados são heterogêneos, com dados governamentais, dados bibliográficos de instituições do patrimônio cultural entre outros.

Observa-se que os processos de avaliação foram conduzidos com propósitos distintos, tais como verificar os níveis de qualidade após determinados processos de limpeza de dados, agregação de dados, anonimização ou ainda identificar possibilidades de melhorias em conjuntos de dados.

Os procedimentos adotados também são heterogêneos, tendo abordagens manuais e automáticas, com ou sem uso de amostragem, utilizando ou não modelos de avaliação de qualidade para estabelecimento de dimensões e métricas.

Um aspecto em comum entre os estudos heterogêneos foi o uso da literatura como fonte para o estabelecimento de dimensões, e a busca pelo alinhamento à diretrizes, recomendações, melhores práticas e instrumentos de padronização do domínio a ser avaliado.

Em relação aos documentos que discutem a criação de artefatos de avaliação de qualidade, observa-se que embora vários estudos tenham discutido a criação de artefatos, muitos deles se referem a um único artefato, fazendo parte de diferentes estratégias de divulgação adotadas pelos autores.

Nesse sentido, observa-se no quadro 3 que ainda são poucos os artefatos elaborados no âmbito da CI brasileira para auxiliar na avaliação de qualidade de dados. Observa-se ainda que esses artefatos possuem forte relação com instrumentos de padronização, com o destaque para os modelos de qualidade de dados, instrumentos que orientam o processo de avaliação, estabelecendo dimensões e métricas para condução do processo.

Em relação a abordagem teórica de qualidade de dados, observou-se que as discussões também possuem uma abordagem heterogênea, sendo discutida a qualidade em cenários como Big Data e *Linked Data* e ainda em dados relacionados à área da Saúde.

Com base no *corpus* construído, observa-se que embora a temática venha aparecendo com mais constância, os estudos de qualidade de dados ainda são poucos e realizados por grupos específicos de pesquisadores. Observou-se também que muitos aspectos de qualidade podem ser explorados por pesquisadores brasileiros da CI.

Visando analisar ainda como a qualidade de dados se relaciona com outros termos no âmbito da CI, foram coletados os termos correlatos apontados pelos autores, a serem apresentados no quadro 4.

Para identificação, foram considerados: 1) termos utilizados nas definições de qualidade de dados apresentadas pelos autores; 2) termos utilizados para explicar a relação entre qualidade de dados e ciência da informação; 3) termos que descrevem os objetos de estudo dos processos de avaliação conduzidos ou das ferramentas de avaliação propostas pelos autores. O quadro 4 apresenta os autores e os termos identificados:

Quadro 4 - Termos correlatos identificados nos estudos aceitos

AUTORES	TERMOS CORRELATOS
Bentancourt e Rocha (2012)	qualidade de projeto; metadados; qualidade dos metadados; recuperação da informação; tratamento da informação; base de dados; política de metadados.
Almeida et al (2016)	anomalias de dados; *datacleaning*; limpeza de dados; dados de acidentes aeronáuticos; melhoria na qualidade dos dados; padrão na entrada de dados; validação na entrada de dados.
Fagundes; Macedo; Freund (2017)	governança de dados; dados de qualidade; modelo de ciclo de vida dos dados; "boas práticas de coleta, armazenamento e recuperação"; dados; diretrizes; princípios e boas práticas; preservação digital; dimensões de qualidade; *big data.*
Melo (2017)	gestão de dados; qualidade aplicada a dados; web semântica; *linked data*; *linked open data*; dados de qualidade; metodologia de avaliação de qualidade; modelo de qualidade de dados; dimensões de qualidade de dados; problemas de qualidade;
Melo; Botega; Santarem Segundo (2017a)	gestão de dados; qualidade aplicada a dados; web semântica; problemas de qualidade; avaliação de qualidade de dados; *linked data*; dados de qualidade; padrões; tecnologias; metodologia para avaliação de qualidade; dados conectados; modelo de qualidade; gestão de qualidade de dados; metodologia de avaliação de qualidade de dados.

Melo; Botega; Santarem Segundo (2017b)	curadoria de dados; padrões; tecnologias. gestão de qualidade de dados; metodologia de avaliação de qualidade de dados; web semântica; *linked data*; *linked open data*; qualidade aplicada a dados; modelo de avaliação de qualidade; métricas de qualidade; problemas de qualidade; dados conectados.
Silva et al (2017)	ontologia; ontologia de domínio; representação de informações; metodologia para avaliação da qualidade de dados; consciência da situação.
Espíndola et al. (2018)	controle da qualidade dos dados; qualidade dos dados; qualidade; governança de dados; dados de qualidade; modelo de ciclo de vida dos dados; "boas práticas de coleta, armazenamento e recuperação"; dados; diretrizes; princípios e boas práticas; preservação digital; processo de qualidade dos dados.
Moura Junior e Aragão (2018)	análise de qualidade de dados; qualidade dos dados; governança corporativa, governança de ti; indicadores de desempenho institucionais; plano de desenvolvimento institucional; qualidade dos dados; meta; indicador.
Piccolo (2018)	garantia de qualidade; garantia da qualidade dos dados; datasus
Juliani, *et al.* (2019)	qualidade dos dados; gestão da informação; governança de dados; boas práticas; catalogação; metadados; organização e representação da informação e do conhecimento; padrão de dados; linguagens documentárias (taxonomias); vocabulários; sistemas de organização do conhecimento; modelo de requisitos; organização e representação da informação.
Moreira et al. (2020)	recuperação de dados; ciclo de vida dos dados; políticas públicas; dados governamentais; dados qualificados; dados abertos; recuperação de dados.
Martins *et al.* (2021a)	modelo de qualidade de dados; modelo de requisitos; curadoria digital; padrões de qualidade em dados e metadados; representação da informação; catalogações descritivas; vocabulários controlados; regras de catalogação; política de qualidade de dados; objetos digitais; catalogação descritiva; catalogação de assunto, classificação; indexação; análise documental; metadados; padronização e descrição de recursos informacionais, padrões de catalogação; linguagem documentária; tesauro; modelo de requisitos; requisitos de qualidade de dados; padrão de metadados; modelo de metadados; representação e da gestão da informação, política de qualidade de dados; avaliação de qualidade de dados.
Martins *et al.* (2021b)	modelo de qualidade de dados; curadoria digital; política de qualidade de dados; avaliação de qualidade de dados; requisitos de qualidade de dados.

Piccolo et al. (2021)	curadoria de dados; gestão de dados de pesquisa; garantia de qualidade; dados de pesquisa; dimensões; métricas; curadoria digital; ciclo de vida dos dados.
Coelho Júnior e Lemos (2022a)	patrimônio cultural; indexação; sistemas de recuperação da informação; descrição e publicação de dados; acervos culturais; qualidade dos dados; metadados; catalogação; museologia; coleções digitais; nível da qualidade; dimensões de qualidade.
Coelho Júnior e Lemos (2022b)	avaliação diagnóstica de qualidade de dados; organização da informação; patrimônio cultural; dados com qualidade; padrões de documentação; metadados
Turi e Comarela (2022)	lei geral de proteção de dados pessoais; ciência de dados; metrificação da qualidade de dados; modelo de qualidade de dados; dimensões de qualidade de dados; métricas de qualidade dos dados.
Coelho Junior (2023)	patrimônio cultural; indexação; sistemas de recuperação da informação; descrição e publicação de dados; acervos culturais; qualidade dos dados; metadados; catalogação; museologia; coleções digitais; nível da qualidade; dimensões de qualidade.
Dias et. al (2023)	"*garbage in, garbage out*" (gigo); ciência de dados; painéis de informação; programa dinheiro direto na escola (pdde); tratamento dados.
Lemos e Coelho Junior (2022a)	avaliação diagnóstica de qualidade de dados; organização da informação; patrimônio cultural; dados com qualidade; padrões de documentação; metadados; museus.
Lemos e Coelho Junior (2022b)	patrimônio cultural; vocabulário controlado; guia de catalogação; práticas de catalogação; organização da informação; representação da informação; metadados; museus.

Fonte: Autora (2025)

Os termos passaram então por um processo de refinamento e categorização. Com base na análise foram estabelecidas 7 categorias de termos: 1) tipos de qualidade de dados; 2) sinônimos e termos alternativos para qualidade de dados; 3) termos relacionados a sistematização da qualidade de dados; 4) termos relacionados a procedimentos de qualidade de dados; 5) termos relacionados às áreas e subáreas da Ciência da Informação; 6) instrumentos utilizados no processo de avaliação de qualidade; e 7) termos relacionados aos temas e enfoques dos documentos.

A análise dos termos e de suas categorias permitiu observar que, embora forneçam *insights* sobre as abordagens de qualidade de dados na Ciência da Informação, as quatro primeiras categorias reúnem termos mais gerais e menos

relacionados com o entendimento de como a qualidade de dados é abordada enquanto um objeto de estudo da Ciência da Informação.

A categoria 1 reuniu os termos que especificam tipos de qualidade de dados, foi composta por apenas dois termos: qualidade de projeto e qualidade de metadados. Já a categoria 2 reúne os termos adotados pelos autores como sinônimos ou termos equivalentes ao conceito de qualidade de dados, tais como: dados de qualidade; qualidade dos dados; qualidade aplicada a dados; dados qualificados; e dados com qualidade. Esses termos refletem certa pluralidade terminológica entorno do termo qualidade de dados.

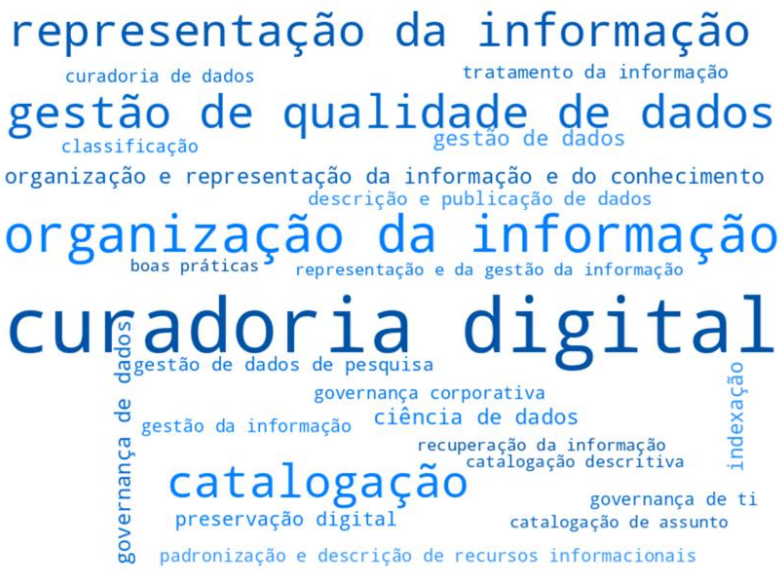
A categoria 3 reúne os termos relacionados a sistematização da qualidade de dados, tais como: dimensões (dimensões de qualidade; dimensões de qualidade de dado); métricas (métricas de qualidade; métricas de qualidade dos dados); problemas de qualidade; garantia de qualidade (garantia da qualidade dos dados); "*garbage in, garbage out*" (gigo); descrição da qualidade; nível da qualidade; critérios de qualidade; categorias de qualidade; questões de qualidade; anomalias de dados. Esses termos refletem a estruturação da qualidade de dados enquanto processo, seguindo a proposta por Wang e Strong (1996).

Ainda relacionados ao processo de avaliação de qualidade, foram identificados termos que refletem procedimentos de avaliação e melhoria na qualidade de dados, como: limpeza de dados; melhoria na qualidade dos dados; padrão na entrada de dados; validação na entrada de dados; avaliação de qualidade de dados; processo de qualidade dos dados; análise de qualidade de dados; avaliação diagnóstica de qualidade de dados; metrificação da qualidade de dados; tratamento dados; processo de avaliação de qualidade.

As últimas 3 categorias reúnem os termos relacionados com os enfoques dos estudos, permitindo entender em que contextos é discutida a qualidade de dados na CI; análise dos instrumentos utilizados pela comunidade no processo de avaliação de qualidade e ainda as áreas e subáreas da Ciência da Informação no âmbito das quais a qualidade de dados é discutida. Esses termos têm grande impacto no entendimento de como a qualidade de dados é abordada como objeto de estudo da CI.

A categoria 5 reúne os termos relacionados com as áreas e subáreas da CI com as quais a qualidade de dados possuem relações interdisciplinares. A figura 3 apresenta a nuvem de palavras dos termos identificados [3].

Figura 3 - Nuvem de palavras da categoria 5



Fonte: Autora (2025)

Como é possível observar na nuvem de palavras, destacam-se nos textos as discussões de qualidade de dados relacionadas à curadoria digital.

> A curadoria digital tem ganhado destaque nos últimos anos, construindo-se como uma prática interdisciplinar abrangente que busca estabelecer diretrizes e um conjunto de ações inter-relacionadas para o tratamento e a manutenção do material com valor informacional. Sendo assim, percebe-se que seu objetivo se alinha com a Organização e a Representação da informação demonstrando sua aproximação com a Ciência da Informação (Triques; Arakaki; Castro, 2020, p. 17).

A qualidade de dados, enquanto um processo que busca avaliação e melhoria de conjuntos de dados, se apresenta, portanto, como uma das ações realizadas no

[3] Os termos foram coletados considerando apenas uma ocorrência por artigo, nesse sentido não refletem quantas vezes um termo aparece em um mesmo texto, mas em quantos textos do corpus esses termos aparecem. Os termos de todas as nuvens de palavras foram coletados manualmente, categorizados e analisados manualmente, sendo aplicada ferramenta de Inteligência Artificial apenas na geração da imagem.

âmbito da curadoria digital visando garantir o tratamento e a manutenção dos recursos informacionais digitais.

Na análise da figura 2 também é possível observar um destaque de termos relacionados à organização e representação da informação e do conhecimento. Outros termos com menor destaque reforçam essa relação, tais como catalogação, classificação, tratamento da informação, catalogação descritiva e catalogação de assunto.

Com base nessa análise é possível observar uma relação de colaboração mútua e interdisciplinar entre as áreas, onde a qualidade de dados contribui para a avaliação e melhoria da qualidade dos produtos da organização e representação da informação e do conhecimento, como dos dados e metadados presentes nos catálogos, dos vocabulários controlados, da classificação e indexação dos recursos informacionais.

Por outro lado, a qualidade de dados enquanto processo se beneficia de instrumentos da organização e representação da informação e do conhecimento, seja utilizando-se dos metadados e das descrições para a melhoria do processo de avaliação de qualidade, seja tomando como base para o processo instrumentos de padronização, tais como vocabulários controlados, políticas e melhores práticas.

Destaca-se, ainda que em menor grau, a relação com os aspectos da gestão e da governança de dados. A governança de dados pode ser entendida como, “[...] um sistema de direitos decisórios e responsabilidades para processos relacionados com a informação, executados conforme modelos acordados. Esses processos descrevem quem pode tomar quais ações, com qual informação, e quando, em que circunstâncias, usando que métodos” (DGI, 2007, p. 1).

Santos e Streit (2018) apontam que a governança de dados envolve o processo de instrumentos como ciclo de vida dos dados e o estabelecimento de princípios e melhores prática voltadas para garantir a estrutura e arquitetura dos dados, a sua segurança e a qualidade dos conjuntos de dados.

A governança de dados envolve, portanto, uma série de decisões visando garantir o controle do acesso, a recuperação, a segurança e a qualidade dos conjuntos de dados ao longo de todo o seu ciclo de vida, envolvendo ainda a explicitação dessas decisões em políticas, ciclos de vida e no estabelecimento da arquitetura dos dados. A figura 4 representa os domínios de decisão da governança de dados:

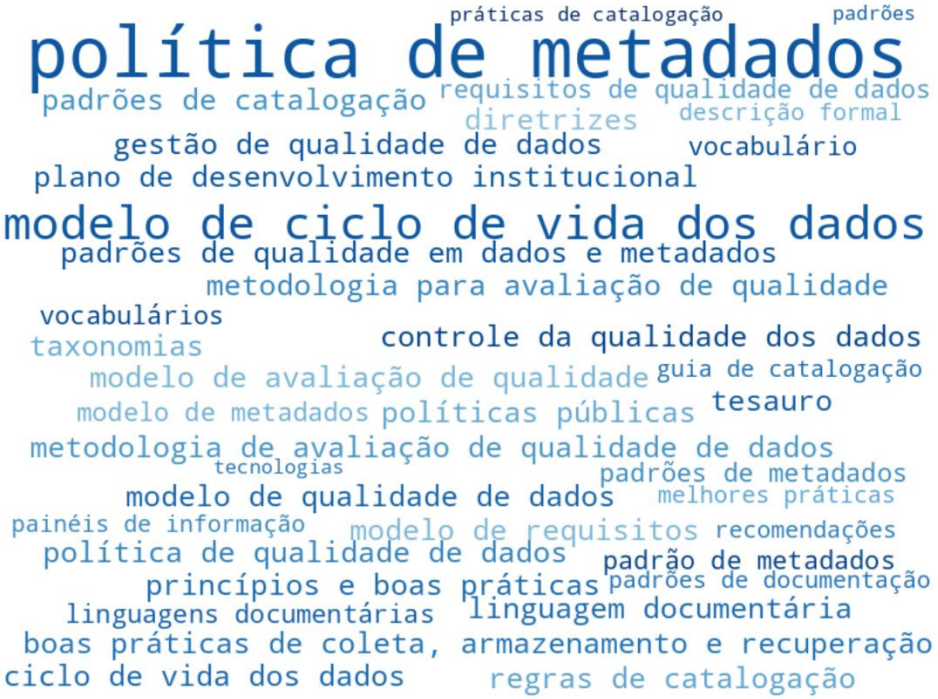
Figura 4 - Domínios relacionados à governança de dados

Fonte: Santos e Streit (2018, p. 66)

A relação entre a qualidade de dados, a organização e representação da informação e do conhecimento e a governança de dados também se destaca na análise dos termos agrupados nas categorias 6 e 7.

A categoria 6 reuniu os termos relacionados aos instrumentos empregados no processo de avaliação de qualidade. Esses instrumentos destacam-se por serem distintos dos encontrados na literatura clássica de qualidade de dados, que se concentra em estabelecer dimensões, apresentar métricas formais e estabelecer instrumentos automáticos/semiautomáticos para auxiliar no processo de avaliação e qualidade. Na CI, a qualidade de dados se destaca pela aplicação de instrumentos de padronização, como boas práticas e diretrizes. A figura 5 apresenta a nuvem de palavras da categoria 6.

Figura 5 - Nuvem de palavras da categoria 6



Fonte: Autora (2025)

Analisando os termos da categoria 6 observa-se que existe uma pluralidade grande de instrumentos, e em especial, de termos utilizados para nomear esses instrumentos. Observa-se a presença de diferentes tipos de instrumentos relacionados à gestão e governança de dados, sendo abordadas políticas, diretrizes, melhores práticas e instrumentos voltados para a governança de dados.

Destaca-se ainda a questão do ciclo e vida dos dados, uma ferramenta importante da governança de dados. A inclusão da qualidade dos dados pode ocorrer em diferentes etapas do ciclo de vida, seja na avaliação inicial de novos dados gerados, na curadoria e preservação dos dados, ou ainda na seleção de novos dados para enriquecimento dos conjuntos de dados. Essa inclusão é importante para garantir o potencial de uso e reuso dos dados e ainda a qualidade dos produtos derivados da utilização desses dados.

Destacam-se ainda entre os instrumentos aqueles relacionados à organização e representação da informação e do conhecimento, tais como políticas de metadados, padrões de metadados, vocabulários e tesauros.

Compreendendo que a qualidade de dados na maioria das vezes é considerada em seu aspecto contextual, no qual a qualidade dos dados depende de seu contexto

de aplicação, foram agrupados na categoria 7 os termos relacionados aos temas por meio dos quais se estuda a qualidade de dados na CI. Essa foi a categoria com maior número de termos, considerando termos que se repetem em mais de um artigo. A figura 6 representa a nuvem de palavras da categoria 7.

Figura 6 - Nuvem de palavras da categoria 7

Fonte: Autora (2025)

Como é possível observar existe um grande destaque para o termo metadados, sendo os metadados muitas vezes o principal objeto do processo de avaliação de qualidade na CI, produzidos no âmbito da organização e representação da informação e do conhecimento, e fundamentais para a recuperação da informação, os metadados muitas vezes são abordados de maneira conjunta aos outros termos de destaque, resultando em análise dos metadados relacionados ao patrimônio cultural, metadados publicados como *Linked Data* etc. Destaca-se ainda na figura 6 a preocupação com a recuperação da informação, objeto de estudo clássico da Ciência da Informação, sendo a qualidade um meio para a melhoria dessa recuperação, e que muitas vezes também perpassando a questão da qualidade dos metadados.

Observa-se na figura 6 o destaque de termos como patrimônio cultural, museus, acervos e coleções culturais. Também é possível observar a preocupação

com qualidade dos dados abertos, dados de pesquisa, dados governamentais, dados no contexto da *Big Data*, *Linked Data* e Web Semântica.

Com base nos estudos identificados no quadro 3 e nos termos correlatos, listados no quadro 4, observa-se a preocupação em representar, organizar, e permitir o uso e o reuso dos conjuntos de dados, preocupações relacionadas com a dimensão representacional da qualidade de dados.

Destacam-se termos relacionados à elaboração e aplicação de instrumentos de padronização, tais como vocabulários, ontologias, códigos de catalogação, padrões de metadados, políticas e boas práticas, que atuam como parâmetros para guiar o processo de avaliação de qualidade de dados na Ciência da Informação.

Quando se observam os conceitos correlatos relacionados aos objetos de estudo dos processos de avaliação de qualidade de dados na Ciência da Informação, destaca-se a abordagem interdisciplinar com as áreas tecnológicas, com os processos de elaboração, seleção e aplicação de artefatos que permitam avaliar e melhorar a qualidade dos dados, sendo possível traçar correlações com a subárea de Informação e Tecnologia.

Na Ciência da Informação, a qualidade de dados muitas vezes é discutida de maneira subordinada a outros termos, como: Curadoria Digital, Curadoria de Dados, Preservação Digital; Gestão de Dados; Governança de Dados; e Ciclo de Vida dos Dados.

Contextualizada a qualidade de dados, apresentados os principais termos necessários para a sua compreensão e a relação entre qualidade de dados e a Ciência da Informação, a próxima seção apresenta o estudo teórico da qualidade de dados *Linked Data*.

4 ESTUDO TEÓRICO DA QUALIDADE DE DADOS *LINKED DATA*

Essa seção apresenta os resultados do estudo teórico e da Revisão Sistemática da Literatura realizada a respeito da qualidade de dados *Linked Data*.

Parte-se de uma breve contextualização a respeito do *Linked Data*, composta pelos conceitos necessários para a sua compreensão, focando especialmente na estrutura de dados publicados de acordo com esses princípios.

Também discute como o contexto e a estrutura dos dados *Linked Data* impactam nos termos e nas definições relacionadas à qualidade de dados, partindo da apresentação e da discussão dos principais problemas de qualidade que afetam esse tipo de dados. Buscou-se ainda abordar o processo de avaliação de qualidade de dados *Linked Data*, discutindo a literatura disponível, as dimensões e os instrumentos que perpassam esse processo.

4. 1 A estrutura de dados publicados como *Linked Data*

Linked Data é um termo apresentado por Tim Bernes-Lee em 2006 para se referir a proposta de publicação de dados estruturados e conectado no ambiente *Web.* Embora os princípios basilares sigam sendo embasados em Bernes-Lee (2006), uma série de outros documentos, publicados anteriormente e posteriormente, tem impacto no entendimento e na adoção desses princípios.

Para um maior entendimento do termo e da estrutura de dados *Linked Data* foi construído um *corpus* documental, composto principalmente pela documentação do W3C, mas também por outros documentos correlatos. Foram considerados os documentos que definem e explicam o *Linked Data*, assim como documentos que definem conceitos, modelos e estruturas que compões os princípios, tais como URIs, RDF, *SPARql*, vocabulários e ontologias.

Também buscou-se identificar documentos do W3C que abordassem diretamente a questão da qualidade de dados *Linked Data*. Dessa busca, destacaram-se 3 documentos principais: "*Best practices for publishing Linked Data*" (W3C, 2014); "*Data on the Web Best Practices: Data Quality Vocabulary*" (W3C, 2016); e "*Data on the web best practices*" (W3C, 2017).

Incluindo a documentação relativa à estruturação do *Linked Data*, a documentação do DQV e de outros vocabulários correlatos, e os documentos

relacionados à qualidade de dados *Linked Data*, o quadro 5 apresenta os documentos que formam a base do *corpus* documental da pesquisa.

Quadro 5 - Corpus documental da pesquisa

Referência	Título do documento	DESCRIÇÃO
IETF	*Hypertext Transfer Protocol*	Discute o protocolo de transferência de recursos da *Web*, o HTTP.
Berners-Lee (2006)	*Linked Data*	Apresenta os princípios basilares da publicação de dados como *Linked Data*.
W3C (2011)	*W3C. URIs, URLs e URNs: Clarifications and Recommendations*	Discute os conceitos de URIs, URLs e URNs, apresentados e discute melhores práticas.
W3C (2013)	*Linked Data Glossary*	Glossário de termos relacionados ao *Linked Data.*
W3C (2004)	*Primer RDF*	Explica a estrutura do modelo RDF.
W3C (2015a)	*Data Quality Vocabulary (DQV)*	Documento que apresenta uma visão geral do DQV.
W3C (2015b)	*Vocabularies*	Apresenta o conceito de vocabulários na perspectiva do W3C.
W3C (2016a)	*Data on the web best practices: Data Quality Vocabulary*	Documentação oficial do DQV.
W3C (2016b)	*List of DQV implementations*	Documento que faz referência a algumas das aplicações do DQV.
W3C (2017)	*Data on the web best practices*	Conjunto de Melhores Práticas (MPs) para a publicação de dados na *Web*.
W3C (2018)	*Links in html documents*	Explica a estrutura de links
W3C (2020)	*Data Catalog Vocabulary (DCAT) - Version 2*	Documentação oficial do DCAT, vocabulário do qual o DQV é uma extensão.
ISO (2022)	ISO/IEC 25012	Apresenta dimensões e métricas de qualidade

Fonte: Autora (2025)

Embora os documentos do W3C destacados no quadro reconheçam a importância da qualidade de dados para a sua recuperação e reutilização e que a

adoção das melhores práticas existentes tenha impacto na qualidade dos dados publicados, observou-se que nenhum desses documentos fornece orientações claras sobre como realizar o processo de avaliação de qualidade e ainda sobre como selecionar conjuntos de dados para diferentes contextos.

Com base no estudo documental, na literatura que compõe o *corpus* teórico da pesquisa e na identificação de demais fontes pertinentes, obtidas por meio de Snowballing, apresenta-se a seguir o *Linked Data*, focando em sua definição e na estrutura dos dados *Linked Data*.

Para Bizer, Heath e Berners-Lee (2009, p. 2, tradução nossa) o *Linked Data* é “sobre como usar a *Web* para criar ligações entre os dados de diferentes fontes” (Bizer; Heath; Berners-Lee, 2009, p. 2).

O termo *Linked Data* está relacionado então a um propósito, o de “criar um ecossistema de produção e consumo de dados com o objetivo de agilizar a descoberta de novos conhecimentos e agregar valor a qualquer informação disponibilizada livremente na Internet.” (Isotani; Bittencourt, 2015, p. 17).

Enquanto um propósito, o *Linked Data* representa o objetivo de permitir um cenário caracterizado por “[...] comunidades concordando com o significado de seus dados e compartilhando-os em um espaço de informações massivamente em rede” (OCLC, 2024, não paginado, tradução nossa).

Esse objetivo atraiu a atenção de diversos setores, como aponta a OCLC (2024, não paginado, tradução nossa) “Essa visão está tomando forma em muitos setores, incluindo comércio eletrônico, medicina, pesquisa científica e serviços governamentais”.

Para compreender a relevância dessa proposta, torna-se necessária uma breve contextualização sobre o ambiente *Web*.

> Em 1989, o físico inglês Sir Timothy John Berners-Lee, no CERN, inventou a WWW (*World Wide Web*) a partir da proposição de três tecnologias fundamentais: o HTML (*Hypertext Markup Language*), o servidor HTTP (*Hypertext Transfer Protocol*) e o URI (*Unified Resource Identifier*). (Isotani; Bittencourt, 2015, p. 25).

O HTML foi a base para garantir o crescimento da *Web*, fornecendo uma estrutura simples para a publicação e visualização de conteúdos diversos.

> A linguagem de marcação *Hypertext Markup Language* (HTML) foi desenvolvida para descrever as estruturas das páginas criadas na *Web*, permitindo a inserção de recursos informacionais e a criação de *hiperlinks*, ou seja, a ligação entre os recursos contidos em uma mesma página ou em páginas diversas. (Jesus, 2021, p. 54).

O conceito de *hyperlinks*, ou simplesmente *links*, é derivado do conceito de hipertexto, cunhado por Theodore Nelson, que representa a "ideia de leitura/escrita não-linear em sistemas informatizados" (Dias, 2019, p. 272). A respeito do hipertexto Dias (2019, p. 269) aponta que:

> O hipertexto se insere no contexto da cibercultura, como uma de suas novas interfaces de comunicação. Na verdade, o hipertexto resgata e modifica antigas interfaces da escrita, como a segmentação em módulos (capítulos e seções), o acesso seletivo e não-linear ao texto (índices e sumários), as conexões a outros documentos (notas de rodapé e referências bibliográficas), implementadas com novas tecnologias. Essa nova maneira de escrever pode ser usada para organizar e divulgar os conhecimentos sobre uma determinada área do saber, sendo especialmente útil nas áreas de gestão de informações, comunicação e educação.

A linguagem HTML, viabiliza por meio dos *links* a relação entre recursos informacionais, permitindo que os usuários, ao clicarem nesses *links*, naveguem entre informações e recursos complementares ou correlatos, disponibilizados por uma mesma fonte ou em fontes diversas.

O W3C (2018, não paginado, tradução nossa) explica o funcionamento dos *links* no contexto da HTML:

> Um link tem duas extremidades - chamadas âncoras - e uma direção. O *link* começa na âncora "recurso" e aponta para a âncora "destino", que pode ser qualquer recurso da *Web* (por exemplo, uma imagem, um videoclipe, uma frase de áudio, um programa, um documento HTML, um elemento dentro de um documento HTML etc.). (W3C, 2018, não paginado, tradução nossa).

A identificação e localização das páginas e recursos disponíveis é realizada por meio de *Uniform Resource Locator* (URLs) e *Uniform Resource Name* (URN), ou ainda de *Uniform Resource Identifier* (URIs).

Os *Uniform Resource Identifier* (URIs) são identificadores únicos que permitem representar a páginas e recursos da *Web*. Também permitem representar e fazer referência, bem como conceitos do mundo real, tais como locais, objetos e pessoas.

> Durante os primeiros anos de discussão sobre identificadores da *web* (início a meados dos anos 90), as pessoas presumiram que um tipo de identificador seria convertido em uma de duas (ou possivelmente mais) classes. Um identificador pode especificar a localização de um recurso (uma URL) ou seu nome (uma URN) independentemente da localização. Assim, um URI era uma URL ou uma URN. (W3C, 2011, não paginado, tradução nossa).

Com o avanço das aplicações na *Web*, essa fragmentação, que fazia com que o URI pudesse ser classificado de maneira excludente em apenas uma das duas classes, perdeu relevância, no contexto atual. Entende-se que o espaço do URI pode ser particionado, sendo composto pelas duas classes, com o objetivo de identificar de maneira única o nome e a localização dos recursos na *web* (W3C, 2011).

O *Hypertext Transfer Protocol* (HTTP) é um protocolo de solicitação e resposta que permite o acesso às informações contidas no ambiente *web*:

> O *Hypertext Transfer Protocol* (HTTP) é um protocolo de nível de aplicação para sistemas de informação de hipermídia distribuídos e colaborativos. É um protocolo genérico, que pode ser usado para muitas tarefas, além de seu uso para hipertexto[...]. Um recurso do HTTP é a digitação e negociação da representação de dados, permitindo que os sistemas sejam construídos independentemente dos dados que estão sendo transferidos (W3C, 1999, não paginado, tradução nossa).

O HTTP orienta a troca de informações entre dois computadores conectados na *web*, o servidor, que contém o recurso solicitado, e o cliente/usuário, solicitante desse recurso.

Em síntese, o funcionamento da *web* está baseado, desde seu princípio, nos seguintes aspectos-chave: o uso do HTML para disponibilização e visualização de recursos multimídia; a adoção do HTTP como protocolo para a troca de informações; a utilização de URIs e URLs para a identificação e localização dos recursos disponibilizados.

O grande diferencial da *web* está justamente na materialização do hipertexto por meio dos *links*, que permitem que os recursos disponibilizados possam ser conectados entre si, possibilitando o acesso não linear a esses recursos.

Embora os documentos HTML sejam processáveis por máquina e apresentem uma estruturação, algumas das limitações dessa linguagem levaram a problemas de

recuperação no contexto da *web,* especialmente quando de sua popularização e massificação.

> Documentos HTML, que implementam a linguagem de marcação central da *Web*, são semiestruturados, mas têm limitações quando se trata de processabilidade por máquina, porque são escritos principalmente para humanos. Embora possam ser usados para exibir dados em um site, ferramentas de *software* não conseguem "entender" a descrição de pessoas e objetos do mundo real descritos em HTML, nem os relacionamentos entre esses *links* (SIKOS, 2015, p. 7).

Ramalho, Martins e Souza (2017, p. 25) apontam que, para além do seu foco em usuários humanos, as limitações do HTML se dão por sua concepção baseada em formatos fixos e pelo foco na apresentação e não na representação dos recursos contidos na *Web*:

> Contudo, sua concepção baseada em marcações fixas trouxe limitações para a representação de informações uma vez que, sendo uma linguagem de marcação voltada, principalmente, para a apresentação dos documentos em ambientes digitais, HTML apenas define a forma como a informação é exibida não se preocupando com o significado da palavra. (Ramalho; Martins; Souza, 2017, p. 25).

Também existem limitações em relação aos *links* criados com base no HTML, "os documentos HTML usam *hiperlinks* para vincular recursos da *web* relacionados ou partes de documentos da *web* entre si; no entanto, não há nenhuma informação sobre o tipo desses *links* (SIKOS, 2015, p. 7)."

Dessa forma, não é possível para as máquinas distinguirem o significado das relações existentes entre esses *links*. (Isotani; Bittencourt, 2015). Essas limitações se ampliam no contexto atual, como apontado por Isotani e Bitterncourt (2015, p. 26):

> É importante frisar que tais problemas ficam mais evidentes nos dias de hoje, pois se calculam dezenas de bilhões de páginas *web* disponíveis e mais de um zettabyte de dados. Esta quantidade de documentos torna praticamente impossível o acesso e a busca por informação de forma eficiente e consistente para os seres humanos.

Embora as páginas da *web* publicadas em HTML sejam processáveis por agentes computacionais, para facilitar o uso desses agentes e melhorar a recuperação no contexto da *web*, "[...] também é necessário que esses dados sejam dotados de uma semântica formal que especifica claramente quais conclusões devem ser

elaboradas a partir das informações coletadas (Hitzler, Krötzsch; Rudolph, 2010, p. 11-12, tradução nossa).

Tim Berner-lee (2006), definiu então o *Linked Data* como uma busca pela *web of data*, ou *web* de dados, em contrapartida a *web* de documentos, ou seja, a *web* baseada exclusivamente nas ligações entre recursos informacionais, criadas com base no formato HTML, com foco na apresentação e disponibilização desses recursos para usuários humanos.

> Assim como a *web* de hipertexto, a *web* de dados é construída com documentos na *web*. No entanto, diferentemente da *web* de hipertexto, onde *links* são âncoras de relacionamentos em documentos de hipertexto escritos em HTML, para dados eles são *links* entre coisas arbitrárias descritas por RDF (Berner-lee, 2006, não paginado, tradução nossa).

É nesse sentido que o *Link Data* recebe ainda uma outra acepção, a de um conjunto de princípios ou melhores práticas criados para viabilizar esse objetivo, "[...] o termo *Linked Data* refere-se a um conjunto de melhores práticas para a publicação e interligação de dados estruturados na *web*" (Heath, Bizer, 2011, não paginado, tradução nossa).

Ao longo dos anos foram publicadas uma série de melhores práticas e recomendações relacionadas à publicação de dados como *Linked Data,* tais como as "melhores práticas para a publicação *Linked Data*" (W3C,2014) e "melhores práticas para a publicação de dados na *Web*" (W3C, 2017).

Essas melhores práticas e recomendações se aprofundaram em diversos aspectos da publicação de dados como *Linked Data*, entretanto, em sua essência os princípios podem ser resumidos em: 1) Uso de identificadores únicos ou *Uniform Resource Indetifier* (URIs); 2) Uso do protocolo padrão de transferência de dados da *Web*, o *Hypertext Transfer Protocol* (HTTP), para compartilhamento de informações sobre os URIs; 3) adoção de formatos padrões, em especial o *Resource Description Framework* e o SPARQL, para fornecer informações úteis sobre os URIs (RDF); 4) Criação de *links* entre os conjuntos de dados, relacionando dados em diferentes fontes (Berners-Lee, 2006).

Os princípios do *Linked Data* são "fundamentados em tecnologias *web*, como HTTP (*Hypertext Transfer Protocol*) e URI *(Uniform Resource Identifier*) com o objetivo

de permitir de forma automática, por agentes de *software*, a leitura dos dados conectados.” (Isotani; Bittencourt, 2015, p. 31-32).

Os dois primeiros princípios do *Linked Data* estão relacionados ao uso de URIs como identificadores e da adoção do protocolo HTTP para a disponibilização de informações úteis sobre esses URIs, tanto para usuários humanos como em formato legível por máquina. O estabelecimento de URIs permite:

> a) conectar e combinar estes dados com outros dados; b) reusar estes dados em outros contextos; c) melhorar a busca e a compreensão dos dados apresentados; d) possibilitar inferência a partir de dados parciais; e) permitir navegação entre documentos, entre outras atividades (Isotani; Bittencourt, 2015, p. 53).

As melhores práticas para a publicação de dados como *Linked Data* do W3C (2014) se aprofundam no uso de URIs, apontando que bons URIs para esse princípio devem: Ser URIs HTTPs; quando desreferenciados, fornecerem ao menos uma versão legível por máquina; serem pautados em uma política interna e na formulação de um plano para sua implementação; serem estáveis e para isso não conter informações que precisem ser alteradas, evitando informações descritivas em linguagem natural.

Os dois primeiros pontos se relacionam pois:

> Quando um cliente HTTP pode procurar um URI usando o protocolo HTTP e recuperar uma descrição do recurso, ele é chamado de URI desreferenciável. URIs desreferenciáveis se aplicam a URIs que são usados para identificar documentos HTML clássicos e URIs que são usados no contexto de *Linked Data* para identificar objetos do mundo real e conceitos abstratos (W3C, 2013, não paginado, tradução nossa).

Para serem considerados persistentes, os URIS devem direcionar para o mesmo conjunto de dados ou recurso, mesmo que esses tenham sido alterados ou removidos. Nesse segundo caso, onde os recursos/dados não estejam mais disponíveis, os URIs devem remeter a uma informação que indique o destino desses dados, sem serem reutilizados.

A própria linguagem HTML fornece soluções para garantir a persistência dos URIs, disponibilizando meios para o fornecimento de informações a respeito de dados indisponíveis ou alterados (W3C, 2014).

Uma dessas possibilidades, é o remetimento a erros pré-estabelecidos, como por exemplo, o 303, que indica que a localização do recurso foi alterada; 410 que indica que o recurso foi definitivamente excluído; e o 503, que indica que o recurso está temporariamente indisponível (W3C, 2017).

Também relacionado à persistência de URIs, tem-se o PURL (*Persistent Uniform Resource Locator*), que permite realizar o direcionamento em casos em que a localização dos recursos foi alterada (W3C, 2013).

Como mencionado, os dois primeiros princípios do *Linked Data* são relacionados a padrões da *web* convencional e estão diretamente relacionados, o primeiro diz respeito ao uso de bons identificadores para recursos, dados, e para a representação de informações do mundo real e “[...] o segundo princípio dos Linked Data aponta o uso de URIs HTTP para identificar objetos e conceitos abstratos, permitindo que esses URIs sejam desreferenciados, ou seja, pesquisados”. (Heath; Bizer, 2011, não paginado, tradução nossa).

O terceiro princípio diz respeito à adoção de padrões para fornecimento de informações relevantes a respeito dos URIs, destacando o modelo *Resource Description Framework* (RDF).

O RDF permite a explicitação das relações entre os dados de maneira formal, legível por máquinas.

> O *Resource Description Framework* (RDF) é uma linguagem para representar informações sobre recursos na *World Wide Web*. Ele é particularmente destinado a representar metadados sobre recursos da *Web*, como o título, autor e data de modificação de uma página da *Web*, informações de copyright e licenciamento sobre um documento da *Web* ou o cronograma de disponibilidade para algum recurso compartilhado. No entanto, ao generalizar o conceito de um "recurso da *Web*", o RDF também pode ser usado para representar informações sobre coisas que podem ser *identificadas* na *Web*, mesmo quando não podem ser *recuperadas* diretamente na *Web*. Exemplos incluem informações sobre itens disponíveis em lojas *on-line* (por exemplo, informações sobre especificações, preços e disponibilidade) ou a descrição das preferências de um usuário da *Web* para entrega de informações. (W3C, 2004, não paginado, tradução nossa).

A descrição dessas relações é realizada através de declarações em formato de triplas <Recurso + Propriedade + Valor>, onde, de acordo com a definição adotada para essa pesquisa, cada tripla representa um dado.

> O RDF é baseado na ideia de identificar coisas usando identificadores da *Web* (chamados *Uniform Resource Identifiers*, ou *URIs*), e descrevendo recursos em termos de propriedades simples e valores de propriedade. Isso permite que o RDF represente declarações simples sobre recursos como um *gráfico* de nós e arcos representando os recursos, e suas propriedades e valores. (W3C, 2004, não paginado, tradução nossa).

A estrutura das declarações em RDF partem do princípio de que para descrever as propriedades de algo é necessário fragmentá-la em três aspectos: o que está sendo descrito (recurso), a característica específica desse recurso que está sendo descrita (propriedade), e o que exatamente essa declaração diz a respeito da coisa descrita (valor). (W3C, 2014).

Segundo o W3C (2004), o uso adequado do RDF implica em:

- Tanto o recurso como as propriedades precisam ser representados por URIs;
- O valor das declarações pode ser descrito usando um URI ou um literal (números, nomes, palavras, frases em linguagem natural);
- Cada tripla representa somente uma propriedade do recurso, para outras propriedades são criadas novas triplas;
- O RDF precisa do suporte de uma linguagem computacional - um formato de serialização – que o torne legível por máquinas; e
- As propriedades precisam ser derivadas de vocabulários.

As triplas RDF podem ser representadas no formato de grafos, chamados também de *knowledge graph*, que facilitam a compreensão para usuários humanos, “por convenção, os recursos são representados como elipses, os valores literais como retângulos e os predicados [propriedades] utilizando arcos direcionados do recurso (sujeito) para o valor (objeto)” (Ramalho; Martins; Souza, 2017, p. 29).

> Às vezes não é conveniente desenhar gráficos ao discuti-los, então uma maneira alternativa de escrever as declarações, chamada de triplas, também é usada. Na notação de triplas, cada declaração no gráfico é escrita como uma tripla simples de sujeito, propriedade e valor, nessa ordem. (W3C, 2004, não paginado, tradução nossa)

Para ser legível por máquina, essa notação depende de um formato de serialização, existindo diversos formatos disponíveis.

Para que as propriedades possam ser descritas é necessário criar/aplicar vocabulários:

> Vocabulários definem os conceitos e relacionamentos (também chamados de “termos” ou “atributos”) usados para descrever e representar uma área de interesse. Eles são usados para classificar os termos que podem ser usados em uma aplicação específica, caracterizar possíveis relacionamentos e definir possíveis restrições ao uso desses termos (W3C, 2017, tradução nossa).

Nesses vocabulários, cada propriedade é identificada com um URI, os vocabulários são criados para distintos contextos, sendo uma recomendação a reutilização de vocabulários publicados por fontes confiáveis.

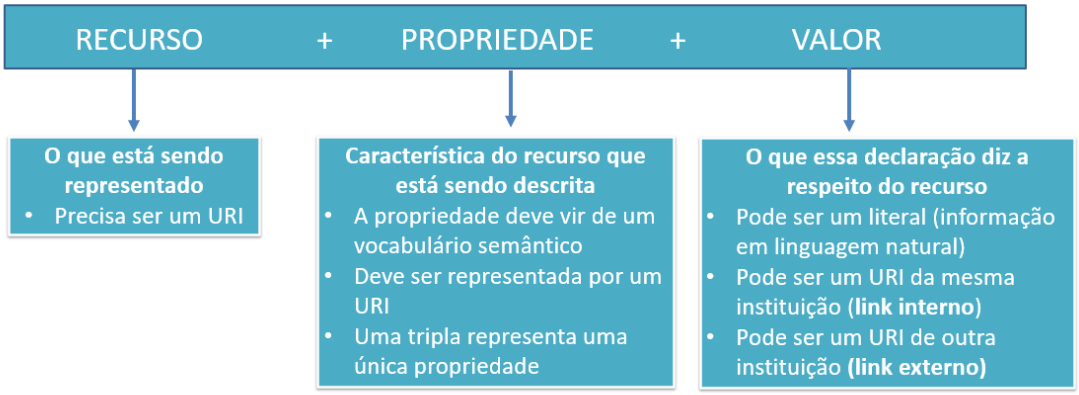
A ligação entre os dados, a que se refere o quarto princípio do *Linked Data* acontece por meio das declarações em RDF, que podem ser divididas entre: as **triplas literais**, onde o valor é apresentado em linguagem natural, e os ***links* RDF**, onde o valor é um URI do mesmo conjunto de dados ou de dados externos. (Heath; Bizer, 2011).

Assim, a ligação pode permitir a navegação interna em conjunto de dados ou ainda a ligação entre diferentes fontes que abordem um mesmo recurso. Como os recursos podem ser representados por diferentes URIs em fontes distintas, uma ligação comum é a feita por meio da propriedade *owl:same*[4] que permite indicar que um URI representa um mesmo recurso que o URI de outra fonte.

Com base nas discussões apresentadas, a figura 7 apresenta uma síntese da estrutura de dados *Linked Data* pautada no modelo RDF.

[4] Propriedade da ontologia *Ontology Web Language* (owl)

Figura 7 - Síntese da estrutura dos dados *Linked Data*



Fonte: Autora (2025)

Embora os dados publicados como *Linked Data* sejam estruturados e pensados para a recuperação e reutilização tanto por agentes computacionais como por usuários humanos “A publicação de dados conectados (seguindo os princípios de Linked Data) não garante a qualidade dos dados” (SIKOS, 2015, p. 60).

Mesmo existindo um conjunto de princípios, melhores práticas e diretrizes para guiar a publicação de dados como *Linked Data*, não existe uma garantia de que os dados sigam completamente todas as orientações disponíveis. Esses conjuntos de dados não estão livres de más formações, de dados imprecisos, incompletos ou incorretos.

Esse cenário se intensifica por uma política de “publicar primeiro e melhorar depois”, que acompanhou muitos dos projetos iniciais de publicação de dados como *Linked Data* (Feeney *et al.*, 2014).

A própria estrutura do *Linked Data* também acrescenta uma série de complexidade para a avaliação dos dados, pois é necessário levar em conta sua adequação às melhores práticas e os contextos de criação e aplicação heterogêneos. Com base nessas questões, a próxima subseção apresenta o estado da arte da qualidade de dados *Linked Data*.

4.2 O estado da arte da qualidade em dados *Linked Data*

Visando identificar o estado da arte da qualidade de dados *Linked Data*, foi realizada a Revisão Sistemática da Literatura a respeito do tema. As buscas baseadas

na estratégia do protocolo retornaram 225 (duzentos e vinte e cinco) documentos, sendo 29 (vinte e nove) duplicados e 104 (cento e quatro) documentos rejeitados.

Entre os documentos rejeitados, 43 (quarenta e três) foram excluídos por não discutirem a temática de interesse em profundidade, apenas mencionando o termo. Esses documentos abordam a qualidade como uma etapa dos processos de publicação e reutilização de conjuntos de *Linked Data,* ou como uma contextualização das problemáticas relacionadas à diversos cenários tecnológicos, tais como *Big Data*, internet das coisas, Inteligência Artificial e *Web* Semântica.

Foram aceitos 92 (noventa e dois) artigos para compor o *corpus* teórico da pesquisa. Depois da seleção, procedeu-se a categorização dos estudos com base na abordagem adotada do tema.

As categorias foram estabelecidas a *posteriori*, foram sumarizados os enfoques dos documentos com base em uma síntese dos seus objetivos, procedimentos metodológicos e dos resultados obtidos. Ao final, foram estabelecidas 3 (três) categorias que refletem as principais abordagens sobre qualidade de dados *Linked Data*, o quadro 6 apresenta o número de documentos aceitos por categoria.

Quadro 6 - Número de documentos por aceitos por categoria

N°	Categoria	Quantidade de artigos incluídos
1	Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*	72
2	Realiza um estudo de avaliação de qualidade em um ou mais conjuntos de dados	12
3	Levantamentos e estudos teóricos sobre qualidade de dados e *Linked Data*	8

Fonte: Autora (2025)

As categorias foram apresentadas no quadro 6 em ordem decrescente de número de artigos incluídos, sendo, portanto, a mais volumosa delas a categoria “1 - Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*.” Nessa categoria, foram incluídos todos os artigos que tinham por objetivo apresentar, propor ou discutir um artefato[5].

[5] Para essa pesquisa um artefato é entendido enquanto instanciação física elaborada com a finalidade de representar determinada realidade, solucionar problemas ou promover melhorias/inovação, podendo se materializar em modelos, sistemas de informação, fluxos, *constructus*, métodos, entre outras formas (March; Smith, 1995; Hevner *et al.*, 2004).

O quadro 7 apresenta a lista dos documentos aceitos incluídos nessa categoria, bem como uma breve síntese de seus objetivos.

Quadro 7 - Documentos aceitos incluídos na categoria “1 Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*”

AUTORES	TÍTULO	DESCRIÇÃO
Fürber e Hepp (2010)	*Using SPARQL and SPIN for Data Quality Management on the Semantic Web*	Apresenta a possibilidade de aplicação do SPARQL no processo de avaliação de qualidade de dados *Linked Data.*
Rula (2011)	*DC proposal: towards linked data assessment and linking temporal facts*	Propõe melhorias em um *framework* de análise de qualidade já existentes, visando a avaliação dos aspectos temporais. Propõe ainda um repositório de dimensões e métricas de qualidade focadas em dados *Linked Data*.
Gamble *et al.* (2012)	*MIM: A Minimum Information Model Vocabulary and Framework for Scientific Linked Data*	Propõe um *checklist (*MIM-*Minimum Information Model*) para guiar a publicação e avaliação de dados relacionados a área de ciências da vida.
Moss; Corsar, Piper (2012)	*A Linked Data Approach to Assessing Medical Data*	Propõe um *framework* para estimar a confiabilidade de dados *Linked Data*, focado em dados da área médica.
Acosta *et al.* (2013)	*Crowdsourcing linked data quality assessment*	Propõe uma metodologia para aplicação de *crowdsourcing* na solução de problemas de qualidade complexos em dados *Linked Data*, que não podem ser identificados automaticamente.
Kontokostas *et al.* (2013)	*TripleCheckMate: a tool for crowdsourcing the quality assessment of linked data*	Propor uma metodologia e semiautomática de avaliação de qualidade de dados *Linked Data* associada a aplicação de *crowdsourcing*
Silva *et al.* (2013)	*Glocal Clinical Registries: Pacemaker Registry Design and Implementation for Global and Local Integration - Methodology and Case Study*	Propõe um *framework* para diversas atividades, incluindo controle de qualidade, focado no ciclo de vida dos dados médicos.
Rinser, Lange e Naumann (2013)	*Cross-lingual entity matching and infobox alignment in Wikipedia*	Propõe um esquema para identificação correção de problemas de qualidade

		relacionados a fontes multilíngues.
Ruckhaus, Baldizán e Vidal (2013)	*Analyzing Linked Data quality with liquate*	Propõe uma ferramenta (LiQuate) com a finalidade de problemas de qualidade e ambiguidades entre conjuntos de dados e *links*.
Cherix *et al.* (2014)	*Lessons Learned - The Case of CROCUS: Cluster-Based Ontology Data Cleansing*	Apresenta uma ferramenta (*Cluster-based Ontology Data Cleansing*- CROCUS) para a identificação e correção de falhas singulares ou erros de instância única em conjuntos de dados *Linked Data*.
Darari, Prasojo e Nutt (2014)	*CORNER: A Completeness Reasoner for SPARQL Queries Over RDF Data Sources*	Propõe uma ferramenta (corner) para seleção de conjuntos de dados com base na avaliação da dimensão *completeness*, visando identificar se as fontes de dados são suficientemente completas para a aplicação.
Feeney *et al.* (2014)	*Improving Curated Web-Data Quality with Structured Harvesting and Assessment*	Propõe um *framework* e uma ferramenta semiautomática para coletar, avaliar e realizar a manutenção de qualidade de dados *Linked Data* com foco na manutenção da qualidade dos dados locais, incluindo a manutenção dos Links com fontes externas.
Fleischhacker *et al.* (2014)	*Detecting Errors in Numerical Linked Data Using Cross-Checked Outlier Detection*	Propõe o uso de uma ferramenta (*Cross-Checked Outlier Detection*) para detecção de problemas de qualidade relacionado a valores numéricos descritos em conjuntos de dados *Linked Data*.
Ibáñez *et al.* (2014)	*Col-Graph: Towards Writable and Scalable Linked Open Data*	Propõe uma ferramenta para identificar e corrigir problemas de qualidade em conjuntos de dados *Linked Data* criados a partir de fragmentos reunidos de diversas fontes.
Kontokostas *et al.* (2014a)	*Databugger: A Test-Driven Framework for Debugging the Web of Data*	Propõe um *framework* (*Databugger*) para *Test-Driven* de qualidade em conjuntos de dados
Kontokostas *et al.* (2014b)	*Test-driven Evaluation of Linked Data Quality*	Propõe a aplicação da Metodologia "*Teste-driven*", consolidada na área de desenvolvimento de *software*, para avaliação automatizada de qualidade dos dados *Linked Data*.

Krompaß, Nickel e Tresp (2014)	*Querying Factorized Probabilistic Triple Databases*	Propõe uma ferramenta para avaliação e correção de dados focada na veracidade dos dados *Linked Data*, a partir da perspectiva dos bancos de dados probabilísticos (PDB).
Paulheim e Bizer (2014)	*Improving the Quality of Linked Data Using Statistical Distributions*	Apresenta dois algoritmos para promover melhorias de qualidade baseado em distribuição estatística, focado em dados extraídos e convertidos de fontes semiestruturadas e não estruturadas
Assaf; Senart e Troncy (2015a)	*Roomba: Automatic Validation, Correction and Generation of Dataset Metadata Enhancing Dataset Search and Spam Detection*	Propõe uma ferramenta (Roomba) para validação, correção e geração de metadados para dados publicados como *Linked Data*. (abordagem representacional)
Assaf, Troncy e Senart (2015c)	*What's up LOD Cloud? Observing the State of Linked Open Data Cloud Metadata*	Apresentar uma ferramenta (Roomba) que permite validação, correção e geração de metadados a respeito de *datasets Linked Data* (abordagem representacional).
Bonner *et al* *(2015)*	*Data Quality Assessment and Anomaly Detection Via Map/Reduce and Linked Data: A Case Study in the Medical Domain*	Apresenta um *framework* para melhorias de qualidade voltado para dados *Linked Data* da área da saúde.
Debattista *et al.* (2015)	*Quality Assessment of Linked Datasets Using Probabilistic Approximation*	Propõe a aplicação de técnica probabilísticas comuns em cenários *Big Data* para a avaliação de qualidade de dados *Linked Data* em grande escala, aplicando ainda a ferramenta LUZZU.
Dimou *et al.* (2015)	*Assessing and Refining Mappings to RDF to Improve Dataset Quality*	Apresenta uma metodologia focada nas questões relacionadas com o formalismo das declarações RDF, buscando avaliar inconsistências relacionadas com a má aplicação dos esquemas, vocabulários e ontologias.
Knuth (2015)	*Linked Data Cleansing and Change Management*	Propõe uma ontologia/vocabulário focado no fornecimento de *feedbacks* dos consumidores, permitindo que esses indiquem problemas de qualidade em dados *Linked Data*.

Tarasowa, Lange e Auer (2015)	*Measuring the Quality of Relational-to-RDF Mappings*	Propõe uma ferramenta (R2RML *Mapping Language*) pra avaliar os mapeamentos utilizados como base do processo de avaliação de qualidade.
Assaf; Senart e Troncy (2016a)	*Roomba: An Extensible Framework to Validate and Build Dataset Profiles*	Propõe um *framework* (Roomba) para extrair, validar, corrigir e gerar metadados para conjuntos de dados *Linked Data.* (abordagem representacional).
Assaf, Senart e Troncy (2016b)	*Towards an objective assessment framework for linked data quality: enriching dataset profiles with quality indicators*	Elaborar um *framework* para avaliação de qualidade e uma ferramenta para auxiliar no processo de seleção de conjunto de dados *Linked Data* com base em critérios de qualidade.
Debattista, Auer e Lange (2016)	*Luzzu - A framework for linked data quality assessment*	Apresentar um *framework* para avaliação automatizada de qualidade de dados *Linked Data* (Luzzu).
Cappiello *et al.* (2016)	*A Quality Model for Linked Data Exploration*	Propõe um modelo baseado na aplicação de dimensões de qualidade para auxiliar no processo de seleção de conjuntos de dados
Feeney *et al.* (2016)	The Dacura Data Curation System	Discute o uso de uma ferramenta (*Dacura system*) voltada para obtenção, avaliação, melhoria e gerenciamento de dados *Linked Data*.
Hassan *et al.* (2016)	*ACRyLIQ: Leveraging DBpedia for Adaptive Crowdsourcing in Linked Data Quality Assessment*	Propor uma ferramenta (*Adapative Crowdsourcing for Linked Data Quality Assessment -ACRyLIQ*) que auxilia na utilização de *crowdsourcing* em processos de avaliação de qualidade de dados *Linked Data*, partindo da estimativa de confiabilidade dos trabalhadores em distintas tarefas.
Muñoz (2016)	*On Learnability of Constraints from RDF Data*	Apresenta uma metodologia baseada no uso de "*constraints*" para permitir avaliação de qualidade de conjuntos de dados *Linked Data*.
Nooghabi e Dastgerdi (2016)	*Proposed metrics for data accessibility in the context of Linked Open Data*	Propõe um modelo e métricas para avaliação da dimensão acessibilidade em dados *Linked Data.*

Rahoman e Ichise (2016)	*Automatic Erroneous Data Detection over Type-Annotated Linked Data*	Propõe um *framework* (*Auto Linked Data Error Detector-ALDErrD*) para a identificação de erros em dados publicados como *Linked Data*
Ahmed (2017)	*Data Quality Assessment in the Integration Process of Linked Open Data (LOD)*	Propõe a elaboração de um fluxo de trabalho para modelagem e avaliação de qualidade de dados *Linked Data*.
Albertoni, Martino e Quarati (2017)	*Linked Thesauri Quality Assessment and Documentation for Big Data Discovery*	Propõe uma metodologia para avaliação de qualidade de tesauros publicados como *Linked Data.*
Beek *et al.* (2017)	*Literally Better: Analyzing and Improving the Quality of Literals*	Propõe uma ferramenta para analisar a qualidade de literais (ou seja, valores em linguagem natural) em grande escala.
Esteves *et al.* (2017)	*Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis*	Propõe um validador de triplas RDF (*Deep Fact Validation-DeFacto*), visando verificar a veracidade de fatos em conjuntos de dados publicados seguindo a estrutura do RDF.
Färber *et al.* (2017)	*Linked Data quality of dbpedia, freebase, opencyc, wikidata, and yago*	Propor um *framework* visando a avaliação de conjuntos de dados de livre acesso, focados em dados de conhecimentos gerais, também chamados de domínio cruzado ou enciclopédico. Também tem como objetivo realizar o processo de avaliação de qualidade em um conjunto de dados.
Feeney, Gleason e Brennan (2017)	*Linked Data schemata: fixing unsound foundations*	Propõe uma ferramenta (*Dacura Quality Service*) e uma metodologia para identificação e resolução de problemas de qualidade relacionados a heterogeneidade de vocabulários aplicados na elaboração de dados *Linked Data*.
Liu *et al.* (2017a)	*Exploiting Source-Object Networks to Resolve Object Conflicts in Linked Data*	Propõe um *framework* para lidar com as questões de conflito de objeto (*ObResolution*) em dados *Linked Data*.
Liu *et al.* (2017b)	*TruthDiscover: Resolving Object Conflicts on Massive Linked Data*	Propõe uma ferramenta (*TruthDiscover*) para a identificação e resolução de problemas de conflitos de objetos em dados *Linked Data*.

Mihindukulasooriya *et al.* (2017)	*A Linked Data Profiling Service for Quality Assessment*	Propõe uma ferramenta configurável para "*data profiling*" de dados *Linked Data*, que pode ser aplicada tanto de maneira manual como ser insumo para avaliação de qualidade automatizada.
Mihindukulasooriya, García-Castro e Gómez-Pérez (2017)	*LD Sniffer: A Quality Assessment Tool for Measuring the Accessibility of Linked Data*	Propõe uma ferramenta de código aberto, baseada na estrutura da W*eb*, para auxiliar na avaliação de qualidade de dados *Linked Data* com uma abordagem de acessibilidade.
Nahari *et al.* (2017)	*A Framework for Linked Data Fusion and Quality Assessment*	Propõe um *framework* (LDIF) para auxiliar nos processos de fusão e na avaliação de qualidade de dados *Linked Data*.
Melo, Botega e Segundo (2017b)	Metodologia de Avaliação de Qualidade para Dados Conectados	Propõe um modelo identificando as principais dimensões relacionadas com a avaliação de qualidade de dados *Linked Data*.
Radulovic *et al.* (2017)	*A comprehensive quality model for Linked Data*	Propõe um modelo de qualidade de dados para dados *Linked Data*, uma ferramenta de avaliação de qualidade e um vocabulário para a comunicação dos resultados (extensão do W3C *Data Quality Vocabulary*).
Van Hoeven *et al.* (2017)	*Validation of multisource electronic health record data: an application to blood transfusion data*	Apresenta um *framework* para validação de dados *Linked Data* provenientes de fontes diversas, focados em dados eletrônicos do sistema de saúde.
Acosta *et al.* (2018)	*Detecting Linked Data quality issues via crowdsourcing: a dbpedia study*	Propõe uma solução para problemas complexos de qualidade baseada em abordagem *crowdsourcing*.
Albertoni *et al.* (2018)	*Quality measures for SKOS: ExactMatch linksets: an application to the thesaurus framework LusTRE*	Propõe uma ferramenta (LusTRE) para avaliar a qualidade da conexão entre tesauros publicados como *Linked Data.*
Braşoveanu et al. (2018)	*Framing Named Entity Linking Error Types*	Propõe uma taxonomia para identificação e correção de problemas de qualidade em dados *Linked Data*, baseada na exploração dos *"Named Entity Linking (NEL)".*
Langer *et al.* (2018)	*SemQuire - assessing the data quality of Linked*	Propor uma ferramenta (*SemQuire*) para auxiliar na condução do processo de

	Open Data sources based on dqv	avaliação de qualidade, tendo como proposito permitir o enriquecimento de dados coorporativos.
Langer e Gaedke (2018)	*DaQAR - an ontology for the uniform exchange of comparable Linked Data quality assessment requirements*	Propõe uma ontologia (*DaQAR Approach*) focada na avaliação de qualidade, visando auxiliar na comparação diferentes conjuntos de dados e no processo de seleção para diferentes aplicações.
Liu *et al.* (2018)	*A new truth discovery method for resolving object conflicts over Linked Data with scale-free property*	Propõe uma ferramenta (*TruthDiscover*) para a identificação e resolução de problemas de conflitos de objetos em dados *Linked Data.*
Mihindukulasooriya e Rico (2018)	*Type Prediction of RDF Knowledge Graphs Using Binary Classifiers with Structural Data*	Propõe uma metodologia usando técnicas de *Machine Learning* para predição de tipos de informação, visando melhoria na qualidade de dados *Linked Data.*
Odoni *et al.* (2018)	*On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance*	Apresenta um *framework* para análise de qualidade baseado em "*Named Entity Linking* (NEL)"
Rico *et al.* (2018)	*Predicting incorrect mappings: a data-driven approach applied to dbpedia*	Propõe uma ferramenta baseada em técnicas de *machine learning* para promover melhoria de qualidade de dados legados.
Ruan *et al.* (2018)	*On Evaluating Web-Scale Extracted Knowledge Bases in a Comparative Way*	Elabora uma metodologia para avaliar a qualidade de dados *Linked Data* chineses, com foco nas dimensões *Richness* e *Correctness*.
Abián *et al.* (2019)	*Using contemporary constraints to ensure data consistency*	Apresenta uma metodologia baseada em "*cross-comparing*" para identificação de inconsistências em conjuntos de dados *Linked Data.*
Arruda *et al.* (2019)	*A Fuzzy Approach for Data Quality Assessment of Linked Datasets*	Apresenta uma ontologia (*Fuzzy*) para representar os resultados do processo de avaliação de qualidade de dados *Linked Data*, baseada no W3C *Data Quality Vocabulary.*
Elbattah e Ryan (2019)	*Learning Sequence Patterns in Knowledge Graph Triples to Predict Inconsistencies*	Propõe uma metodologia para identificação de inconsistências em conjuntos de dados *Linked Data.*
Gürdür, El-Khoury e	*Methodology for linked enterprise data quality*	Propõe uma metodologia para a integração e avaliação de

Nyberg (2019)	*assessment through information visualizations*	qualidade de dados *Linked Data* para aplicação no contexto empresarial.
Heling (2019)	*Quality-Driven Query Processing over Federated RDF Data Sources*	Propõe uma metodologia para avaliação de qualidade de dados resultantes da integração de fontes heterogêneas
Rashid *et al.* (2019)	*A quality assessment approach for evolving knowledge bases*	Propõe um artefato para a análise de qualidade centrado na etapa de validação, usando perfis de dados para identificação de alterações e problemas de qualidade.
Haller *et al.* (2020)	*What are links in Linked Open Data? a characterization and evaluation of links between knowledge graphs on the web*	Propõe uma metodologia para realizar avaliação quantitativa de conjuntos de dados por meio da análise dos *links.*
Huang *et al.* (2020)	*An RDF Data Set Quality Assessment Mechanism for Decentralized Systems*	Propõe um modelo de qualidade baseado em tecnologia de armazenamento da *blockchain* visando permitir o compartilhamento e a auditoria dos resultados da avaliação de qualidade.
Homburg (2020)	*Connecting Semantic Situation Descriptions with Data Quality Evaluations—Towards a Framework of Automatic Thematic Map Evaluation*	Propõe uma ontologia e modelo para seleção de dimensões e métricas adequadas para o processo de seleção de dados *Linked Data*.
Albertoni, Martino e Quarati (2021)	*Documenting Context-Based Quality Assessment of Controlled Vocabularies*	Propõe uma metodologia para avaliação de qualidade de controle de vocabulário para dados governamentais publicados como *Linked Data.*
Hanlon *et al.* (2021)	*Towards an effective user interface for data exploration, data quality assessment and data integration*	Propõe uma ferramenta para a realização de exploração, integração e avaliação de qualidade de dados *Linked Data.*
Mangel *et al.* (2021)	*Data Reliability and Trustworthiness Through Digital Transmission Contracts*	Propõe um artefato para avaliação de qualidade de dados *Linked Data* focado em dados industriais e em casos em que existe uma alta demanda de confiabilidade de dados.

Fonte: Autora (2025)

A análise do quadro ressalta a multiplicidade e heterogeneidade dos artefatos existentes para auxiliar no processo de avaliação e qualidade, mesmo quando

restringidos apenas a artefatos focados em dados publicados como *Linked Data*. Existem diferentes tipos de artefatos, com propósitos e públicos-alvo distintos, que diferem entre si tanto nas atividades que desempenham, como na forma como desempenham essas atividades.

A categoria "2 - Realiza um estudo de avaliação de qualidade em um ou mais conjuntos de dados" reúne documentos cuja proposta é avaliar os níveis de qualidade de um ou mais *datasets*, aplicando dimensões e métricas. O quadro 8 apresenta os artigos incluídos nessa categoria, seguidos de uma breve descrição.

Quadro 8 - Documentos aceitos incluídos na categoria "2 - Realiza um estudo de avaliação de qualidade em um ou mais conjuntos de dados".

AUTORES	TÍTULO	DESCRIÇÃO
Font, Zouaq e Gagnon (2015)	*Assessing the quality of domain concepts descriptions in dbpedia*	Realiza a avaliação de qualidade da descrição de conceitos de domínio nos dados da DBpedia
Mihindukulasooriya *et al.* (2015)	*An analysis of the quality issues of the properties available in the spanish DBpedia*	Realiza uma avaliação de qualidade nos dados da *DBpedia*, focada na dimensão *concisão.*
Tallerås (2017)	*Quality of linked bibliographic data: the models, vocabularies, and links of data sets published by four national libraries.*	Realiza a avaliação de qualidade dos dados qualidade dos dados bibliográficos publicados como *Linked Data* pelas bibliotecas nacionais da Espanha, França, Reino Unido e Alemanha.
Abián *et al.* (2018)	*Wikidata and DBpedia: A comparative study*	Realiza a análise de qualidade de dois conjuntos de dados derivados da *Wikipedia*
Debattista *et al.* (2018)	*Evaluating the quality of the LOD cloud: an empirical investigation*	Utiliza o *software* Luzzo para realizar uma avaliação extensiva da qualidade de dados registrados na *Linked Open Data Cloud* (avaliando 130 conjuntos de dados).
Healy *et al.* (2018)	*The accuracy of chemotherapy ascertainment among colorectal cancer patients in the surveillance,*	Realiza a avaliação de qualidade de um conjunto de dados relacionado a quimioterapia, focado na dimensão acurácia.

	epidemiology, and end results registry program	
Harper (2018)	*Linkage of maternity hospital episode statistics data to birth registration and notification records for births in england 2005–2014: quality assurance of linkage of routine data for singleton and multiple births*	Realiza a análise de qualidade de um conjunto de dados que registra o nascimento e notificação registro de nascimento todos os nascimentos na Inglaterra 2005 – 2014.
Ibáñez *et al.* (2019)	*An assessment of adoption and quality of Linked Data in european open government data*	Realiza uma avaliação de qualidade quantitativa dos conjuntos de dados governamentais publicados como *Linked Data* e indexados pelo de dados indexados pelo Portal Europeu de Dados (EDP).
Kahlawi (2020)	*An ontology driven ESCO LOD quality enhancement*	Realiza a avaliação de conjunto de dados relacionados a uma ontologia criada pela União Européia para a descrição de dados relacionados ao mercado de trabalho (*European Skills, Competences, Qualifications and Occupations* - ESCO LOD).
Kamdar e Musen (2021)	*An empirical meta-analysis of the life sciences Linked Open Data on the web*	Realiza uma análise de dados provenientes das ciências da vida publicados como *Linked Data*, visando verificar a heterogeneidade e qualidade desses dados.
Varmdal *et al.* (2021)	*Data from national health registers as endpoints for the tromso study: correctness and completeness of stroke diagnoses*	Realiza uma análise de qualidade focada nas dimensões *correctness and completeness* e compara conjuntos de dados com base nessa análise.

Fonte: Autora (2025)

A partir da análise do quadro, é possível identificar estudos voltados para avaliar apenas um conjunto de dados, estudos com o propósito de avaliar e comparar dois ou mais conjuntos de dados.

Entre esses estudos, destaca-se uma abordagem contextual, sendo a maioria voltados para avaliar conjuntos de dados de domínios específicos, como dados governamentais, dados da área da saúde, dados bibliográficos e enciclopédicos. Também se destacam estudos focados em analisar apenas uma ou duas dimensões de qualidade, como as dimensões correção e completude, consistência e acurácia.

A identificação dos estudos da categoria "2 - Realiza um estudo de avaliação de qualidade em um ou mais conjuntos de dados" é importante especialmente para o processo de construção do fluxo para a seleção de dados *Linked Data*, pois esses estudos permitem a observação das etapas, decisões e ferramentas utilizadas pelos autores para realizar a avaliação de dados *Linked Data*.

A última categoria, "3 - Levantamentos e estudos teóricos sobre qualidade de dados e *Linked Data*" é uma categoria mais diversa, que reúne diferentes levantamentos bibliográficos e discussões teóricas sobre a qualidade de dados *Linked Data*. O quadro 9 apresenta os estudos reunidos nessa categoria

Quadro 9 - Documentos aceitos incluídos na categoria "3 - Levantamentos e estudos teóricos sobre qualidade de dados e *Linked Data*"

AUTORES	TÍTULO	DESCRIÇÃO
Behkamal *et al.* (2014)	*Data Accuracy: What does it mean to LOD?*	Realiza um levantamento para identificar as métricas necessárias para mensurar os aspectos sintáticos e semânticos de qualidade de dados *Linked Data*, focado na acurácia dos dados.
Isaac e Baker (2015)	*Linked Data Practice at Different Levels of Semantic Precision: The Perspective of Libraries, Archives and Museums*	Discute os diferentes níveis de precisão semântica através da perspectiva dos vocabulários controlados.
Zaveri *et al.* (2015)	*Quality assessment for Linked Data: A Survey*	Apresenta os resultados de uma Revisão Sistemática

		da Literatura sobre melhorias de qualidade em dados *Linked Data*.
Sapna, Rani e Mishra (2018)	*An Investigative Study on the Quality Aspects of Linked Open Data*	Realiza um levantamento sobre a questão de qualidade no contexto do *Linked Data*, abordando desafios do processo de avaliação.
Hadhiatma (2018)	*Improving data quality in the linked open data: a survey*	Realiza um levantamento sobre métodos a serem aplicados no processo de avaliação de dados *Linked Data.*
Possemato (2018)	*How RDA is essential in the reconciliation and conversion processes for quality Linked Data*	Discutir como o *Resource Description and Access* (RDA) pode melhorar a qualidade de dados bibliográficos publicados como *Linked Data*.
Catania, Guerrini e Yaman (2019)	*Exploiting Context and Quality for Linked Data Source Selection*	Discute o papel de contexto e da qualidade no processo de seleção de fontes de dados *Linked Data*, levantando ferramentas que possam ser aplicadas para avaliação de qualidade.
Bailey *et al.* (2020)	*How Well Do Automated Linking Methods Perform? Lessons from US Historical Data*	Realiza uma revisão de literatura a respeito da qualidade de dados provenientes de métodos automáticos de ligação.
Issa *et al.* (2021)	*Knowledge Graph Completeness: A Systematic Literature Review*	Realiza uma Revisão Sistemática da Literatura sobre a dimensão completude no contexto de dados *Linked Data*.

Fonte: Autora (2025)

Os estudos incluídos nessa categoria são fundamentais para a compreensão aprofundada do contexto da qualidade de dados *Linked Data*, compreensão dos desafios, identificação de metodologias e ferramentas que possam auxiliar no processo de seleção de fontes de dados *Linked Data*.

Os textos dessa categoria também se destacam em relação ao processo de identificação de definições e da complementação do *corpus* através da técnica de *Snowballing*. Embora todos os documentos identificados pela RSL tenham sido considerados para esse fim, os dessa categoria, por sua abordagem teórica e característica de serem levantamentos (em alguns casos, inclusive Revisões Sistemáticas) foram importantes para a identificação de autores basilares e para a identificação dos termos relevantes para o entendimento da qualidade de dados *Linked Data.*

Com base nesse *corpus* teórico, buscou-se identificar os principais desafios de qualidade que afetam dados publicados como *Linked Data*, a próxima subseção apresenta os resultados dessa análise.

4.3 Problemas de qualidade em dados *Linked Data*

Para a análise dos desafios de qualidade relacionados a dados *Linked Data,* foram considerados todos os artigos que compõe o *corpus* da RSL. Ao longo da leitura flutuante dos artigos aceitos foram coletados os problemas de qualidade mencionados pelos autores, em seguida foram construídas categorias *a posteriori* para a sistematização dessas problemáticas.

As problemáticas identificadas puderam ser agrupadas em três principais categorias, sendo elas: 1 - Desafios relacionados às características das fontes de dados; 2 - Desafios relacionados com a estrutura dos dados *Linked Data*; e 3 - Desafios relacionados ao processo de avaliação de qualidade.

Os problemas de qualidade incluídos na categoria 1 - Desafios relacionados às características das fontes de dados, abordam problemas de qualidade associados a origem dos dados, como a heterogeneidade dessas fontes e os desafios do processo de seleção de fontes de dados. O quadro 10 apresenta a relação de subcategorias e o número de artigos que mencionam essa problemática.

Quadro 10 - Problemas de qualidade relacionados as fontes de dados

Subcategoria	Descrição	N° de artigos
Heterogeneidade das fontes	As fontes de dados *Linked Data* são diferentes em muitos aspectos, elas são provenientes de domínios distintos, criam seus dados com propósitos diferentes. Como a qualidade é majoritariamente abordada em um aspecto contextual, um dado considerado de qualidade para uma fonte pode não ser de qualidade para outra fonte.	14
Diferentes níveis de Curadoria de dados	Uma consequência da heterogeneidade das fontes são os diferentes níveis de curadoria. Existindo dados que passam por um processo rigoroso, sistemático e constante de avaliação e dados que apenas são convertidos para *Linked Data* de forma automática, e nunca passam por processos de revisão e atualização.	13
Dificuldade em selecionar as melhores fontes de dados para ligação	Justamente por essa diversidade, existe uma dificuldade em selecionar as melhores fontes de dados, incluindo a incapacidade de encontrar e usar dados, que se relaciona à problemas de acessibilidade desses dados.	5
Aspectos de temporalidade	Com os diferentes níveis de curadoria, existem problemas relacionados com a frequência de atualização dos conjuntos de dados, que precisa ser avaliada para determinar se os dados estão suficientemente atualizados para a aplicação pretendida.	2
Desafios relacionados a dados provenientes de *crowndsourcing*	Esses dados têm um grande impacto em volume na nuvem de dados *Linked Data*, e a forma como eles são criados e mantidos, de maneira colaborativa, apresenta desafios específicos que precisam ser considerados no processo de avaliação de qualidade.	4
Impacto do uso de fontes externas na qualidade do conjunto de dados	Por toda a diversidade mencionada, existe ainda uma preocupação com o impacto dos dados provenientes de fontes externas na qualidade dos conjuntos de dados	1

Fonte: autora (2025)

É possível observar por sua descrição, que as subcategorias identificadas são interrelacionadas e que possuem em comum a problemática da heterogeneidade e variedade existente entre as fontes de dados *Linked Data.* Como a qualidade de dados muitas vezes só pode ser mensurada com uma abordagem contextual, a diversidade das fontes implica em um desafio para o processo de seleção de dados para ligação.

O *Linked Data* se baseia justamente nesse processo de ligação, tornando a preocupação sobre como selecionar fontes de dados e como esses dados de fontes externas vão impactar na qualidade final dos dados publicados de grande relevância para o cenário analisado.

A categoria 2 reúne os desafios relacionados com a estrutura dos dados *Linked Data*. O quadro 11 apresenta esses desafios:

Quadro 11 - Problemas de qualidade relacionados com a estrutura de dados *Linked Data*

Subcategoria	Descrição	N° de artigos
Conversão automática e semiautomática de dados legados	O processo de conversão de dados utilizando ferramentas automática e semiautomáticas é comumente adotado na publicação de dados *Linked Data*. Esse processo não é livre de erros, e quando os dados resultantes não são auditados, ou fazem uso apenas de técnicas de validação automáticas, nem sempre eficientes, a qualidade dos dados pode ser comprometida	21
Heterogeneidade dos dados	Assim como são heterogêneas as fontes de dados, os dados publicados como *Linked Data* também são estruturalmente muito diversos. São adotados diferentes vocabulários, os conjuntos de dados possuem tamanhos, níveis de granularidade, e constância de atualização diferentes, o que pode dificultar o processo de avaliação, seleção e reutilização desses dados.	9
Inconsistência no conteúdo dos dados	Os dados podem ser mal interpretados, redundantes, incompletos, inconsistentes ou incorretos e esses aspectos precisam ser considerados no processo de avaliação de qualidade.	14
Dados conflitantes	Essa problemática está relacionada aos desafios de lidar com dados de valores conflitantes em uma mesma fonte ou em fontes diversas. Como os dados *Linked Data* estão em ambiente aberto, diferentes fontes podem trazer dados conflitantes sobre uma mesma entidade ou objeto. Também pode ocorrer inconsistências entre objetos e predicados em relação as entidades do mundo real.	7
Dinamicidade dos dados *Linked Data*	Os dados estão constantemente sujeitos a mudanças e atualizações, existindo casos em que a completude é essencial e outros em que é inviável, fazendo com que a avaliação de qualidade precise ser um processo constante. Essa dinamicidade também leva a preocupações com a permanência e acessibilidade desses dados, que podem ser excluídos, modificados ou se tornarem inacessíveis sem comunicação prévia.	9

Emprego incorreto de propriedades, vocabulários e ontologias	O uso incorreto das propriedades, vocabulários e ontologias pode levar a dados conflitantes ou incorretos, resultando em problemas de qualidade que precisam ser identificados.	5
Dificuldade na identificação da proveniência das informações	Com o enriquecimento semântico das fontes, os dados de diferentes fontes são combinados dificultando assim a identificação da origem dos dados	3
Aspectos relacionados aos *Links*	Os *links* são parte da estrutura de dados *Linked Data* e podem levar a problemas como *links* quebrados, redundantes, mudanças inesperadas em *links* externos, dificuldade de diferenciar *links* internos e externos	5
Dificuldades de integração	A heterogeneidade da estrutura dos dados, já mencionada em outra subcategoria, pode dificultar o processo de integração de dados de diferentes fontes, especialmente quando se trata do uso de vocabulários distintos	4
Ausência da avaliação de qualidade nos ciclos e vida de dados	Existe uma discrepância entre a preocupação com a publicação e a manutenção dos dados *Linked Data*, onde muitas vezes é adotada uma política de "publicar primeiro e corrigir depois", mas que resulta em dados que são publicados e não recebem a manutenção necessária para que possam ser reutilizados sem causar problemas de qualidade	5
Aplicação dos dados em contextos diferentes dos que foram criados	Dados *Linked Data* muitas vezes são aplicados em um contexto diferente e com propósitos diferentes dos quais foram criados, gerando problemas de qualidade que não poderiam ser previstos pelos publicadores, por não serem pertinentes no contexto original.	2
Heterogeneidade no nível da entidade	A estrutura de dados *Link Data* prevê múltiplos URIs para um mesmo objeto do mundo real, o que pode ser um desafio para o processo de avaliação de qualidade	1
Qualidade dos literais	A estrutura do RDF permite que literais, textos em linguagem natural, sejam utilizados como valor das declarações. Muitas vezes a checagem da qualidade e da consistência desses literais é negligenciada	1
Questões relacionadas com a estrutura em RDF	A base da estrutura de dados *Linked Data* é o RDF, problemas com a estruturação de dados nesse formato podem afetar a qualidade final do conjunto de dados.	1
Ausência ou insuficiência de restrições lógicas	As bases de conhecimento de *Linked Data* contêm apenas algumas restrições lógicas ou não são bem modeladas.	1
Possibilidade de criação de novos dados a partir de	O uso do *SPARQL* permite a coleta, atualização e disponibilização de fragmentos e dados, que podem gerar problemas de qualidade	1

fragmentos SPARQL		

Fonte: autora (2025)

Se destaca na literatura uma preocupação com o processo de conversão de dados legados (dados publicados anteriormente em outros formatos), sobre como avaliar a qualidade dos dados convertidos utilizando ferramentas automáticas e semiautomáticas, identificar e corrigir problemas derivados desse processo.

Também são abordados problemas relacionados com a heterogeneidade estrutural de dados, levando em consideração os diferentes vocabulários adotados, considerando os potenciais conflitos e incompatibilidade entre eles, ou a sua aplicação incorreta.

Foram mencionados ainda desafios relacionados ao próprio processo de avaliação de qualidade, incluídos na categoria 3. O quadro 12 apresenta as subcategorias relacionadas a esses desafios.

Quadro 12 - Problemas de qualidade relacionados com o processo de avaliação de qualidade de dados *Linked Data*

Subcategoria	Descrição	N° de artigos
Ausência de metadados descritivos adequados	A ausência de representação adequada dos conjuntos de dados, especialmente disponibilizada de maneira formal, dificulta o processo de avaliação de qualidade, fazendo com que sejam necessários vários artifícios para obtenção de informações necessárias para essa avaliação, como por exemplo: a proveniência, o conteúdo, frequência de atualização e etc	10
Volume, Variedade e Velocidade com que são gerados os dados	Os dados *Linked Data* muitas vezes podem ser associados a um contexto de B*ig Data*, e esses aspectos tornam o processo de avaliação mais complexo, especialmente quando o tempo disponível para esse processo é curto devido a rápida obsolescência dos conjuntos de dados	9
Limitações das ferramentas de avaliação existentes	Nem sempre estão disponíveis as ferramentas adequadas e necessárias para o processo de avaliação de qualidade, muitas das ferramentas existentes dependem de interferência humana. Dependendo do tamanho do conjunto de dados fornecido, o processo de avaliação manual por especialistas do domínio será demorado e dispendioso. As ferramentas	6

	muitas vezes são criadas para um domínio específico e não podem ser adaptadas	
Discrepâncias entre as necessidades dos usuários de dados *Linked Data*	Essas diferentes necessidades dificultam a criação de soluções genéricas para o processo de avaliação de qualidade, tornando necessárias as abordagens contextuais focadas em domínios, ou mesmo em aplicações específicas	2
Dificuldade de comparação dos resultados de diferentes processos de avaliação de qualidade	Os resultados das ferramentas existentes muitas vezes são exportados em formatos e estruturas que dificultam uma comparação direta entre esses resultados	2
Ausência de padrões oficiais para analisar a qualidade de dados *Linked Data*	Embora o W3C disponibilize um vocabulário para avaliação de qualidade, ele não recomenda diretamente dimensões, critérios e métricas para o processo de avaliação, o que dificulta a realização padronizada do mesmo	2
Dificuldade sobre como estabelecer dimensões e métricas para um domínio	Ausência de orientações sobre como selecionar as dimensões e métricas, bem como o peso para cada uma delas no contexto do *Linked Data* dificulta a condução de processos de avaliação de qualidade	1
Alguns problemas só são detectáveis no momento da aplicação dos dados	Mesmo quando os publicadores realizam a avaliação, a maior parte dos problemas de inconsistência só se tornam obvias quando os dados são processados	1
Incompatibilidade entre o tempo disponível pra avaliação e a obsolescência dos dados	O processo de seleção e curadoria das fontes é longo, quando esse processo é concluído os dados podem já estar desatualizados	1
Avaliações de qualidade majoritariamente disponibilizadas pelo provedor	Geralmente o nível de qualidade é determinado do ponto de vista do provedor, existindo uma defasagem de bases de avaliação que levem em conta a perspectiva dos consumidores	1

Fonte: autora (2025)

Nessa categoria se destacam as limitações das ferramentas e as dificuldades relacionadas a criação de novas ferramentas, como discrepância nas necessidades dos usuários, ausência de orientações oficiais, dificuldade de comparação dos resultados do processo de avaliação.

A compreensão aprofundada dessas problemáticas é importante para o objetivo geral da presente pesquisa, não só para justificar a necessidade do artefato que será elaborado como produto, mas também para embasar a construção do mencionado artefato, possibilitando que esses aspectos sejam levados em consideração no momento da sua elaboração.

Discutidos os principais problemas de qualidade que afetam dados publicados como *Linked Data*, buscou-se compreender melhor os artefatos que podem auxiliar no processo de avaliação e melhoria da qualidade desses dados. A próxima subseção apresenta os resultados dessa análise.

4.4 Ferramentas para avaliação de qualidade de dados *Linked Data*

Para compor o *corpus* de coleta a respeito das ferramentas, foram considerados os artigos aceitos na RSL e incluídos na categoria “Propõe um artefato para avaliação ou melhorias de qualidade em dados publicados como *Linked Data*”. O *corpus* foi submetido então a uma leitura flutuante, onde se observou que os artefatos, embora tenham em comum a busca pela avaliação de qualidade de dados *Linked Data*, são diversos entre si, desde o objetivo, até as atividades que realizam e como realizam essas atividades.

Com base nas descrições dos artefatos apresentados, elaborou-se uma síntese das principais características individualizadoras, que permitem diferenciar os artefatos existentes para auxiliar no processo de avaliação de qualidade de dados *Linked Data*, o quadro 13 apresenta essas características.

Quadro 13 - Características individualizadoras para artefatos que auxiliam na avaliação de dados *Linked Data*

Característica dos artefatos	Descrição
Tipo	Foram adotados diferentes termos para descrever os artefatos propostos e apresentados para auxiliar na qualidade de dados Linked Data. Destacaram-se: *frameworks*; *softwares*; métodos/metodologias; ontologias; vocabulários; instrumentos baseados em inteligência artificial/machine learning; modelos; Instrumentos baseados em *crowdsourcing;* guias para avaliação e melhoria de qualidade; fluxos de trabalho; metodologias complexas (termo utilizado para se referir a abordagens compostas por mais de um artefato);

Atividade	Em relação às atividades que desempenham, os artefatos se dividem em quatro vertentes: **Avaliação** – artefatos que se propõe a identificar problemas de qualidade, mensurar os níveis de qualidade ou ainda comparar os níveis de qualidade entre dois ou mais *datasets*; **Melhoria** – artefatos voltados para a correção de problemas específicos de qualidade de dados; **Avaliação e melhorias** – artefatos que desempenham as duas etapas, gerando relatórios de qualidade e corrigindo os problemas identificados; **Comunicação dos resultados** – artefatos de abordagem representacional que visam facilitar a comunicação e o reuso dos resultados do processo de avaliação de qualidade
Forma como desempenha a atividade	Quanto à forma como desempenham as atividades propostas, os artefatos podem ser divididos em: **Automáticos -** que não necessitam de interferência humana durante todo o processo; **Semiautomáticos -** que realizam parte do processo de maneira automática, mas necessitam de decisão humana em determinadas etapas; **Manuais -** são artefatos que atuam como guias para orientar a avaliação, dependendo da atividade de pessoas especializadas ou de usuários, podendo ainda serem insumos para a elaboração de artefatos automáticos e semiautomáticos.
Dimensões	Em relação a forma como abordam as dimensões, os artefatos se diferem entre **Específicos** – cujo funcionamento é direcionado para avaliar apenas uma dimensão de qualidade; **Multidimensionais** – cujo funcionamento abrange uma série de dimensões, existindo, inclusive, artefatos que permitem a customização das dimensões a serem utilizadas no processo de avaliação.
Domínio	Em relação ao domínio para os quais foram criados, os artefatos variam entre: **Gerais -** que podem ser aplicados em conjuntos de dados provenientes de diversas áreas; **Específicos** – criados visando atender a dados provenientes de um domínio específico, como dados da área da saúde ou dados governamentais.
Público a que se destina	Quanto ao público a que se destinam os artefatos se diferem em três: **Voltados para os publicadores de dados** – ou seja para que esses realizem uma avaliação dos próprios dados e possam promover melhorias de qualidade visando atender de maneira mais satisfatória a demanda de seus consumidores; **Voltados para os consumidores de dados** – ferramentas pensadas para auxiliar na seleção de

	datasets LD tanto para aplicações como para ligação entre *datasets* **Abrangentes** – que não especificam um público-alvo.

Fonte: Autora (2025)

A identificação dessas características individualizadoras será importante para a construção do glossário e do fluxo para a seleção de dados *Linked Data*, pois compreender os aspectos que individualizam esses artefatos auxilia tanto na sua seleção para diferentes contextos como na sua adaptação ou na criação de novos artefatos para atender demandas específicas.

As categorias foram estabelecidas *a posteriori*, tendo como base o problema inicial que levou ao desenvolvimento do artefato, seus objetivos e a as atividades que ele desempenha, sendo sempre priorizado o objetivo central do artefato para a categorização.

Foram considerados na análise 74 artigos, distribuídos em 11 categorias. As categorias e o número de artigos cujos artefatos podem ser nelas incluídos são apresentados no quadro 14:

Quadro 14 - Categorias e número de artigos por categoria

N° da categoria	Categoria	N° de artigos incluídos
1	Artefatos focados em vocabulários, ontologias, tesauros e metadados descritivos	13
2	Artefatos que são baseados em adaptação de técnicas de outros domínios	13
3	Artefatos que buscam explorar características estruturais de dados *Linked Data*	12
4	Artefatos focados em categorias e dimensões específicas	7
5	Artefatos focados em um domínio específico	5
6	Artefatos voltados para seleção e curadoria de fontes para a ligação	4
7	Artefatos para avaliação geral de qualidade de dados *Linked Data*	4
8	Artefatos *voltados para a c*onversão de dados legados	3
9	Artefatos voltados para a identificação e solução de conflito de objetos	2
10	Artefatos que realizam a adaptação de um outro artefato já existente	1
11	Artefatos focados na Integração de dados	1

Fonte: Autora (2025)

A análise do quadro permite observar que existe um destaque para artefatos de organização e representação, como vocabulários, ontologias, tesauros e metadados descritivos.

Dentro dessa categoria, os vocabulários são abordados de três maneiras distintas: 1) como um aspecto da qualidade – onde o objetivo do artefato é verificar a aplicação correta dos vocabulários no processo de formação de triplas em RDF; 2) Como um meio para checagem de outros aspectos de qualidade – onde esses vocabulários e metadados fornecem informações a respeito dos conjuntos de dados, que podem ser exploradas na realização do processo de avaliação e; 3) Como um meio para compartilhamento formal dos resultados do processo de avaliação de qualidade – o que permite que esses resultados sejam reaproveitados e processados com maior facilidade por usuários máquina.

Também se destacam os artefatos que buscam a adaptação de técnicas já consolidadas em outros domínios para viabilizar o processo de avaliação, sendo exemplos de técnicas: o *test-drive* proveniente da engenharia de *software*, uso de técnicas de *Machine Learning*, aplicação de técnicas de *data profile*, exploração das possibilidades de aplicação de *crowdsourcing,* exploração da *blockchain* etc.

Observa-se que foram muito explorados os aspectos estruturais de dados *Linked Data*, onde os artefatos buscam avaliar problemas de qualidade que ocorrem por conta das especificidades estruturais desses dados (como a utilização do RDF e a estrutura de *links*) ou se utilizam dessas características para a criação do artefato, como na utilização do SPARQL para identificar problemas de qualidade ou da propriedade *owl:sameAs* para identificar anomalias.

Observou-se ainda que mesmo quando os vocabulários não são o foco principal dos artefatos, parte dos artefatos automáticos e semiautomáticos se utilizam de vocabulários em alguma etapa de seu funcionamento, seja para agilizar o processo de análise, coletando metadados que auxiliem no processo de avaliação, seja para exportar os relatórios gerados.

Discutidos os aspectos teóricos da qualidade de dados *Linked Data*, a próxima seção apresenta os resultados do estudo terminológico da qualidade de dados.

5 ESTUDO TERMINOLÓGICO DA QUALIDADE DE DADOS *LINKED DATA*

A presente seção apresenta os resultados do estudo terminológico realizado a respeito da qualidade de dados *Linked Data* visando permitir a construção do um glossário.[6] O estudo buscou ainda contribuir com a clareza conceitual necessária para a discussão dos aspectos processuais da seleção de dados *Linked Data*.

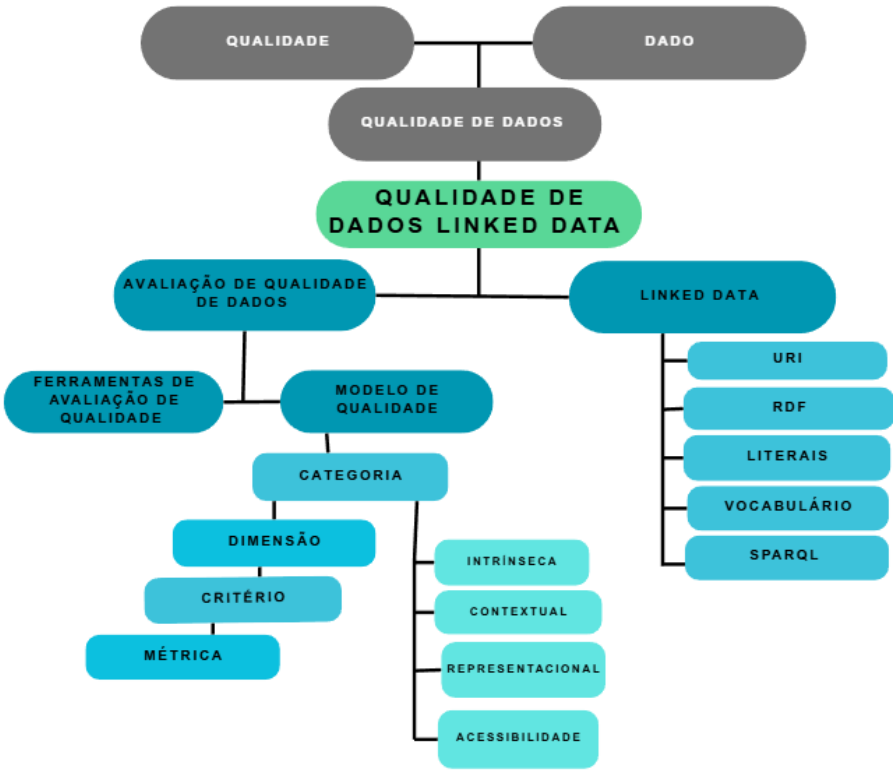
A base para a condução do estudo foi o *corpus* teórico e documental identificado nas etapas anteriores da pesquisa. O estudo foi conduzido em duas etapas principais: identificação/seleção dos termos e elaboração das definições.

A seleção dos termos seguiu a estrutura da árvore de domínio, e foi baseado na análise da literatura de qualidade de dados *Linked Data* e da documentação oficial do W3C, sendo identificados e organizados de maneira hierárquica os termos necessários para compreensão da qualidade de dados *Linked Data*.

Com base no estudo terminológico e seguindo metodologia para construção de Árvores de Domínio foram selecionados 22 termos para compor o glossário, apresentados na figura 8.

[6] Considerando a sua extensão e estrutura, optou-se pela disponibilização das grades como dados de pesquisa, disponibilizado no repositório institucional da UNESP.

Figura 8 - Árvore de domínio da qualidade de dados *Linked Data*



Fonte: Autora (2025)

A construção das definições foi baseada na garantia literária, tendo como fonte a Revisão Sistemática da Literatura. Nos casos em que informações adicionais a respeito do termo se mostraram necessárias, foram realizadas pesquisas exploratórias e aplicada técnica *Snowballing*. Para o estabelecimento dos termos relacionados com o *Linked Data*, partiu-se principalmente da garantia organizacional, baseada na documentação oficial do W3C.

As definições foram construídas por meio da fragmentação das definições identificadas, seguindo a estrutura da Grade. Foi criada uma grade para cada um dos termos. [7]

Os termos foram divididos em grupos, para facilitar a sua discussão e apresentação, sendo eles: termos relacionados a qualidade de dados, termos relacionados ao *Linked Data* e avaliação de qualidade de dados *Linked Data*.

[7] Considerando a sua extensão e estrutura, as grades foram disponibilizadas como dados de pesquisa. Podem ser acessadas no repositório da UNESP ou por meio do *link*: https://drive.google.com/drive/folders/1q4F9vNkoWHlTjExlzUfAoGMCbU2AFKRG?usp=sharing

Para uma melhor compreensão, a presente subseção foi estruturada de maneira a refletir essa organização, onde cada subseção apresenta as informações necessárias para a compreensão das definições cunhadas e uma sistematização em formato de figura, que serviu como auxílio no processo de elaboração da definição.

5.1 Termos relacionados à Qualidade de dados

A presente subseção reúne os termos iniciais necessários para a compreensão da Qualidade de dados *Linked Data*, sendo eles: “qualidade”; “dado” e “qualidade de dados”.

Muitas discussões e discordâncias circundam a definição do termo qualidade (Juran *et al.*,1999). O conceito de dado também não está livre de discussões e discordâncias. Redman (2013) aponta que a discussão do termo envolve um considerável nível de desacordos, sendo objeto de discussão de filósofos, cientistas da computação, estatísticos e outros grupos ao longo de gerações.

Santos e Sant’ana (2013, p. 201) acrescentam ainda que “Muitas foram as tentativas de definir o conceito de dado, e uma das dificuldades inerentes à definição reside na condição do termo e do próprio dado serem utilizados de maneiras distintas por praticamente todas as áreas da Ciência.”

Portanto, os termos qualidade e dados são de complexa definição, não existindo um consenso na literatura a respeito de como defini-los e delimitá-los. Esse desafio se justifica, em partes, por que ambos são termos cunhados em contextos interdisciplinares, sendo definidos de maneira fortemente contextual, com variações significativas em sua definição a depender da área e dos autores adotados.

Buscou-se então discutir cada um desses conceitos, visando a construção de uma definição significativa para o contexto da qualidade de dados, derivada do discurso da comunidade. As próximas subseções representam cada um dos termos, sendo compostas pelas discussões, sistematização e definição.

5.1.1 Qualidade

Em relação ao termo qualidade, destacam-se as definições propostas por Juran *et al.* (1998). Os autores apontam que embora não exista um consenso, no âmbito dos processos de avaliação e melhoria de qualidade, o termo qualidade pode ser encarado com base em duas definições, que se resumem em: qualidade como

características dos produtos que atendem às necessidades do cliente ou como ausência de deficiências.

A qualidade enquanto “características dos produtos que atendem às necessidades do cliente” possui uma abordagem contextual, e é definida pelos autores como:

> “Qualidade” significa aquelas características dos produtos que atendem às necessidades do cliente e, portanto, fornecem satisfação ao cliente. Neste sentido, o significado de qualidade está orientado para o rendimento. O propósito de tal qualidade superior é fornecer maior satisfação do cliente e, espera-se, aumentar a renda. No entanto, fornecer mais e/ou melhores características de qualidade geralmente requer um investimento e, portanto, geralmente envolve aumentos nos custos. Maior qualidade nesse sentido geralmente “custa mais” (Juran *et al.,* 1998, p. 2.1).

Melo (2017, p. 47) também define a qualidade a partir de uma abordagem contextual, como um sinônimo de adequação ao uso.

> Qualidade pode ser encarada como um conceito subjetivo sobre a percepção de um indivíduo em relação a um serviço, produto, dado, informação etc. De modo geral, pode ser definida como medidas para que o produto oferecido esteja de acordo com o que se espera dele, podendo este ser uma informação, um dado, um serviço ou um processo.

A definição de qualidade como um sinônimo de adequação ao uso, “*fitness for use*”, é amplamente adotada pelo domínio da qualidade de dados (Juran *et al.*, 1988; Wang; Strong,1996).

> O conceito de "adequação para uso" é agora amplamente adotado na literatura de qualidade. Ele enfatiza a importância de adotar um ponto de vista do consumidor sobre a qualidade porque, em última análise, é o consumidor que julgará se um produto é ou não adequado para uso. Nessa perspectiva contextual, para serem considerados de qualidade os dados devem atender satisfatoriamente a demanda para determinada aplicação (Wang; Strong, 1996, p. 6, tradução nossa).

A segunda definição para qualidade perpassa uma abordagem intrínseca, onde:

> Qualidade significa liberdade de deficiências — liberdade de erros que exigem fazer o trabalho repetidamente (retrabalho) ou que resultam em falhas de campo, insatisfação do cliente, reclamações do cliente,

> e assim por diante. Nesse sentido, o significado de qualidade é orientado para custos, e maior qualidade geralmente “custa menos” (Juran *et al.*, 1998, p. 2.1).

Como é possível notar, essas definições inicialmente são conflitantes entre si, já que, em seu primeiro sentido, qualidade significa um investimento (de tempo, recurso, estrutura, pessoal etc.) e em sua segunda acepção qualidade significa economia. Embora inicialmente conflitantes, essas definições não são excludentes, elas se complementam na medida em que representar as duas principais formas por meio das quais a qualidade pode ser encarada em uma organização.

Independente da perspectiva adotada, outro aspecto relevante para a definição do termo qualidade, é que a qualidade precisa ser avaliada, e essa avaliação ocorre em momentos diferentes. Radulovic *et al.* (2017, p. 1, tradução nossa) ressaltam a importância da avaliação da qualidade:

> A qualidade é amplamente reconhecida como uma necessidade crucial em todos os domínios (por exemplo, engenharia civil, *software*) e, para fornecer produtos e serviços de alta qualidade, a especificação e a avaliação da qualidade são de grande importância.

Ao abordar a avaliação da qualidade, Juran *et al.* (1998, p. 2.1) acrescenta que essa avaliação pode ser conduzida através de 3 processos principais, denominados posteriormente como “*Juran trilogy*”: 1) **planejamento de qualidade** – envolve estabelecer os objetivos, identificar os consumidores, determinar as necessidades dos consumidores, desenvolver as características que atendam a essas necessidades; 2) **controle de qualidade** – envolve a avaliação da performance atual, comparação da performance com os objetivos de qualidade e resolver o *gap* entre o objetivo e o estado atual; e 3) **melhoria de qualidade** – envolve provar a necessidade de melhoria, estabelecer a infraestrutura necessária; estabelecer projetos de melhoria; estabelecer equipe, promover treinamento e manter a motivação da equipe.

Esses processos são intercambiáveis, podem ser processos crônicos ou com abordagens pontuais. Nesse sentido, se agrega ainda um aspecto processual, por meio do qual se planeja, controla, avalia e melhora a qualidade de determinado produto ou serviço.

A perspectiva adotada para a definição de qualidade influência no processo de mensuração da qualidade de determinado produto/serviço, adicionando ao termo uma

característica multidimensional com diferentes acepções, que alteram a forma como se estabelece se um produto/serviço possui ou não qualidade, e ainda se qualidade é um investimento ou uma economia de recursos.

Langer *et al.* (2018, p. 165) ressaltam como a complexidade da avaliação da qualidade impacta na definição do termo:

> O termo qualidade no contexto da análise de fontes de dados é difuso e abrange aspectos que vão além de uma simples validação sintática ou verificação de correção para a ausência de contradições e erros em conjuntos de dados locais. Pesquisas anteriores já se concentraram nesse desafio e, diversas vezes, investigaram as diferentes dimensões da qualidade.

A avaliação da qualidade pode ocorrer com base em requisitos pré-estabelecidos ou ainda por meio do processo de identificação e explicitação das necessidades e objetivos de qualidade pretendidos para determinado contexto de aplicação. A ABNT (1994), na ABNT NBR ISO "**9000**: Sistemas de gestão da qualidade, ao definir o conceito de qualidade, ressalta esse aspecto intrínseco/extrínseco que pode ser relacionado às necessidades de qualidade.

> Qualidade: Totalidade de características de uma entidade que lhe confere a capacidade de **satisfazer as necessidades explícitas e implícitas**. Numa situação contratual ou numa área regulamentada, tal como na área de segurança (2.8) nuclear, as necessidades são especificadas, enquanto em outras áreas as necessidades implícitas devem ser identificadas e definidas. 2) Em muitos casos, as necessidades podem mudar no decorrer do tempo o que implica análises críticas periódicas dos requisitos para a qualidade. 3) As necessidades são traduzidas normalmente em características com critérios especificados (requisitos para a qualidade). (ABNT NBR ISO, 1994, p. 3).

Essa definição aborda ainda outro aspecto relevante: o de que as necessidades de qualidade podem ser traduzias em requisitos, ou critérios de qualidade. Esses requisitos podem ser preestabelecidos, derivando em políticas, normas, regulamentos e melhores práticas. Esses instrumentos podem ser estabelecidos a nível nacional, regional ou ainda de abrangência interna da organização. Nos casos em que não existem tais documentos, identificar quais são esses requisitos passa a ser parte do processo de avaliação.

A figura 9 apresenta uma síntese dos principais aspectos necessários para a definição do conceito de qualidade no âmbito da qualidade de dados.

Figura 9 - Sistematização da análise do termo qualidade

Fonte: autora (2025)

Com base nos aspectos apresentados na figura 9 e na fragmentação das definições fragmentadas utilizando o método da Grade, foi elaborada a seguinte definição:

> Qualidade pode ser definida como a medida de adequação de determinada entidade à requisitos de qualidade. A qualidade é objeto dos processos de avaliação, melhoria e controle. Os requisitos podem ser pré-estabelecidos ou implícitos, de caráter contextual ou intrínseco. Requisitos preestabelecidos são formalizados em um documento que irá orientar os diferentes processos, sendo exemplos as melhores práticas, políticas, normas e regulamentos. Esses documentos possuem diferentes níveis de abrangência, como nacional, regional e local. Os requisitos implícitos precisam ser identificados antes da condução dos processos mencionados. Em uma abordagem contextual, a qualidade está relacionada a capacidade da entidade de atender às necessidades dos usuários, sendo considerada um investimento, que terá um custo mais elevado. Em uma abordagem

> intrínseca, a qualidade está relacionada a ausência de erros e deficiências que possam levar a falhas ou retrabalho, resultando na economia de recursos. O termo qualidade não pode ser empregado como sinônimo de excelência ou para comparação entres diferentes entidades, para denotar essa acepção, pode ser empregado um termo composto, como: alta qualidade, baixa qualidade, boa qualidade, má qualidade, melhor qualidade ou pior qualidade.

Apresentada a definição de Qualidade, a próxima subseção apresenta as discussões e a definição do termo Dado para a qualidade de dados.

5.1.2 Dado

Fox, Levitin e Redman (1992) realizam uma ampla discussão a respeito das diferentes definições relacionadas à dados, com objetivo de defini-los no contexto da qualidade de dados. Para auxiliar na compreensão dessas definições e de suas problemáticas, o quadro 15 apresenta uma sistematização dessas definições, suas referências e as problemáticas identificadas pelos autores:

Quadro 15 - Potenciais definições e problemáticas relacionadas a definição do termo dado para qualidade de dados

DEFINIÇÃO	AUTORES	DESCRIÇÃO
Dado como sinônimo de fato	Blumenthal, 1969; Fry; Sibley, 1976	Definir dados como sinônimo de fatos implicaria na não existência de dados falsos ou imprecisos
Dado como resultado dos processos de mensuração ou observação	Davis e Rush (1979); Ralston e Reilly (1983)	A existência de dados que podem ser obtidos de outras formas, além de mensuração ou observação, tornam a definição limitada
Dado como representação ou um símbolo de objetos, conceitos e eventos do mundo real	Burch *et al.* (1983)	Implica desconsiderar o aspecto abstrato que compõe os dados, tornando possível que um mesmo dado possa ser representado de maneiras distintas
Dado como um insumo para a geração de informação	Dorn (1981)	Consiste em uma definição incompleta e interdependente, que não permite identificar as fronteiras do que seria um dado e do que seria uma informação
Dado como um item composto pela tripla (e, a, v), onde o valor é selecionado do domínio do atributo para representar o valor do atributo para a entidade	Tsichritzis and Lochovsky (1982)	Essa definição, resgatada da área de modelagem de dados não considera a distinção necessária entre aspectos conceituais (e, portanto, abstratos) e representacionais dos dados.

Fonte: adaptado de Fox, Levitin e Redman (1992)

Como é possível observar nas discussões realizadas por Fox, Levitin e Redman (1992) um dos principais desafios ao cunhar uma definição para o termo dado está em definir o que são dados. Considerando o método da grade, o desafio se encontra em estabelecer o gênero do termo. Boa parte das definições analisadas focam em definir dado a partir de seus aspectos representacionais e simbólicos.

Para Tourino (2023, p. 166) dados são “utilizados para registrar fatos e/ou representar atributos de entidades e seus contextos”. A ISO/IEC (2015) define dados como “representação reinterpretável de informações”. Nessas definições, destaca-se o conceito de dado por seus aspectos simbólico e representacional, materializado e registrado, que podem ser interpretados, reinterpretados, analisados e tratados de maneiras distintas.

Fox, Levitin e Redman (1992) não refutam o aspecto representacional e simbólico dos dados, entreato ressaltam que definir dados apenas como representações implicaria em desconsiderar o aspecto abstrato que compõe os dados. Os autores ressaltam “como os dados são abstratos, eles devem ser representados de alguma forma”.

Nessa linha, o termo dado seria compreendido como uma unidade de conteúdo, como apontado por Santos e Santana (2013). Essa unidade de conteúdo seria composta pelo aspecto conceitual e abstrato do dado, podendo ser registrada, permitido assim a sua materialização.

Redman (2013) apresenta uma definição que considera ambos os aspectos, destacando a necessária distinção entre os aspectos conceituais e representacionais do termo dado. Para o autor, dados consistem em:

> [...] dois componentes: um modelo de dados e valores de dados. Modelos de dados são abstrações do mundo real que definem do que se tratam os dados, incluindo especificações de “entidades” e “atributos” (propriedades) importantes dessas coisas e o relacionamento entre elas. [...] Por si só, um modelo de dados é muito parecido com um calendário de reuniões em branco. Há uma estrutura, mas nenhum conteúdo. Os valores de dados completam o quadro. Eles são atribuídos a atributos para entidades específicas. Assim, um único dado assume a forma: <entidade, atributo, valor> [...] (Redman, 2013, p. 22, tradução nossa).

Portanto, para que um dado possua sentido ele precisa ser contextualizado com base no domínio ao qual pertence por meio de um modelo de dados. “Embora geralmente estejamos interessados em dados sobre objetos do mundo real (físicos ou

abstratos) e relacionamentos entre objetos, trabalhamos com uma visão abstrata, ou modelo do mundo" (Fox; Levitin; Redman, 1992, p.12, tradução nossa).

Esse modelo é tradicionalmente representado, pela comunidade de banco de dados, em uma estrutura de triplas: "entidade, atributos e valor" ou "<e, a, v>".

O modelo possibilita a compreensão dos dados ao realizar uma abstração do mundo real, ou seja, refletindo o contexto no qual se inserem esses dados, descrevendo quais são as "entidades" desse domínio e quais as características possíveis para essas entidades (os seus "atributos" ou propriedades). Os valores correspondem ao valor de determinado atributo para a entidade estabelecida, podendo cada atributo possuir um conjunto de valores esperados/permitidos (Fox; Levitin; Redman, 1992; Santos; Sant'ana, 2013; Redman, 2013).

Santos e Santana (2013) ressaltam ainda que mesmo que esse modelo não seja apresentado de maneira explícita, ele precisa estar claro de maneira implícita, permitindo a interpretação dos dados.

Portanto um dado corresponde a composição entre o conteúdo conceitual e abstrato e sua representação, que ocorre por meio de um modelo, sendo cada tripla <e, a, v> correspondente a um dado. Cada novo atributo da entidade resultada em uma nova tripla. Nessa definição, um mesmo dado pode possuir mais de uma representação.

Fox, Levitin e Redman (1992, p. 12) ressaltam a necessidade de considerar os aspectos abstratos e representacionais dos dados para o contexto da qualidade de dados, apontando que a importância dessa abordagem:

> [...] é que ela leva a três conjuntos de questões de qualidade: aquelas relacionadas à qualidade do modelo ou visão, aquelas relacionadas à qualidade dos próprios valores dos dados e aqueles relacionados à qualidade da representação e registro dos dados e gravação de dados (Fox; Levitin; Redman, 1992).

Nesse sentido, para o contexto da qualidade de dados, ambos os aspectos são importantes, sendo necessário avaliar tanto a qualidade do conteúdo abstrato dos dados como os aspectos que dizem respeito a suas representações. Entretanto, o objeto principal de interpretação e avaliação de qualidade são os dados registrados, como aponta Redman (2013).

Para serem significativos os dados precisam poder ser armazenados, recuperados e processados, tanto por usuários humanos como por máquinas

(ISO/IEC, 2015; Nooghabi; Dastgerdi, 2016; Tourino, 2023). Embora os dados possam ser representados em formato analógico ou digital, para as análises de qualidade de dados destacam-se os dados representados em formato digital.

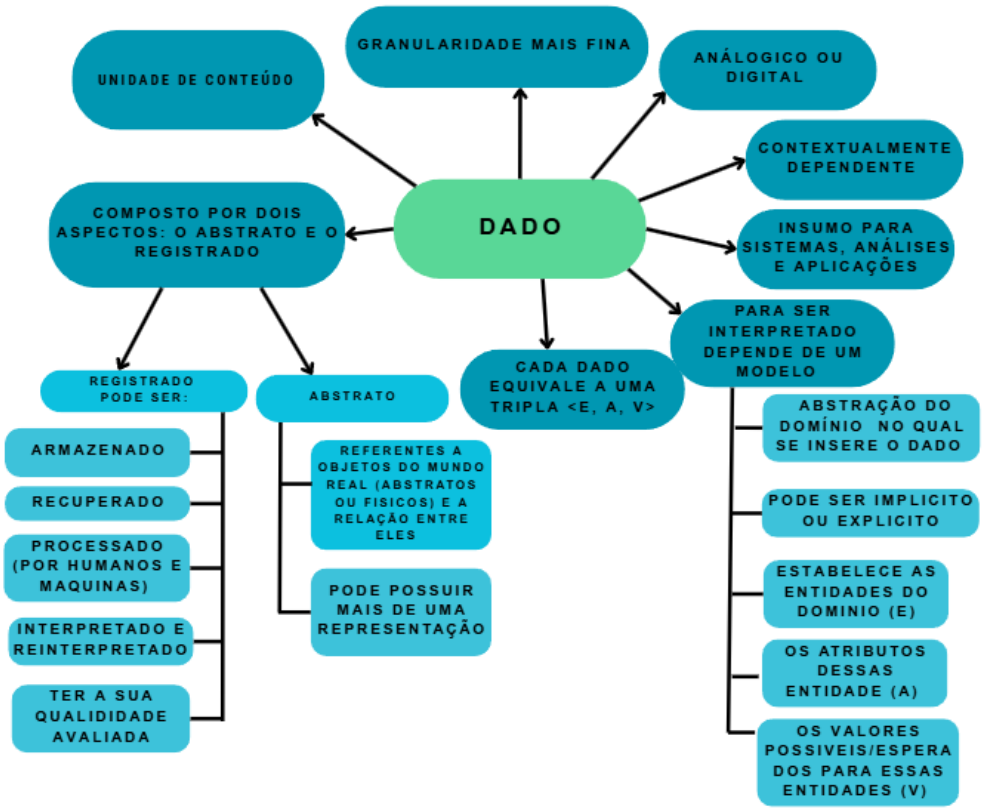
Santos e Sant'ana (2013, p. 205) definem dado com a "granularidade mais fina possível de determinado contexto de uso". Essa definição ressalta dois aspectos importantes: a de que os dados são contextualmente dependentes e a de que possuem um grande potencial de uso como insumo. Por serem a granularidade mais fina possível, os dados podem ser objeto de interpretação e reinterpretação, possuindo assim um grande potencial de processamento.

Nesse sentido, destaca-se a capacidade dos dados de atuarem como insumos, pois permitem a compreensão de fenômenos, a predição de cenários e a identificação de padrões, sendo valioso para diferentes sistemas, ferramentas e aplicações.

> As organizações sempre precisaram gerenciar seus dados, mas as mudanças na tecnologia expandiram o escopo dessa necessidade de gerenciamento, pois mudaram a compreensão das pessoas sobre o que são dados. Essas mudanças levaram a novas maneiras para criar produtos, compartilhar informações, gerar conhecimento e aprimorar o sucesso organizacional. Mas o rápido crescimento da tecnologia e, com ela, da capacidade humana de produzir, capturar e extrair significado de dados intensificou a necessidade de gerenciar dados de forma eficaz (DAMA International, 2015, p. 19).

A figura 10 apresenta uma síntese dos principais aspectos necessários para a definição do conceito de dado no âmbito da qualidade de dados.

Figura 10 - Sistematização do termo dado no âmbito da qualidade de dados.



Fonte: Autora (2025)

Com base na fragmentação das definições e na análise da sistematização apresentada, elaborou-se a seguinte definição:

> Um dado é a unidade de conteúdo de granularidade mais fina, composto por dois aspectos: o abstrato e o registrado. Em seu aspecto abstrato, o dado se refere a entidades do mundo real (conceituais ou físicas) e a relação existente entre elas, podendo um dado possuir mais de uma representação. Para permitir a sua interpretação, o dado depende de um modelo, abstração do domínio no qual se insere, podendo esse modelo estar implícito ou explícito. O modelo estabelece as entidades (e) do domínio, os atributos (a) dessa entidade e os valores (v) possíveis/esperados desse atributo para a entidade em questão. Nesse sentido, cada dado equivale a uma tripla <e,a,v>. Quando registrado, seja em meio analógico ou digital, o dado se torna passível de armazenamento, recuperação, processamento (por humanos ou máquinas), interpretação e reinterpretação e avaliação de qualidade. Por suas características, o dado é um insumo para sistemas, análises e aplicações, permitindo a compreensão de fenômenos, a predição de cenários e a identificação de padrões.

Apresentada a definição do termo Dado, a próxima subseção apresenta as discussões e a definição do termo Qualidade de Dados.

5.1.3 Qualidade de dados

Ao fragmentar e analisar as definições de qualidade de dados, observa-se uma grande divergência em relação ao estabelecimento do gênero do termo qualidade de dados. Para estabelecer a definição do glossário, partiu-se então da categorização dos gêneros identificados nas definições analisada, tendo sido estabelecidas quatro categorias principais; a qualidade de dados enquanto campo de estudo ou domínio; a qualidade de dados como um problema de pesquisa ou desafio prático a ser enfrentado; a qualidade de dados como uma medida; e a qualidade de dados enquanto um processo. O quadro 16 apresenta a sistematização realizadas.

Quadro 16 - Os gêneros da qualidade de dados organizados em categorias

Categoria (a qualidade de dados pode ser entendida enquanto)	**Gênero** (qualidade de dados pode ser definida como:)	**Autores**
campo de estudo ou domínio	Um tema	Gualdani (2022)
	Um objeto de estudo	Gualdani (2022)
	Um campo consolidado do conhecimento	Piccolo *et al.* (2021)
	Um domínio	Sadiq (2013)
	Foco da atenção dos pesquisadores	Kahlawi (2020)
	Campo de pesquisa amplamente investigado	Rashid, *et al.* (2019)
Um problema/desafio a ser superado	Uma questão	Rula (2011)
	Um problema	Knuth (2015) Arnold (1992); Juran e Godfrey, (1999);
	Um desafio	Issa (2021) Batini (2004); Rakov (1998).
	Uma questão	Rula (2011); Esteves *et. al* (2019)
Uma medida	Um fator	Almeida *et al.* (2016, p. 74)
	Um parâmetro	Piccolo *et al.* (2021)
	Uma medida	Rahoman; Ichise (2016)
	Sinônimo de adequação ao uso	Knight; Burn (2005) (Juran et al, 1974; Wang e Strong, 1996) Zaveri *et al* (2015, p. 2); Heling (2019, p. 2)
Um processo	Processo	Espíndola *et al.* (2018, p. 282); Barata (2015).

Fonte: Autora (2025)

Nesse sentido, considerando o campo “gênero” da grade de qualidade de dados e a análise geral das definições apresentadas para o termo pela literatura, entende-se que o termo qualidade de dados se caracteriza como um termo poliédrico.

Definir um termo como poliédrico implica na compreensão de que, em determinados contextos, a construção dos conceitos especializados se estabelece de forma a que um termo contenha em si mais de uma acepção. Nesse cenário, o princípio da unicidade, onde um termo deve representar a apenas uma acepção, promovido pelas escolas clássicas de terminologia, deve resignar-se em benefício da descrição fiel de uma área especializada do saber (Cabré, 1999).

A **Qualidade de Dados se estabelece como um domínio** por sua comunidade discursiva e por sua produção cientifica, técnica e tecnológica.

> É importante considerar que a qualidade de dados é um conceito de sentido delimitado tradicionalmente na literatura científica, de modo que constitui um campo consolidado do conhecimento, que possui arcabouço teórico próprio (Piccolo *et al.,* 2021, p. 6).

O domínio da qualidade de dados é embasado por “[...] várias décadas de contribuições de pesquisa de alta qualidade e inovações comerciais” (Sadiq, 2013, p. 9).

A qualidade de dados se destaca por seu caráter interdisciplinar. É considerada “[...] um tema propício como objeto de estudo de diversos campos da ciência tornando-se parte de um escopo interdisciplinar que abrange diversas áreas do conhecimento” (Gualdani, 2022, p. 485). Esse caráter interdisciplinar foi a base para a sua formação e a sua comunidade segue sendo interdisciplinar.

O caráter interdisciplinar da qualidade de dados enquanto domínio e do processo de formação de sua comunidade está intrinsicamente relacionado à um outro aspecto, o da **qualidade de dados como um problema de pesquisa ou um desafio a ser superado** por pessoas, empresas e instituições.

> A pesquisa e a prática em qualidade de dados e informações são caracterizadas por diversidades metodológicas e temáticas. A natureza interdisciplinar dos problemas de qualidade de dados, bem como um forte foco em soluções baseadas no princípio da adequação ao uso, diversificou ainda mais o corpo de conhecimento relacionado (Sadiq, 2013, p. IX).

“O problema da qualidade dos dados tem sido abordado desde que os dados foram levantados e coletados” (Knuth, 2015, p. 202). Essa abordagem da qualidade de dados como um problema está intrinsicamente relacionada com a aplicação dos dados:

> [...] uma questão importante para aplicações orientadas a dados, que deve ser profundamente investigada e compreendida. Como consequência da qualidade não controlada dos dados que fluem pelos sistemas de informação, os dados em geral podem se degradar rapidamente ao longo do tempo (Rula, 2011, p. 341).

Nesse sentido, os problemas de qualidade são diversos e os meios para superá-los precisam se adaptar tanto ao contexto de criação e aplicação desses dados, como às suas caraterísticas sintáticas e semânticas.

Quando se aborda a qualidade **de dados enquanto uma medida**, entende-se que esses dados podem possuir níveis distintos de qualidade, onde:

> a alta qualidade dos dados garante que um sistema ou produto ofereça benefícios aos seus usuários, satisfazendo suas necessidades. Em contraste, a baixa qualidade dos dados tem várias implicações negativas para os usuários (Hassan *et al.,* 2022, p.3).

A qualidade dos dados não é uma medida absoluta, não existindo uma maneira única e universal para medi-la (Paulheim; Bizer, 2014; Rahoman; Ichise, 2016).

Essa medida pode estar relacionada com a estrutura sintática dos dados, e ser caracterizada como a busca pela “menor quantidade possível de anomalias” (Almeida *et al*., 2016, p. 74). Pode ser definida em relação a conformidade dos dados com boas práticas, políticas e normas (Bentancourt; Rocha, 2012, p. 88).

As medidas de qualidade podem estar relacionadas ainda com o quão adequados os dados estão para o uso pretendido ou para o domínio de aplicação desses dados, sendo essa definição de qualidade de dados mais adotada pela literatura (Juran *et al,* 1974; Wang; Strong, 1996; Zaveri *et al*.; Rahoman; Ichise, 2016; Heling, 2019).

Enquanto uma medida, a qualidade de dados pode ser tratada como sinônimo dos conceitos “dados de qualidade”, “qualidade dos dados” ou “qualidade aplicada a dados”, onde a qualidade dos dados é um aspecto que pode ser mensurado e melhorado com base na perspectiva adotada.

Enquanto um processo, a qualidade de dados pode ser abordada pela literatura como um sinônimo dos processos de avaliação de qualidade e controle de qualidade.

> O controle da qualidade dos dados nas organizações é um dos processos compreendidos pela governança de dados, são definidas métricas, procedimentos e requisitos que auxiliam a organização a atingir a qualidade de dados necessária para cumprir suas demandas e alcançar seus objetivos (Barata, 2015, p. 276).

O controle de qualidade consiste em um processo constante de gerenciamento, que "[...] se dá pela "definição de papéis, responsabilidades, políticas e procedimentos relacionados à aquisição, manutenção, representação e disseminação de dados e informações" (Botega *et al*., 2017).

O processo de avaliação de qualidade pode ser considerado como uma etapa do controle de qualidade ou como um processo isolado, realizado com o propósito de seleção de fontes ou de resolução de problemas de qualidade específicos. "Diversas metodologias foram desenvolvidas para aprimorar bem como avaliar a qualidade dos dados" (Kahlawi, 2020, p. 60).

Considerando que os problemas de qualidade possuem um caráter fortemente contextual, e que a qualidade não pode ser medida de maneira única e absoluta, o processo de avaliação de qualidade também precisa ser intrinsecamente plural.

O processo de avaliação de qualidade dos dados se caracteriza, portanto, como um processo multidimensional (Esteves et. al, 2019). Os modelos que permitem a avaliação de qualidade acompanham essa característica.

> Observa-se na literatura, tanto estudos que propõem que a qualidade dos dados seja considerada de forma genérica, onde a partir de um determinado modelo a qualidade pode ser gerida independente do domínio de aplicação, como também pesquisas que procuraram definir e operacionalizar as dimensões de qualidade de dados específicas para determinados contextos (Fagundes; Macedo; Freund, 2018, p. 196).

O *framework* da qualidade de dados proposto por Wang e Strong (1996) segue sendo a base para orientar o processo de avaliação de qualidade. Nessa estrutura, a avaliação é realizada por meio da escolha de dimensões, critérios e métricas que permitem medir a qualidade dos dados. As dimensões permitem agrupar os diferentes

aspectos que serão avaliados nos conjuntos de dados, sendo compostas por um conjunto de critérios, que descrevem os atributos específicos que serão analisados. As métricas são os indicadores que permitem mensurar a qualidade dos dados em relação a um critério.

> Desse modo, dimensões funcionam como aspectos pelos quais a qualidade de dados pode ser avaliada de maneira adequada, dentro de um domínio. Isso pressupõe que, para determinado domínio, sejam definidas essas dimensões, o que pode ser concretizado no âmbito de uma metodologia de avaliação construída com foco nesse contexto específico (Piccolo *et al*., 2021, p. 6).

Nesse sentido, entende-se que o processo de avaliação de qualidade é guiado pelo estabelecimento das dimensões a serem avaliadas e pela identificação e criação das ferramentas necessárias para a sua avaliação.

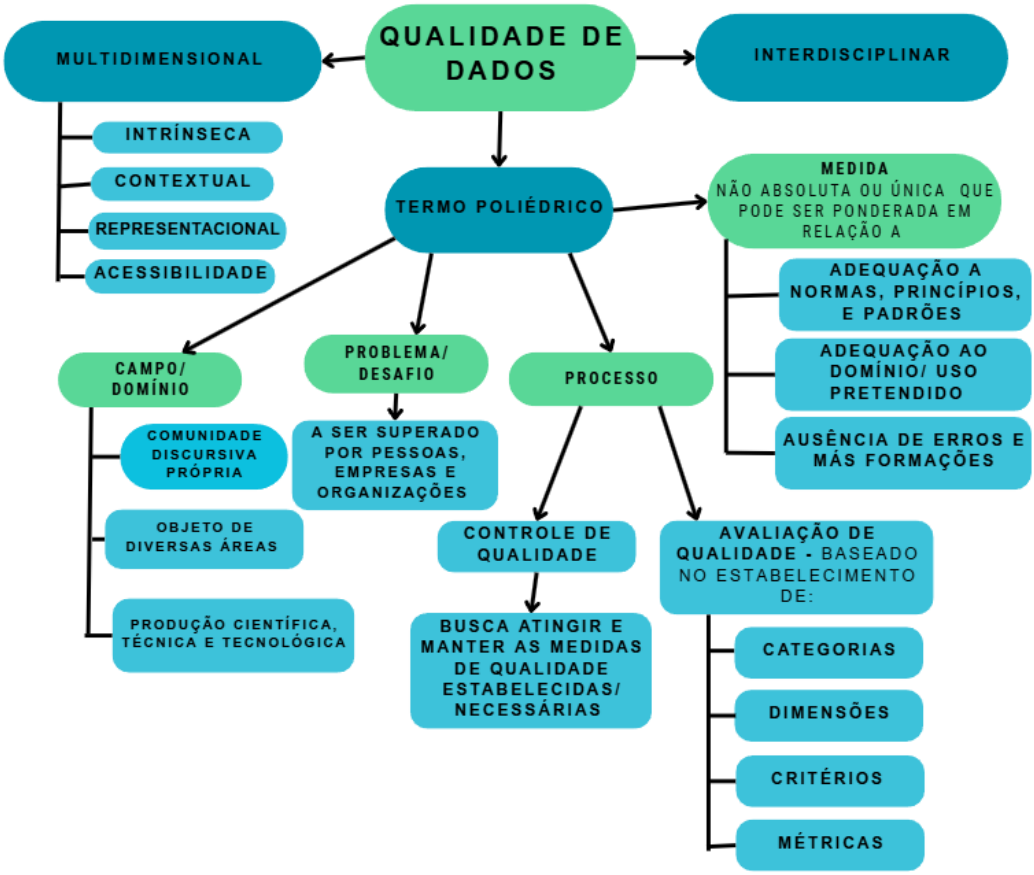
O aspecto multidimensional da qualidade ultrapassa o seu entendimento enquanto um processo, diferentes perspectivas, ou dimensões, afetam a forma como se realiza o processo de avaliação de qualidade e como se identifica se um conjunto de dados é ou não de boa qualidade.

Em relação a essas dimensões, elas geralmente seguem o processo de categorização apresentado por Wang e Strong (1996), que é a referência mais frequente nas definições, tanto para apresentar dimensões, como para descrever o processo de avaliação de qualidade.

Como discutido, as principais dimensões, ou ainda perspectivas, do conceito de qualidade de dados são: contextual, intrínseca, representacional e acessibilidade.

Com base na fragmentação das definições de qualidade de dados, elaborou-se uma sistematização desse termo, apresentada na figura 11:

Figura 11 - Sistematização da análise do termo qualidade de dados



Fonte: Autora (2025)

Com base na fragmentação das definições e na sua sistematização apresentada na figura 11, foi construída uma definição para compor o glossário:

> O termo qualidade de dados pode ser considerado poliédrico, de caráter interdisciplinar e multidimensional, contando com as acepções: **1. Enquanto um campo, ou domínio**, a qualidade de dados possui uma comunidade de discurso e uma produção cientifica, técnica e tecnológica bem estabelecida, de escopo interdisciplinar, de interesse de diversas áreas do conhecimento no qual são discutidas, investigadas e propostas formas de mensurar e melhorar os níveis de qualidade dos conjuntos de dados, assim como formular artefatos e metodologias que permitam auxiliar na condução de processos de avaliação e melhoria de qualidade de dados. **2. Enquanto um problema/desafio** a qualidade de dados é entendido como uma barreira a ser superada por pessoas, empresas e organizações para o uso efetivo e eficiente de dados em diversos contextos. **3. Enquanto uma medida ou uma ponderação** pode ser abordada como sinônimo dos conceitos “dados de qualidade”, “qualidade dos dados” ou “qualidade aplicada a dados”, onde a qualidade dos dados é um aspecto que pode ser mensurado e melhorado. A medida de qualidade dos dados não pode ser mensurada de maneira única ou absoluta, e pode estar relacionada à: adequação a normas, princípios, padrões e melhores práticas; a ausência de erros e anomalias; e a adequação ao domínio ou uso pretendido dos dados **4. Enquanto processo**, a

> qualidade de dados é abordada pela literatura como um sinônimo dos processos de avaliação e controle de qualidade de dados. O processo de controle de qualidade busca atingir e manter as medidas de qualidade necessárias ou estabelecidas. O processo de avaliação de qualidade é realizado por meio do estabelecimento de critérios, dimensões e métricas que permitem mensurar os níveis de qualidade dos conjuntos de dados. O caráter interdisciplinar e multidimensional do termo afeta todas as acepções mencionadas, uma vez que tanto o domínio da qualidade de dados, como os problemas, a sua medida, manutenção e avaliação são objeto de interesse de diversas áreas do conhecimento. Podem ser considerar as principais dimensões de qualidade de dados a contextual, intrínseca, representacional e a acessibilidade.

Apresentada a definição do termo Dado, a próxima subseção apresenta as discussões e a definição dos termos relacionados ao *Linked Data*.

5.2 Termos relacionados ao *Linked Data*

A presente subseção reúne os termos relacionados com o *Linked Data*, necessários para compreender como a estrutura e o contexto desses dados impactam na avaliação de qualidade de dados.

Os termos apresentados são: *Linked Data*; *Universal Resource Identifiers* (URIs); *Literal*; *Resource Description Framework* (RDF); Vocabulário e SPARQL. As próximas subseções representam cada um dos termos, sendo compostas pelas discussões, sistematização e definição dos termos.

5.2.1 *Linked Data*

A fragmentação das definições permitiu observar que existem apenas pequenas variações em relação ao gênero do termo *Linked Data*, que é definido como conjunto de práticas, boas práticas ou melhores práticas (Bizer; Heath; Berners-Lee, 2009; Bizer; Heath, 2011; W3C, 2014; Isotani; Bittencourt, 2015; W3C, 2023).

Arakaki (2016) buscou definições e abordagens a respeito de *Linked Data* e sua relação com a Ciência da Informação, concluindo que a maioria dos autores definem *Linked Data* “[...] como melhores práticas para estruturação de dados na *Web*.” (Arakaki, 2016, p. 57). O estudo constatou que os trabalhos identificados se baseavam, em sua maioria, nas definições de Berners-Lee (2006), Bizer, Heath e Berners-Lee (2009) e Heath e Bizer (2011).

O objetivo dessas práticas é promover “**publicação e conexão** de dados estruturados na *Web*, usando padrões internacionais recomendados pelo W3C.”

(Isotani; Bittencourt, 2015, p. 34, grifo nosso). Buscam ainda promover "**acesso** tanto por humanos quanto por máquinas.

> Em resumo, *Linked Data* consistem simplesmente em usar a *Web* para criar **conexões tipificadas** entre dados de diferentes fontes. Estas [fontes] podem ser tão diversas quanto bancos de dados mantidos por duas organizações em diferentes localizações geográficas, ou simplesmente sistemas heterogêneos dentro de uma organização que, historicamente, não interoperaram facilmente no nível dos dados. Tecnicamente, *Linked Data* refere-se a dados publicados na *Web* de forma que sejam legíveis por máquinas, **seu significado seja explicitamente definido**, estejam conectados a outros conjuntos de dados externos e possam[...] (Bizer; Heath; Berners-lee, 2009, p. 3, tradução nossa, grifo nosso).

Essa definição ressalta a importância da tipificação das relações existentes entre os dados, que permite que o significado das relações seja explicitamente definido de maneira legível para máquinas. Ressalta ainda a importância das conexões entre diferentes fontes que podem ser internas ou externas.

W3C (2023, não paginado, tradução nossa, grifo nosso) definem *Linked Data* como "[...] um conjunto de práticas **recomendadas** para publicação de dados estruturados na *We*b. Ressalta-se nessa definição o caráter não prescritivo do *Linked Data*, baseado sempre em boas práticas e recomendações.

O "*Linked Data*, foi criado por Tim Berners-Lee pela necessidade de padronizar a conexão entre dados na *Web* (Isotani; Bittencourt, 2015, p. 13).

Berners-Lee (2006) não apresenta uma definição direta *para Linked Data*, mas sim um conjunto composto por 4 princípios:

> 1. Use URIs como nomes para coisas;
> 2. Use URIs HTTP para que as pessoas possam pesquisar esses nomes;
> 3. Quando alguém procura um URI, forneça informações úteis, usando os padrões (RDF*, SPARQL)
> 4. Inclua *links* para outros URIs para que eles possam descobrir mais coisas.

Esses princípios posteriormente seriam completados e ampliados.

> Compreende-se que o uso dos padrões criados pelos Grupos de Trabalho do W3C e o trabalho da comunidade de desenvolvedores, de gestores governamentais e da sociedade interessada no desenvolvimento *Web* são essenciais para que se alcance

> efetivamente dados abertos e conectados (Isotani; Bittencourt, 2015, p. 13).

Nesse sentido, embora a base para publicação de dados *Linked Data* sejam as quatro práticas apresentadas por Berners-lee (2006), a adoção do *Linked Data* perpassa outras recomendações e boas práticas, que derivam das produções dos grupos de trabalho do W3C, mas também das experiências dos usuários, e das boas práticas, políticas e normas derivadas do domínio de criação e uso dos dados.

Em síntese, “Estas práticas são fundamentadas em tecnologias *Web*, como HTTP (*Hypertext Transfer Protocol*) e URI (*Uniform Resource Identifier*), com o objetivo de permitir a leitura dos dados conectados, de forma automática, por agentes de *software*” (Isotani; Bittencourt, 2015, p. 31).

A relação entre o RDF e o *Linked Data* é reforçada pelo W3C (2014), que destaca que RDF e *Linked Data* não são sinônimos, entretanto, não existem dados *Linked Data* sem adoção do modelo.

Para o Intercâmbio dos dados, o *Linked Data* se utiliza dos formatos de serialização do RDF, como por exemplo: RDF/XML, RDFa, Turtle. Em relação a busca e recuperação o *Linked Data* “Permite consultas SPARQL distribuídas dos conjuntos de dados e uma abordagem de navegação ou descoberta para encontrar informações (em comparação com uma estratégia de busca)” (W3C, 2014, não paginado, tradução nossa).

O *Linked Data* consiste, portanto na estruturação dos “recursos em redes conectadas e semanticamente enriquecidas, favorecendo a **interoperabilidade** entre sistemas e domínios” (Arakaki *et al*., 2025, p. 2, grifo nosso).

Destacam-se dois aspectos importantes do *Linked Data*: a busca pela interoperabilidade de dados estruturados entre fontes, domínios e sistemas distintos, que antes poderiam ser dificultadas pela adoção de modelos especificamente criados para esses domínios e sistemas, facilitada pela adoção do RDF, e o enriquecimento semântico promovido “pela formalização de propriedades e de axiomas que tornam o nível de semântica formal mais elevado” (Torino *et al*., 2020, p. 8). Para isso, tornam-se necessários os vocabulários e as ontologias que possibilitam a definição formal das propriedades e axiomas.

Outro aspecto é a sua relação com os princípios de dados abertos, “[...] É importante frisar que “Dados Conectados” não necessariamente precisam ser abertos”

(Isotani; Bittencourt, 2015, p. 34). Quando são adotados em conjunto, os princípios de Dados Abertos e do *Linked Data* “geralmente são chamados de Linked Open Data” (W3C, 2013).

Embora o conceito de *Linked Open Data* inicialmente seja utilizado para se referir a um projeto de materialização dos princípios do *Linked Data* (Santarem Segundo, 2015), o termo é comumente utilizado pela comunidade discursiva para se referir a junção dos princípios de Dados Abertos e *Linked Data*.

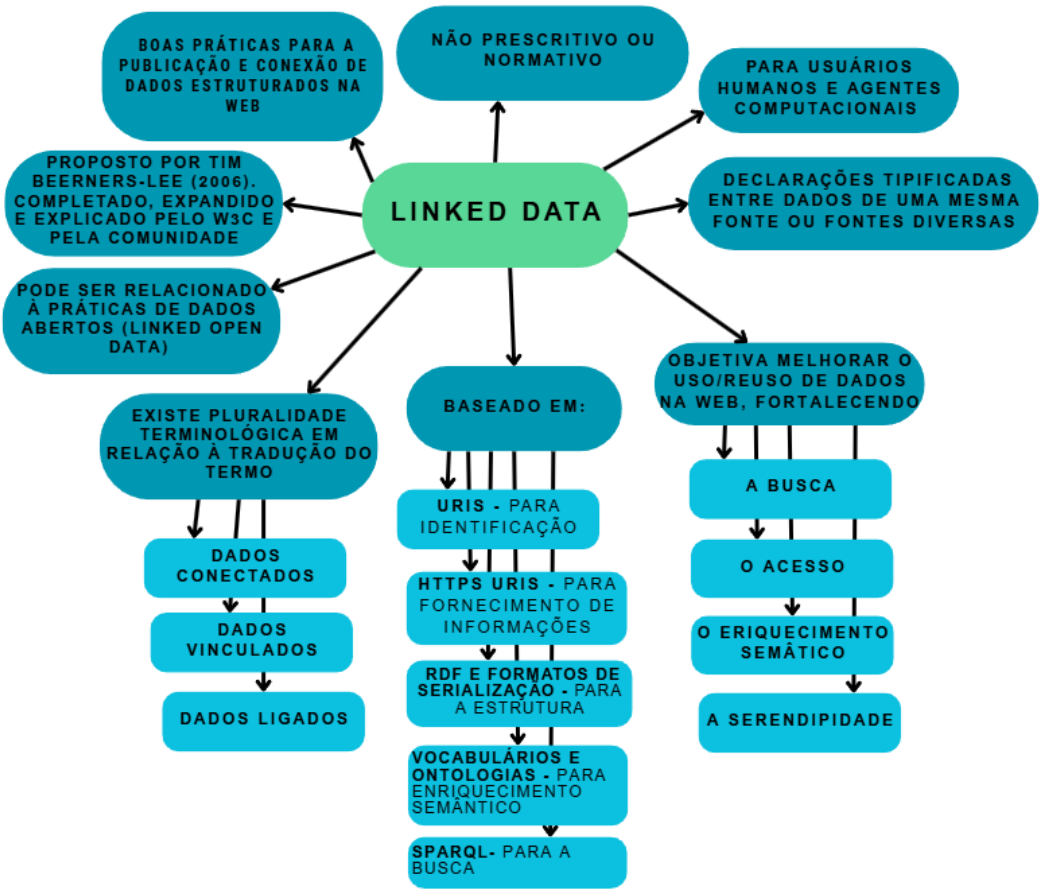
> O *Linked Data* está focado na interoperabilidade Técnica e o *Linked Open Data* focado na interoperabilidade Legal. Ou seja, o primeiro está mais relacionado às melhores práticas para estruturação dos dados a partir das ferramentas e a garantia de troca de informações e o segundo com as questões de abertura dos dados e do uso de licenças de código aberto. Ambos os termos referem à ligação de dados e a diferença primordial concentra-se na questão de estarem abertos ou não (Arakaki, 2016, p. 119).

Existe uma pluralidade semântica em relação a tradução do termo para o português, como apontam Isotani e Bittencourt (2015, p. 31):

> Apesar de estarmos usando o termo Dados Conectados, outros pesquisadores fazem menção ao mesmo conceito usando os termos dados interligados ou dados ligados. Vários dicionários traduzem o termo *Linked* como ligado, conectado, associado, entre outros. Nós acreditamos que o termo ligado não é o mais adequado para transmitir o significado atribuído ao termo “*linked*” dentro do nosso contexto. Ao observar o termo “Ligar”, há muitos cenários que não estão adequadamente aplicados a *Linked Data*. No entanto isso não acontece ao considerarmos a palavra “conectar”.

Considerando essa pluralidade, e o fato de o termo ser majoritariamente adotado em inglês, inclusive pela comunidade da Ciência da Informação, optou-se por manter o termo em inglês. Com base nas discussões e na fragmentação das definições elaborou-se a figura 12.

Figura 12 - Sistematização do termo Linked Data



Fonte: Autora (2025)

Com base nas definições analisadas e na sistematização apresentada na figura 12, elaborou-se a seguinte definição para compor o glossário:

> O *Linked Data* é um conjunto de boas práticas, de caráter não prescritivo ou normativo, para a publicação e conexão de dados estruturados na W*eb*. Tem como objetivo fortalecer e facilitar a busca, recuperação, acesso, reuso, intercambio e interoperabilidade de dados tanto para usuários humanos como para agentes computacionais. Busca prover a serendipidade (descoberta acidental) e o enriquecimento semântico dos dados. As boas práticas se utilizam de padrões da *Web*, como o protocolo HTTP e o uso de identificadores únicos, como URIs e IRIs. O modelo RDF é a base para a estrutura dos dados *Linked Data*, tendo seu uso combinado ao de formatos de serialização. O RDF prevê o uso de vocabulários e ontologias, que permitem o enriquecimento semântico dos dados. Essas conexões ocorrem por meio de *Links* que conectam dados de uma mesma fonte ou de fontes diversas. O *Linked Data* foi proposto em 2006 por Tim Berners-Lee, sendo posteriormente expandido, explicado e completado por uma série de documentos e boas práticas do W3C e pela própria comunidade de usuários. Sua adoção também é afetada por princípios e normas do domínio ao qual se relacionam os dados. As práticas do *Linked Data* podem ser adotadas em conjunto com práticas de dados abertos, sendo, nesse caos, nomeado pela

> comunidade como *Linked Open Data*. Existe uma pluralidade em relação a tradução do termo para o português, que pode ser traduzido como dados conectados, dados vinculados, dados ligados ou dados interligados.

Apresentada a definição de *Linked Data*, a próxima subseção apresenta as discussões e a definição do termo *Universal Resource Identifiers (Uris)*.

5.2.2 UNIVERSAL RESOURCE IDENTIFIERS (URIS)

Um *Universal Resource Identifier* (URI) pode ser definido como um "Um identificador global padronizado pela ação conjunta do *World Wide Web Consortium* e da *Internet Engineering Task Force*" (W3C, 2013, não paginado, tradução nossa). Consistem em "sequências curtas que identificam recursos na *Web*" (W3C, 2001, não paginado, tradução nossa).

Os URIs são adotados na *Web* para identificar os recursos, ou entidades, que podem ser "qualquer coisa, incluindo uma construção física ou conceitos mais abstratos, como cores" (W3C, 2013, não paginado, tradução nossa).

Os URIs possuem uma variedade de formas, podendo ser classificados como:

> **URL (*Uniform Resource Locator*),** que basicamente define um localizador/endereço para um determinado recurso a partir de um protocolo existente; e ii) **URN (*Unified Resource Name*),** que representa um nome para um determinado recurso, garantindo unicidade e persistência de forma global mesmo quando o recurso não está disponível (Isotani; Bittencourt, 2015, p. 57).

Tem-se ainda o conceito de International Resource Identifier (IRI):

> IRIs são uma generalização de URIs que permite uma gama mais ampla de caracteres Unicode. Todo URI e URL absolutos são IRIs, mas nem todo IRI é um URI. Quando IRIs são usados em operações definidas apenas para URIs, eles devem primeiro ser convertidos de acordo com o mapeamento definido (W3C, 2014, não paginado, tradução nossa, grifo nosso).

Os URIs podem ser desreferenciados, ou resolvidos, isso implica na possibilidade de que as pessoas busquem por esses URIs na *Web* e tenham como retorno uma representação do recurso (no caso de URIs que representam recursos do mundo real) ou acesso ao próprio recurso, bem como informações úteis a respeito desse recurso. Para que possam ser desreferenciados, os URIs podem ser

associados ao protocolo de compartilhamento de informações da *Web*, sendo chamados de HTTPs URIs.

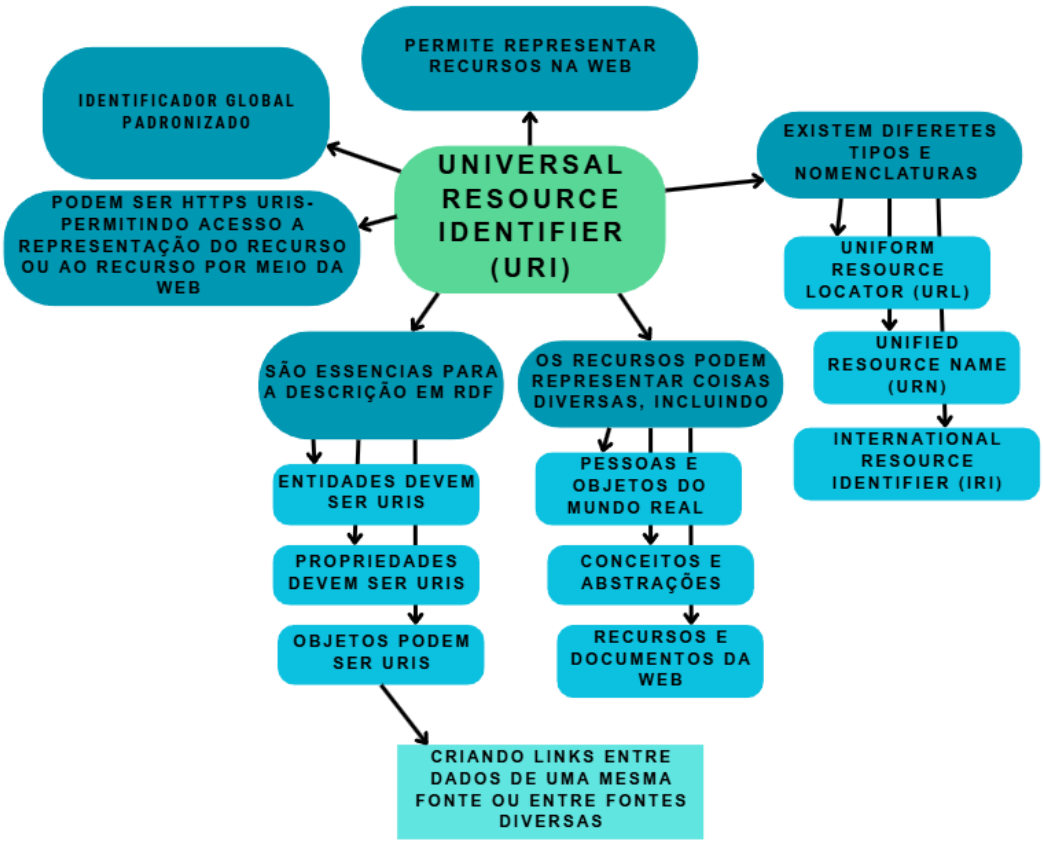
Os URIs desempenham um papel fundamental na publicação de dados *Linked Data*, pois são essenciais para a estrutura do RDF.

Como para o modelo RDF “[... todo recurso preferencialmente deve ter um URI/IRI, então o IRI pode representar um <sujeito>, <predicado> ou <objeto> (Isotani; Bittencourt, 2015 p. 61).

Enquanto o objeto pode ou não ser representado por um URI, os sujeitos (ou recursos) e os predicados (ou propriedades) precisam necessariamente ser descritos como URIs/IRIs.

Com base nas discussões realizadas e na fragmentação das definições selecionadas, elaborou-se a sistematização apresentada na figura 13:

Figura 13 - Sistematização do termo URI



Fonte: Autora (2025)

Com base nas discussões, na fragmentação das definições e na sistematização apresentada na figura 13, elaborou-se a seguinte definição para compor o glossário:

Os *Universal Resource Identifiers* (URIs) são identificadores globais padronizados que representam, por meio do uso de uma estrutura e caracteres preestabelecidos, os recursos na *Web*. Recursos podem ser coisas diversas, como pessoas e objetos do mundo real, conceitos e abstrações e ainda recursos e documentos da *Web*. Existem diferentes tipos e nomenclaturas, sendo exemplos os *Uniform Resource Locator* (URL), que definem a localização do recurso a partir de um protocolo, o URN (*Unified Resource Name*) que definem o nome dos recursos, e ainda os *International Resource Identifier* (IRIs) que estendem as possibilidades de caracteres permitidos para a representação dos recursos. Os URIs podem ser URIs HTTPs, que permitem que esses sejam desreferenciados/ resolvidos, utilizando o procolo de compartilhamento da *Web* para permitir acesso a informações sobre as entidades, no caso de representações de objetos e pessoas do mundo real, ou ainda aos próprios recursos no caso de recursos *Web*. Os URIs são essenciais para a estruturação dos dados em RDF pois as entidades e propriedades devem, necessariamente, ser URIs. Os objetos das declarações em RDF podem ser URIs ou Literais. Quando são utilizados URIs como objetos, criam-se *Links* entre dados de uma mesma fonte ou de fontes diversas.

Apresentada a definição do termo *Linked Data*, a próxima subseção apresenta as discussões e a definição do termo Literal.

5.2.3 LITERAL

No contexto do *Linked Data*, um literal consiste em um conjunto de valores em linguagem natural, “Literais são usados para valores como *strings*, números e datas.” (W3C, 2014, não paginado, tradução nossa).

Isotani e Bittencourt (2015, p. 63) definem os literais como “[...] todos os valores identificados em um grafo RDF que não tem um IRI [ou URI] associado.”

Os literais também representam entidades do mundo real:

> Qualquer IRI ou literal denota algo no mundo (o "universo do discurso"). Essas coisas são chamadas de recursos. Qualquer coisa pode ser um recurso, incluindo coisas físicas, documentos, conceitos abstratos, números e *strings*; o termo é sinônimo de "entidade" como é usado na Semântica RDF 1.2 [RDF12-SEMANTICS]. O recurso denotado por um IRI é chamado de referente, e o recurso denotado por um literal é chamado de valor literal. (W3C, 2025, não paginado, tradução nossa).

Os Literais são importantes pois “permitem que informações baseadas em texto sejam integradas ao modelo de dados RDF.” (Beek *et al*., 2017, p. 3).

Os Literais podem ser acompanhados de uma *tag* opcional que permite a identificação do seu idioma (Heath; Bizer, 2011). Também podem ser associados a um tipo de dados, descrito por um URI.

> Tipos de dados são usados com literais RDF para representar valores como *strings*, números e datas. A abstração de tipo de dado usada em RDF é compatível com o *XML Schema*. Qualquer definição de tipo de dado que esteja em conformidade com essa abstração PODE ser usada em RDF, mesmo que não esteja definida em termos do XML Schema (W3C, 2025, não paginado, tradução nossa).

Ressalta-se que a utilização de literais no modelo RDF é permitida apenas para a representação dos valores das declarações, não podendo ocupar os espaços de recurso e das propriedades.

Com base nas discussões e na fragmentação das definições selecionadas foi elaborada a sistematização apresentada na figura 14.

Figura 14 - Sistematização do termo Literal

Fonte: Autora (2025)

Com base nas discussões, na fragmentação das definições e na sistematização apresentada na figura 14, elaborou-se a seguinte definição:

No contexto do *Linked Data* um Literal é um conjunto de valores em linguagem natural e que não é representado por um URI em uma declaração RDF. Os literais são números, datas, textos, *strings*, dentre outros. Referem-se a entidades do mundo real, como pessoas, objetos, conceitos e abstrações. Literais podem ser associados a uma *tag* que indique em que idioma estão. Devem ser acompanhados de uma indicação de tipo de dados em algum dos esquemas aceitos pelo RDF, representados por um

URI, essa identificação facilita a avaliação da qualidade dos literais. Literais podem ser utilizados apenas na posição de valor em uma declaração RDF, não podendo ser utilizado nas posições de recurso e propriedade.

5.2.4 Resource Description Framework (RDF)

O RDF é definido pelo W3C como um *framework*, uma estrutura de dados. (W3C, 2014; W3C; 2025). Pode ser definido ainda como um modelo padrão (W3C; 2014) ou como o “equivale a uma linguagem de representação de informação na Web” (Isotani; Bittencourt, 2015, p. 57).

O RDF “baseia-se na ideia **de identificar objetos** usando identificadores da *Web* ou URIs HTTP **e descrever recursos** em termos de propriedades simples e valores de propriedades.” (W3C, 2013, não paginado, tradução nossa, grifo nosso).

O modelo tem como objetivo “**expressar descrições** de recursos” W3C (2014, não paginado, tradução nossa, grifo nosso). Isotani e Bittencourt comentam ainda que o RDF visa permitir a **descrição formal** dos recursos bem como a **relação existente entre recursos**.

A busca pela descrição das relações é reforçada pelo W3C: “O RDF estende a **estrutura de ligação da Web** para usar URIs para **nomear o relacionamento entre as coisas** [...]” (2014, não paginado, tradução nossa, grifo nosso). “Usando esse modelo simples, ele permite que dados estruturados e semiestruturados sejam **combinados**, **expostos** e **compartilhados** entre diferentes aplicações.” (W3C, 2014, não paginado, tradução nossa, grifo nosso).

A estrutura do RDF é baseada em declarações a respeito dos recursos:

> Qualquer IRI ou Literal denota algo no mundo (o "universo do discurso"). Essas coisas são chamadas de recursos. Qualquer coisa pode ser um recurso, incluindo coisas físicas, documentos, conceitos abstratos, números e *strings*; o termo é sinônimo de "entidade", como é usado na especificação (W3C, 2014, não paginado, tradução nossa).

Nesse sentido, o RDF tem como proposito descrever as entidades por meio de declarações que as identifique ou que expressem de maneira formal e nomeada o tipo de relação existente entre a entidade sujeito da declaração e uma outra entidade. A estrutura do RDF se baseia no estabelecimento de um:

> [...] conjunto de triplas, cada uma composta por um sujeito, um predicado e um objeto. Um conjunto dessas triplas é chamado de grafo

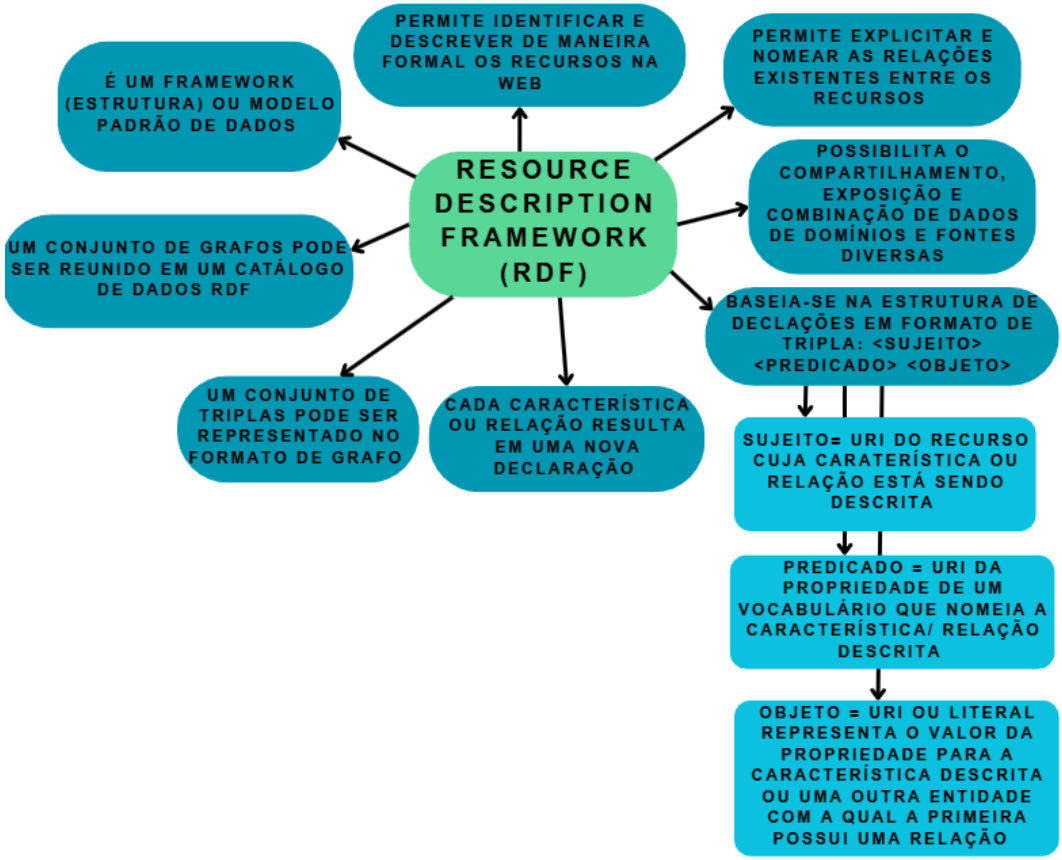
> RDF. Um grafo RDF pode ser visualizado como um diagrama de nó e arco direcionado, no qual cada tripla é representado como uma ligação nó-arco-nó (W3C, 2014, não paginado, tradução nossa).

Cada tripla pode ser representada no formato de <sujeito> <predicado> <objeto>, ou ainda <recurso> <propriedade> <valor>. Onde o recurso é a entidade descrita, a propriedade é a característica desse recurso a ser descrita e o valor consiste no valor dessa propriedade para o recurso descrito. Cada nova caraterística, ou valor, da entidade deve ser expressa por uma nova tripla (Isotani; Bittencourt, 2015).

O RDF pode ser ligado ainda ao "[...] conceito de Múltiplos Grafos. Ao se criar um documento RDF, pode-se adicionar outros grafos que se conectam ao grafo original, proporcionando assim catálogos de dados" (Isotani; Bittencourt, 2015, p. 66).

Com base nas discussões e na fragmentação das definições selecionadas foi elaborada a sistematização apresentada na figura 15.

Figura 15 - Sistematização do termo RDF



Fonte: Autora (2025)

Com base nas discussões, na fragmentação das definições e na sistematização apresentada na figura 15, elaborou-se a seguinte definição:

O RDF é um *framework* (estrutura), ou modelo padrão de dados, que permite identificar e descrever as características dos recursos na *Web*, bem como explicitar e nomear as relações existentes entre recursos. Essa estrutura facilita o compartilhamento, exposição e combinação de dados de domínios distintos. Baseia-se na estrutura de declarações em formato de tripla <sujeito> <predicado> <objeto>. O sujeito é representado por um URI do recurso cuja caraterística ou relação está sendo descrita. O predicado deve ser o URI da propriedade de um vocabulário que nomeei a característica ou a relação descrita. O objeto pode ser um URI ou um Literal que represente o valor da propriedade para a característica descrita ou um outro recurso com o qual o primeiro possui uma relação. Cada característica ou relação resulta em uma nova declaração. Um conjunto de triplas pode ser representado no formato de grafo. Um conjunto de grafos pode ser reunido em um catálogo de dados RDF.

Apresentada a definição do termo RDF, a próxima subseção apresenta as discussões e a definição do termo Vocabulário.

5.2.5 Vocabulário

No contexto do *Linked Data*, vocabulário pode ser definido como uma coleção de termos criados para um propósito específico. (W3C, 2017; W3C, 2013). Vocabulários podem ser entendidos ainda como “coleções de classes e propriedades” (Bizer; Heath; Berners-Lee, 2009, p. 5). Um vocabulário no *Linked Data* consiste em:

> [...] um conjunto de classes e propriedades (chamadas simplesmente de termos do vocabulário), úteis para descrever tipos específicos de coisas, ou coisas em um determinado domínio ou indústria, ou coisas em geral, mas para um uso específico ou uma reunião de definições (LOV, [s.d.], não paginado).

Em relação a sua aplicação, os vocabulários:

> [...] definem os conceitos e relacionamentos (também chamados de "termos" ou "atributos") usados para **descrever e representar uma área de interesse**. Eles são usados para **classificar os termos** que podem ser usados em uma aplicação específica, **caracterizar possíveis relacionamentos e definir possíveis restrições ao uso**

> **desses termos** (W3C, 2017, não paginado, tradução nossa, grifo nosso)

Nesse sentido, os vocabulários têm o propósito de apresentar e organizar termos, que podem ser classes e propriedades, que buscam descrever os recursos e os potenciais relacionamentos existentes entre recursos de uma área de interesse, geralmente sendo criados para atender a um domínio ou contexto de aplicação específico.

São essenciais para a aplicação do modelo RDF, pois:

> O RDF fornece um modelo de dados genérico e abstrato para descrever recursos usando triplas de sujeito, predicado e objeto. No entanto, não fornece termos específicos de domínio para descrever classes de coisas no mundo e como elas se relacionam entre si. Essa função é atendida por taxonomias, vocabulários e ontologias. (Heath; Bizer, 2011, não paginado, tradução nossa).

Os vocabulários possuem uma relação direta com a semântica dos dados, como aponta LOV ([s.d.], não paginado):

> As definições de termos fornecidas pelos vocabulários trazem uma semântica clara para descrições e *links*, graças à linguagem formal que utilizam (algum dialeto de RDF, como RDFS ou OWL).

Em relação a sua estrutura, no contexto do *Linked Data* os vocabulários "[...] são expressos em RDF, usando termos de RDFS e OWL, que fornecem vários graus de expressividade na modelagem de domínios de interesse" (Bizer; Heath; Berners-Lee, 2009, p. 5, tradução nossa).

Embora qualquer pessoa possa criar um vocabulário para ser utilizado em dados *Linked Data*, a recomendação do W3C é que sejam priorizados vocabulários já existentes e bem estabelecidos no domínio (W3C, 2017).

O principal desafio de uma definição para o termo vocabulário no domínio do *Linked Data* está em sua relação com outros termos como: ontologias, vocabulários controlados e Sistemas de Organização do Conhecimento.

Esses termos são muitas vezes tratados pela comunidade do *Linked Data* como sinônimos, e utilizados de maneira intercambiável para se referir a um mesmo instrumento, como no caso da OWL, onde a documentação oficial utiliza simultaneamente os termos vocabulário e ontologia.

Ao apresentar boas práticas para a publicação de vocabulários o W3C (2008, não paginado, tradução nossa), aponta que “vocabulário e ontologia são usados indistintamente no contexto desta especificação”. O glossário de termos do W3C (2011, não paginado, tradução nossa), ao se referir ao termo vocabulário, menciona que: “O uso deste termo se sobrepõe ao de Ontologia”. O W3C (2017, não paginado, tradução nossa) ao abordar a relação entre os termos ontologia e vocabulário afirmam que “não há uma divisão estrita entre os artefatos referidos por esses nomes” e que em relação ao termo vocabulário “o uso deste termo se sobrepõe ao de Ontologia”.

Embora possam ser utilizados de maneira intercambiada pela literatura, o W3C (2017) trata os termos como quase sinônimos:

> Vários quase sinônimos para "vocabulário" foram cunhados, por exemplo, ontologia, vocabulário controlado, tesauro, taxonomia, lista de códigos e rede semântica. Não há uma divisão estrita entre os artefatos referidos por esses nomes. 2) "Ontologia", no entanto, tende a denotar os vocabulários de classes e propriedades que estruturam as descrições de recursos em conjuntos de dados (vinculados). Ontologias são os principais blocos de construção para técnicas de inferência na Web Semântica. O primeiro meio oferecido pelo W3C para a criação de ontologias é a linguagem RDF Schema. É possível definir ontologias mais expressivas com axiomas adicionais usando linguagens como as da *The Web Ontology Language* (W3C, 2017, não paginado, tradução nossa).

Nesse sentido, pode-se compreender que o termo vocabulário atua como um termo guarda-chuva para se referir a diferentes tipos de coleções de termos. A ontologia pode ser entendida nesse domínio como um tipo de vocabulário, caracterizado por uma maior complexidade estrutural, atuando como um:

> Um modelo formal que permite a representação do conhecimento para um domínio específico. Uma ontologia descreve os tipos de coisas que existem (classes), os relacionamentos entre elas (propriedades) e as maneiras lógicas pelas quais essas classes e propriedades podem ser usadas em conjunto (axiomas). (W3C, 2017, não paginado, tradução nossa).

A ontologia pode se diferenciar pela presença de formalismos complexos, como axiomas e restrições, e por seu papel estrutural, na medida que atua como base para a criação de outros vocabulários.

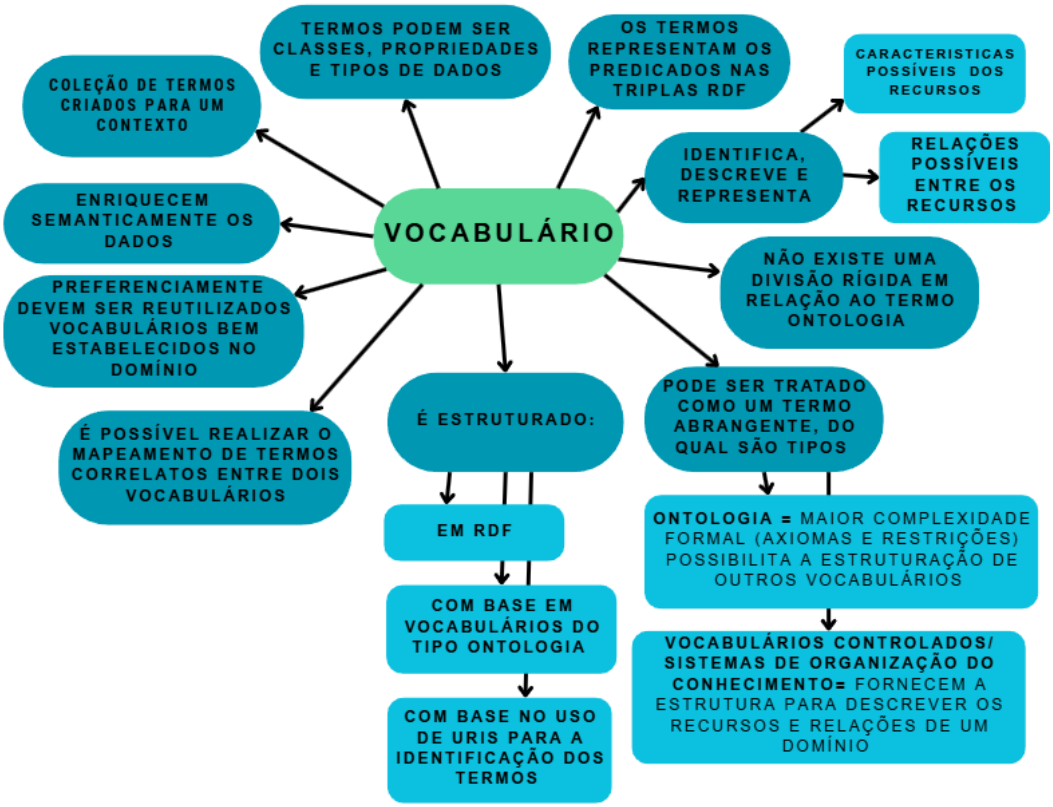
Outros termos também podem ser utilizados para referir-se a tipos específicos de vocabulários, sendo exemplos Vocabulários Controlados, Esquemas Conceituais

e Sistemas de Organização do Conhecimento. Esses termos são utilizados pela comunidade do *Linked Data* para se referir a vocabulários que:

> [...] enumeram e definem recursos que podem ser empregados nas descrições [...] ou seja, vocabulários que estruturam as descrições de recursos em conjuntos de dados (conectados). Um conceito de um tesauro, por exemplo, “arquitetura”, será usado, por exemplo, no campo de assunto para a descrição de um livro (onde “assunto” foi definido em uma ontologia para livros). Para definir os termos nesses vocabulários, formalismos complexos geralmente não são necessários (W3C, 2017, não paginado, tradução nossa).

Com base nas discussões e na fragmentação das definições selecionadas foi elaborada a sistematização apresentada na figura 16.

Figura 16 - Sistematização do termo vocabulário



Fonte: Autora (2025)

Com base nas discussões, na fragmentação das definições e na sistematização apresentada na figura 16, elaborou-se a seguinte definição:

Vocabulários são coleções de termos criados para um contexto específico. Os termos podem ser classes, propriedades e tipos de dados, que permitem identificar, descrever e representar características possíveis dos recursos de um domínio bem como as potenciais relações entre recursos. Os vocabulários são utilizados como predicados nas triplas RDF, permitindo o enriquecimento dos dados. Podem ser criados por qualquer pessoa para atenderem a necessidade de um domínio específico, entretanto, preferencialmente deve-se reutilizar um vocabulário já estabelecido. Podem ser feitos mapeamentos entre termos correlatos de distintos vocabulários. No contexto do *Linked Data* não existe uma diferenciação rígida entre os termos “vocabulário” e “ontologia”, sendo comumente utilizado pelo W3C e pela comunidade de maneira intercambiável. Entretanto, é possível considerar o termo vocabulário como um termo abrangente, do qual são tipos ontologias e vocabulários controlados/sistemas de organização do conhecimento. Nessa acepção, as ontologias são vocabulários caracterizados pela complexidade formal, pela presença de axiomas e restrições e pelo caráter estruturante, fornecendo a base para a descrição de outros vocabulários. Os vocabulários controlados/sistemas de organização do conhecimento, nesse contexto, fornecem a base para caracterizar e descrever as características dos recursos e as relações de um determinado domínio. Os vocabulários são elaborados em RDF, com base na estrutura de vocabulários do tipo ontologia, e utilizando URIs para representar os termos.

Apresentada a definição do termo Vocabulário, a próxima subseção apresenta as discussões e a definição do termo SPARQL.

5.2.6 SPARQL

O termo SPARQL é utilizado para se referir a “um conjunto de especificações do W3C “que fornece linguagens e protocolos para consultar e manipular conteúdo de grafos RDF [...]” (W3C, 2013, tradução nossa).

Esse conjunto de especificações é composto por um protocolo, uma linguagem de consulta, e por especificações para a criação de SPARQL *Endpoints*.

Os “SPARQL *Endpoints* são mecanismos de acesso por meio de consultas à base de dados utilizando a linguagem SPARQL.” (Isotani; Bittencourt, 2015, p. 82).

A linguagem de consulta SPARQL fornece os elementos necessários para a realização de buscas em grafos RDF individuais e em catálogos de dados (Heath; Bizer, 2011).

A busca em um *endpoint* SPARQL é realizada por meio da criação de uma consulta utilizando a linguagem SPARQL. “A maioria das formas de consulta SPARQL contém um conjunto de padrões triplos, denominado padrão básico de grafo. Padrões triplos são como triplos RDF, exceto que cada sujeito, predicado e objeto pode ser uma variável.” (W3C, 2013, não paginado, tradução nossa).

Com base nas discussões e na fragmentação das definições selecionadas foi elaborada a sistematização apresentada na figura 17.

Figura 17 - Sistematização do termo SPARQL

Fonte: Autora (2025)

Com base nas discussões, na fragmentação das definições e na sistematização apresentada na figura 17, elaborou-se a seguinte definição:

O SPARQL pode ser entendido como um conjunto de especificações e protocolos do W3C, composto por uma linguagem de consulta e por orientações para a criação de *Endpoints* SPARQL. Tem o objetivo de permitir a busca e manipulação de conjuntos de dados e catálogo de dados RDF. A linguagem de consulta permite diferentes tipos de buscas e é baseada no uso de triplas, como no RDF, entretanto, cada uma das partes da tripla <sujeito> <predicado> <objeto> pode ser substituída por uma variável.

Apresentada a definição do termo SPARQL, a próxima subseção apresenta as discussões e a definição dos termos relacionados com a Avalição de Qualidade de Dados *Linked Data.*

5.3 Termos relacionados à Avaliação De Qualidade de Dados Linked Data

A presente subseção reúne os termos *relacionados com* a Qualidade de dados *Linked Data*. Os termos apresentados são: *"Qualidade de Dados Linked Data;* Avaliação de qualidade de dados; Ferramentas de Avaliação de Qualidade de Dados; Modelo de Qualidade; Categorias de Qualidade; Dimensão; Critério; e Métrica.

Apresentam-se ainda os termos relacionados às principais categorias de qualidade de dados: Qualidade Contextual; Qualidade Intrínseca; Qualidade Representacional; Qualidade de Acessibilidade.

As próximas subseções representam cada um dos termos, sendo compostas pelas discussões, sistematização e definição dos termos.

5.3.1 Qualidade de dados *Linked Data*

Os termos "*Linked Data Quality*" e "*Quality of Linked Data*" são frequentemente adotados no *corpus* teórico levantado para essa pesquisa, entretanto, não são apresentadas definições para o termo, sendo em geral, fornecidas definições abrangentes focadas no termo qualidade de dados.

Buscando, portanto, compreender como as características dos dados *Linked Data* afetam o entendimento da qualidade de dados, e a ausência de definição, buscou-se a construção de uma definição pauta na sistematização das discussões realizadas. Para isso, foi considerada a estrutura estabelecida para ao termo qualidade de dados e incluindo os desafios e caraterísticas atribuídas a esse termo pelo domínio do *Linked Data.*

Com base na análise dos enfoques dos estudos identificados, é possível observar que enquanto **um domínio**, a qualidade de dados *Linked Data* é composta também por uma comunidade interdisciplinar, focada principalmente na construção e discussão de ferramentas e metodologias para a avaliação e melhoria da qualidade de dados. Dedica-se ainda a avaliar e comparar fontes de dados *Linked Data* relacionadas a distintas áreas, sendo a maior parte dessas avaliações de abordagem contextual, com o objetivo de avaliar conjuntos de dados de domínios específicos, como dados governamentais, dados da área da saúde, dados bibliográficos e enciclopédicos.

Enquanto um problema ou desafio a ser superado, a qualidade de dados possui desafios muito caraterísticos, que podem ser organizados em problemas relacionados: com a estrutura dos dados, com as fontes publicadoras e com o processo de avaliação desses dados.

Os problemas relacionados com estrutura têm relação com a aplicação incorreta dos princípios e boas práticas que circundam esse domínio, com falhas na estruturação em RDF, na criação de URIs, na escolha ou criação dos vocabulários e ainda com o desrespeito as características e restrições das propriedades. Destacam-

se ainda problemas relacionados ao processo de conversão, considerando que a maior parte dos dados *Linked Data* provém da conversão de dados legados.

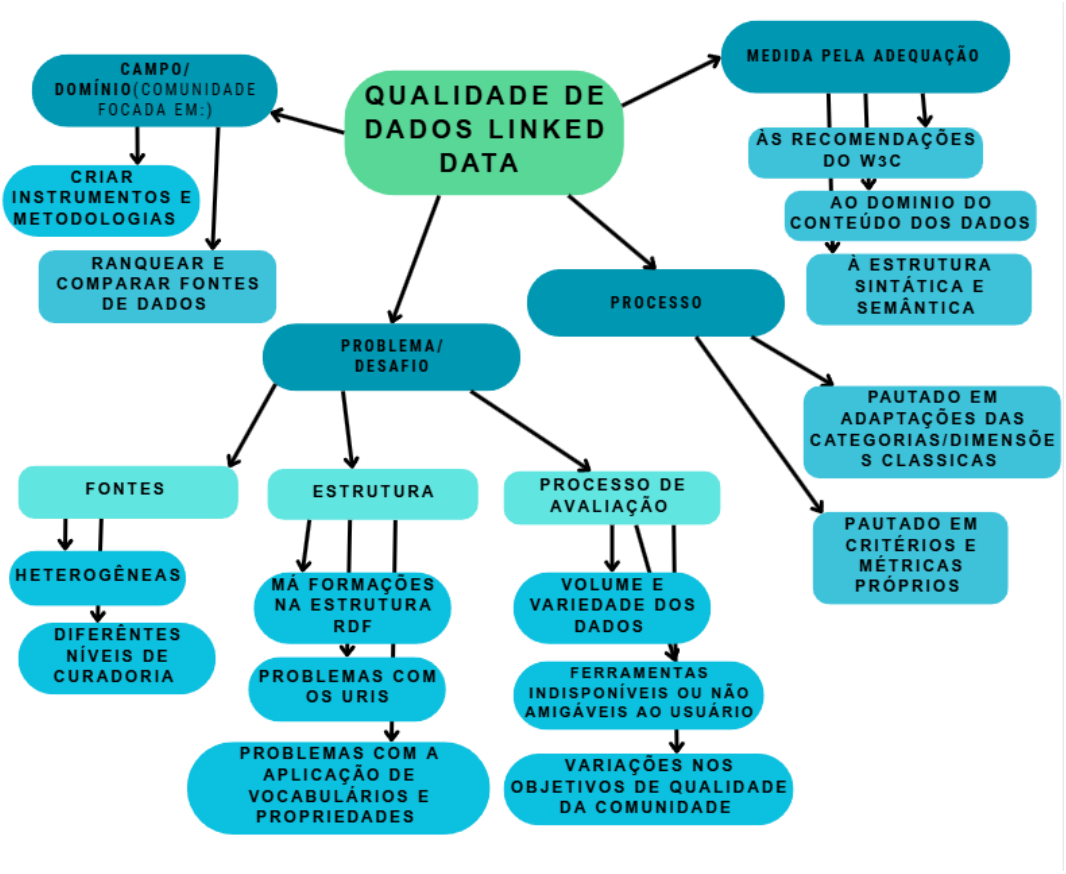
Em relação as fontes dos dados, os problemas de qualidade de dados se relacionam com o processo de seleção das fontes, que são heterogenias quanto aos seus domínios e propósitos, e por isso possuem diferentes níveis de curadoria. Tem-se ainda o desafio ligado política de “publicar primeiro e melhorar depois”, fortemente adotada pela comunidade.

O processo de avaliação enfrenta ainda problemas relacionados com a diversidade de domínios. Mesmo se tratando de dados estruturados seguindo um mesmo princípio, os dados *Linked Data* são heterogêneos em seu contexto, fazendo com que as ferramentas tenham que ser adaptadas também a aos domínios dos quais os conteúdos dos dados derivam. Tem-se ainda questões relacionadas com o volume e variedade desses dados e às limitações das ferramentas existentes, que em muitas situações se encontram indisponíveis ou são pouco amigáveis aos usuários.

Enquanto um processo, a avaliação de qualidade de dados *Linked Data* adota, em geral, a estrutura proposta por Wang e Strong (1996) ou ainda a estrutura proposta pelas normas ISO para qualidade de dados vigentes. São utilizadas as categorias e dimensões clássicas da qualidade de dados, embora o significado e as definições dessas categorias e dimensões seja afetado pela estrutura e pelas características da comunidade de dados *Linked Data.* Os critérios e as métricas adotados para avaliar a qualidade são específicos desse domínio, como poderá ser observado na análise dos modelos de qualidade disponíveis para dados *Linked Data.*

Com base nessa análise, foi construída uma sistematização para a definição do termo, apresentada na figura 18.

Figura 18 - Sistematização do termo “Qualidade de Dados *Linked Data*”



Fonte: Autora (2025)

Com base na sistematização da literatura e seguindo a estrutura da definição de qualidade de dados, entende-se que a Qualidade de dados *Linked Data* pode contar com as seguintes definições:

> O termo qualidade de dados *Linked Data* refere-se às implicações da adoção do *Linked Data* na qualidade dos dados, e pode contar com as seguintes acepções: **1. Enquanto um campo ou domínio** a Qualidade de Dados *Linked Data* possui uma comunidade própria, heterogênea e interdisciplinar, focada principalmente na criação de ferramentas, metodologias e modelos para possibilitar a realização dos processos de avaliação e controle de qualidade de dados. A comunidade também se concentra em avaliar e comparar os níveis de qualidade das fontes de dados disponíveis. **2. Enquanto um problema a ser superado**, a Qualidade de Dados *Linked Data* possui desafios relacionados principalmente com as fontes de dados, com a estrutura dos dados e com o próprio processo de avaliação. As fontes de dados são heterogêneas e possuem diferentes níveis de curadoria dos dados. Os problemas de estrutura em dados *Linked Data* são relacionados principalmente com aplicação incorreta do RDF, a criação de URIs inconsistente e com problemas relacionados a seleção e aplicação de vocabulários e propriedades. Em relação ao processo de avaliação de qualidade, os dados *Linked Data* enfrentam problemas relacionados ao volume e variedade dos dados, às ferramentas que estão indisponíveis ou não são amigáveis para os usuários, e as variações nos

> objetivos de qualidade da comunidade, que dificultam a criação de ferramentas únicas e generalistas. **3. Enquanto uma medida** a Qualidade de Dados *Linked Data* pode ser medida com base na adequação dos dados aos princípios e melhores práticas disponibilizados pelo W3C e pela comunidade ao qual os dados se relacionam, pode ser medida em relação a adequação ao uso ou ainda em relação a sua adequação sintática. **4. Enquanto um processo a avaliação da Qualidade de dados *Linked Data*,** em sua maioria, adota a estrutura proposta por Wang e Strong (1996) ou a estrutura proposta pela norma ISO de qualidade de dados vigente. São adotadas as categorias e dimensões clássicas, sendo adaptadas as suas definições para o contexto. São criados critérios e dimensões próprios para o contexto do *Linked Data*, levando em consideração principalmente sua estrutura pautada em triplas seguindo o modelo RDF.

Apresentada a definição do termo Qualidade de Dados Linked Data, a próxima subseção apresenta as discussões e a definição do termo 2 Avaliação de Qualidade de Dados

5.3.2 Avaliação de qualidade de dados

Em relação ao seu gênero, a avaliação de qualidade pode ser definida como o "[...] processo de mensuração de diferentes critérios de qualidade dos dados" (Nahari *et al.*, 2017, p.68, tradução nossa).

A avaliação de qualidade pode ser entendia ainda como uma atividade ou tarefa (Zaveri *et al.*, 2015; Nayak *et al.*, 2021) que integra outros processos como o de gestão de dados e o controle de qualidade. Nessa perspectiva, a avaliação de qualidade pode ser uma tarefa pontual ou cíclica, necessária em diversos momentos do ciclo de vida dos dados (Rula, 2011; Nooghabi; Dastgerdi, 2016).

A avaliação de qualidade pode ser conduzida com o objetivo de selecionar fontes de dados em potencial ou permitir a identificação e correção de problemas de qualidade.

> "[...] a avaliação da qualidade é necessária antes de usar os dados para uma determinada tarefa, a fim de garantir que os dados tenham um nível de qualidade adequado. Além disso, os resultados da avaliação podem ser usados para auxiliar o processo de melhoria da qualidade, e corrigindo deficiências nos dados (Mihindukulasooriya *et al.*, 2017, tradução nossa).

Nesse contexto, a avaliação de qualidade possui dois agentes principais: consumidores (usuários) e publicadores, onde respectivamente um está mais focado na atividade de seleção de fontes e o outro na identificação e correção de problemas

de qualidade. “O objetivo da atividade de avaliação da qualidade dos dados é analisar a relevância de um conjunto de dados para seus consumidores e auxiliar na publicação de dados de melhor qualidade” (Nayak *et al*., 2021, p. 3).

A avaliação de qualidade pode ser organizada de maneira hierárquica, seguindo a estrutura de categorias, dimensões e métricas (Arruda *et al*., 2019). Essa estrutura hierárquica se baseia no *framework* proposto por Wang e Strong (1996), o nível mais alto, e, portanto, menos granular, da qualidade são as categorias, que organizam dimensões de qualidade, que são compostos critérios de qualidade que podem ser mensurados por uma ou mais métrica de qualidade, sendo as métricas o nível mais granular da qualidade.

A avaliação de qualidade “envolve a formulação de aspectos necessários em termos de métricas de qualidade como indicadores e o teste de conjuntos de dados em relação a esses requisitos de qualidade” (Langer *et al*., 2018, p. 164, tradução nossa).

Rula (2011) aponta como principais etapas da avaliação de qualidade: a definição das dimensões de qualidade, o estabelecimento e aplicação das métricas e a análise dos resultados. A definição das dimensões e métricas a serem utilizadas se baseia na criação ou aplicação de um modelo de qualidade.

A avaliação de qualidade consiste na “[...] análise de dados para mensurar a qualidade dos conjuntos de dados em dimensões de qualidade relevantes” (Zaveri *et al*, 2015, p. 7, tradução nossa). Pode envolver “a comparação entre as medições obtidas e os valores de referência, a fim de permitir um diagnóstico de qualidade” (Zaveri *et al*, 2015, p. 7, tradução nossa) ou a comparação entre os resultados de distintos conjuntos de dados.

Em ambas as acepções, a avaliação de qualidade é intrinsecamente multifacetada, ou ainda multidimensional, tendo suas etapas, atividades e objetivos afetados pelas distintas perspectivas da qualidade de dados, que são refletidas nas categorias propostas por Wang e Strong (Wang; Strong, 1996; Assaf; Senart; Troncy, 2016; Issa *et al.*; 2021).

Quando abordado em uma perspectiva contextual, pode ser entendido como o processo ou a atividade “[...] de avaliar se um dado atende às necessidades dos consumidores em um caso de uso específico” (Assaf; Senart; Troncy, 2016, p. 112, tradução nossa). Pode ainda ser abordada em uma perspectiva intrínseca, onde a

avaliação de qualidade busca verificar se os dados estão livres de erros, más formações (Wang; Strong, 1996).

Mesmo em uma perspectiva intrínseca, onde a avaliação se concentra em caraterísticas diretamente relacionadas dos conjuntos de dados “para avaliar a qualidade de qualquer conjunto de dados, é fundamental definir as dimensões de qualidade com base no domínio de uso” (Behkamal *et al*., 2014, p. 2).

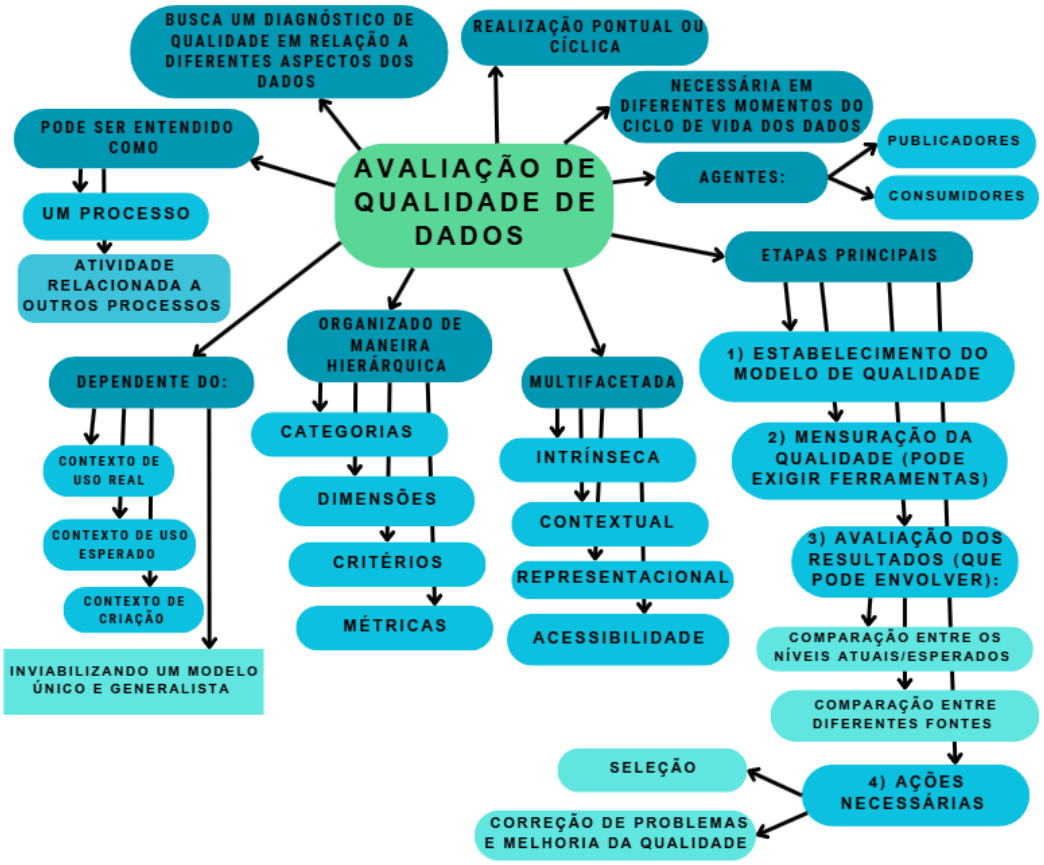
Deve-se, portanto, considerar domínio de criação e uso, mesmo que se tenha apenas a previsão de uso esperada pelos publicadores, pois a forma como esses dados são estruturados, os padrões adotados e as políticas desse domínio afetam o conceito de qualidade e terão impacto inclusive na avaliação dos níveis de qualidade sintáticos e semânticos dos dados.

No contexto do *Linked Data*, precisam ser levados em consideração tanto os princípios e melhores práticas do W3C, quanto os princípios, melhores práticas e normas relacionados ao domínio ao qual o conteúdo dos dados se relaciona. Justamente por essa característica, é difícil criar modelos de avaliação generalistas, tornando “controverso quais métricas de qualidade são de maior interesse e se um conjunto básico de métricas de uso geral faz sentido[...]” (Langer *et al.*, 2018, p. 164, tradução nossa).

A avaliação se beneficia do uso de ferramentas, que podem ser automáticas, semiautomáticas ou manuais (Kahlawi, 2020, p. 61).

Com base na fragmentação das definições e na análise da literatura, realizou-se a sistematização apresentada na figura 19.

Figura 19 -Sistematização do termo avaliação de qualidade



Fonte: Autora (2025)

Com base na fragmentação das definições e na sistematização apresentada na figura 19, elaborou-se a seguinte definição:

A avaliação de qualidade de dados pode ser definida como um processo ou como uma atividade realizada no âmbito de outros processos, como gestão e controle de qualidade de dados, que busca realizar um diagnóstico da qualidade dos dados em relação a diferentes aspectos preestabelecidos. A avaliação pode ser realizada de maneira pontual ou cíclica, sendo necessária em diversos momentos do ciclo de vida dos dados. É organizada de maneira hierárquica, composta por categorias, dimensões, critérios e métricas. É considerada multifacetada, pois pode ser avaliada com base em diferentes perspectivas, sendo elas: intrínseca, contextual, representacional e acessibilidade. A avaliação de qualidade possuí quatro etapas principais: estabelecimento do modelo de qualidade, mensuração, avaliação dos resultados e realização de atividades necessárias. O modelo de qualidade estabelece quais dimensões, critérios e métricas serão avaliados. Para mensurar a qualidade podem ser necessárias ferramentas automáticas, semiautomáticas e/ou manuais. A avaliação dos resultados pode ser feita por meio de uma comparação entre os resultados obtidos e os resultados esperados, que devem ser pré-estabelecidos. Pode ainda ter como base a comparação dos resultados de diferentes conjuntos de dados. Ao final da avaliação, os resultados podem ser utilizados para selecionar fontes de

dados ou para identificar e corrigir problemas de qualidade. Os principais agentes da avaliação de qualidade são os consumidores e os publicadores. A avaliação de qualidade depende do domínio de criação e uso dos dados, ou da previsão de uso esperada pelos publicadores. No contexto do *Linked Data*, precisam ser levados em consideração tanto os princípios e melhores práticas do W3C, quanto os relacionados ao domínio do conteúdo dos dados. Essa característica diminui a relevância de modelos de qualidade muito abrangentes ou genéricos.

Apresentada a definição do termo Avaliação de Qualidade de Dados, a próxima subseção apresenta as discussões e a definição do termo Ferramenta de Avaliação de Qualidade de Dados.

5.3.3 Ferramenta de avaliação de qualidade de dados

Ao fragmentar as definições encontradas para ferramentas de avaliação de qualidade de dados, observou-se que a maioria se concentra em apresentar enunciados acidentais e informativos a respeito do funcionamento desse tipo de ferramenta.

Partiu-se então da definição da palavra ferramenta, que pode ser entendida como “Conjunto de instrumentos, peças e utensílios empregados num ofício ou num trabalho manual ou mecânico (Priberam, 2025, não paginado). Pode ser definida ainda como “Aquilo que serve de meio ou auxílio para determinado fim (Priberam, 2025, não paginado).

Nesse contexto, as ferramentas de avaliação de qualidade podem ser compreendidas como instrumentos criados para realizar ou auxiliar na avaliação da qualidade de dados.

Para que essas ferramentas possam auxiliar na avaliação de qualidade, elas:

> [...] devem abordar e mensurar os diversos aspectos da qualidade que levam a dificuldades na comparação, no *benchmarking* e na avaliação dos resultados, juntamente com a seleção da fonte de dados adequada com base nas necessidades de qualidade (Rani; Sapna; Mishra, 2018, p. 35).

Considerando a complexidade da avaliação de qualidade, as ferramentas “são caracterizadas por uma alta diversidade em termos de características e medidas avaliadas” (Radulovic *et al.,* 2017, p. 4-5).

Elas podem variar em relação aos seus objetivos e às atividades que realizam. Elas podem se concentrar em uma dimensão específica, na avaliação genérica da

qualidade dos dados, ou ainda na inspeção exploratória de problemas de qualidade (Radulovic *et al.,* 2017).

As ferramentas podem variar ainda em relação ao "[...] número de métricas para avaliar a qualidade, as abordagens para processar os dados, o tipo de dados usados para avaliar, a flexibilidade do usuário para escolher a métrica e o peso correspondente e o relatório de avaliação" (Nayak *et al*. 2021, p. 5).

Outra variação importante em relação as ferramentas de avaliação de qualidade têm relação com a forma como desempenham suas atividades, podendo ser classificadas em automáticas, semiautomáticas e manuais (Kontokostas *et al*., 2013; Nayak *et al*, 2021; Salem; Benchikha, 2022).

Em relação a essa variação pode-se entender que nas ferramentas:

> Automáticas: os recursos selecionados são fornecidos como entrada para uma ferramenta automática que realiza a avaliação da qualidade sem qualquer envolvimento do usuário. Semiautomáticas: os recursos selecionados são atribuídos a uma ferramenta semiautomática que realiza a avaliação da qualidade dos dados empregando alguma forma de *feedback* do usuário. Manuais: os recursos selecionados são atribuídos a uma pessoa (ou grupo de indivíduos) que então procederá à avaliação manual dos recursos individualmente (Kontokostas *et al*., 2013, p.3).

Radulovic *et al.* (2017) acrescenta que diversas ferramentas foram desenvolvidas especificamente para o contexto do *Linked Data*, o que também pode ser observado na análise teórica da qualidade de dados.

Com base na análise e na categorização das ferramentas identificadas, foi possível observar que elas variam principalmente em relação ao tipo, a atividade que desempenham (avaliação; melhoria; avaliação e melhoria; e comunicação dos resultados), a forma como desempenham essas atividades (automática, semiautomática e manual), quanto as dimensões e métricas avaliadas e ainda ao público a que se destinam (publicadores, consumidores ou gerais).

Com base na análise e fragmentação das definições e na análise teórica da qualidade de dados *Linked Data*, realizou-se a sistematização apresentada na figura 20.

Figura 20 - Sistematização do termo “ferramenta de avaliação de qualidade”

Fonte: Autora (2025)

Com base na análise das ferramentas identificadas, na fragmentação das definições e na sistematização apresentada na figura 20, construiu-se a seguinte definição:

> As ferramentas de avaliação de qualidade de dados são instrumentos criados para auxiliar ou realizar completamente a avaliação de qualidade de dados. Existe uma grande pluralidade em relação as ferramentas, com diferentes tipos, que possuem nomenclaturas próprias. Elas variam em relação ao domínio para o qual foram criadas, existindo ferramentas para domínios específicos e ferramentas de aspecto geral. Existem diversas ferramentas criadas especificamente para a avaliação de dados *Linked Data*. As ferramentas variam quanto as dimensões e métricas avaliadas, existindo ferramentas focadas em dimensões específicas, permitindo ou não a personalização dessas dimensões e métricas e do peso que cada uma delas tem em relação ao resultado da avaliação. Elas podem desempenhar diferentes atividades, como avaliação, a avaliação e correção dos problemas de qualidade identificados ou ainda concentrar-se na exportação de resultados do processo de avaliação de qualidade, como é caso dos vocabulários de qualidade de dados. Variam quanto a forma como desempenham essas atividades, existindo ferramentas com abordagens automáticas, semiautomáticas e manuais. Elas podem ser criadas ainda para um público específico, existindo ferramentas focadas nas necessidades dos consumidores e dos publicadores

bem como ferramentas voltadas para ambos, sendo mais comum a criação de ferramentas focadas em publicadores de dados.

Apresentada a definição do termo Ferramenta de Avaliação de Qualidade de Dados, a próxima subseção apresenta as discussões e a definição do termo Modelo de Qualidade

5.3.4 Modelo de qualidade

Embora diversos modelos tenham sido identificados na análise teórica da qualidade de dados *Linked Data*, poucas definições foram apresentadas para o termo. A ISO25012 enfatiza que um modelo de qualidade é a base para a realização da avaliação de qualidade dos dados.

Candela *et al.* (2021) aponta que os modelos fornecem as **características de qualidade** e as **medidas** relacionadas aos dados, juntamente com fórmulas para calcular as medidas mencionadas.

Radulovic (2017, p. 3, tradução nossa, grifo nosso) aponta que um modelo de qualidade é composto por "[...] um conjunto de **características** específicas de qualidade, **sub-características** de qualidade, **medidas** de qualidade e pelas relações entre essas características e medidas".

Já a ISO25012 (2022, não paginado, tradução nossa) indica que no modelo de qualidade "são estabelecidas as principais características de Qualidade de Dados que devem ser consideradas na avaliação das propriedades do produto de dados pretendido".

ISO/IEC 25012 (2022, não paginado, tradução nossa, grifo nosso) aponta ainda que um modelo de qualidade é um "Conjunto definido de **características** que fornece uma estrutura para especificar **requisitos** de qualidade de dados e **avaliar** a qualidade dos dados".

Nesse sentido, é possível observar pela análise das definições que um modelo de qualidade deve, necessariamente, indicar quais as características que serão avaliadas em relação a qualidade dos dados e as formas utilizadas para medi-las.

Traçando um paralelo com a estrutura hierárquica na qual se organiza a avaliação de qualidade de dados, as características específicas a serem avaliadas equivalem às dimensões de qualidade, enquanto as sub-características, ou requisitos,

podem ser entendidas como os critérios de qualidade a serem avaliados. As medidas, ou formas de avaliação, equivalem as métricas de qualidade.

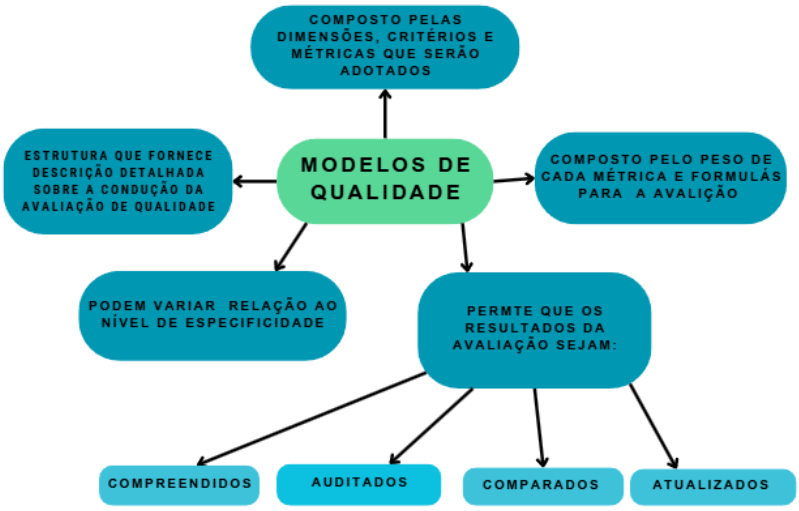
Para além de indicarem quais dimensões e métricas serão adotadas, os modelos fornecem “detalhes importantes sobre as medidas de qualidade, como definições, escalas ou fórmulas, os modelos de qualidade fornecem uma orientação sobre quais medidas são importantes para a avaliação e como medi-las” (Radulovic, 2017, p.3, tradução nossa).

Radulovic (2017, p.3) acrescenta que os modelos de qualidade são importantes por “fornecer terminologia e orientação consistentes para a avaliação da qualidade”.

Essa descrição detalhada do processo é fundamental para a compreensão dos resultados, possibilitando que esses resultados sejam posteriormente auditados e atualizados. Além disso, o fornecimento das dimensões adotadas e da terminologia relacionada ao processo permite que os resultados sejam compartilhados, reutilizados e comparados.

Com base nas discussões e na fragmentação das definições selecionadas, elaborou-se a sistematização apresentada na figura 21.

Figura 21 -Sistematização do termo “modelos de qualidade”



Fonte: Autora (2025)

Com base na fragmentação das definições e na sistematização apresentada na figura 21, elaborou-se a seguinte definição para compor o glossário:

> Um modelo de qualidade é uma estrutura que fornece descrição detalhada sobre como será conduzida a avaliação de qualidade. É composto pelas dimensões,

> critérios e métricas que serão utilizados e por detalhes relevantes, como os pesos das métricas, definições e orientações. O fornecimento detalhado dessas informações é a base da avaliação de qualidade, sendo fundamental para que os resultados possam ser compreendidos, auditados, atualizados, compartilhados e comparados. Os modelos de qualidade podem variar em relação ao nível de especificidade, existindo modelos genéricos e modelos criados para aplicação apenas em um cenário específico de uso de dados. Mesmo os modelos genéricos precisam ser ajustados para aplicação, geralmente são mantidas as dimensões, sendo feitos ajustes nos critérios e especialmente nas métricas e nos pesos que recebem essas métricas em relação ao resultado geral da avaliação de qualidade.

Apresentada a definição do termo Modelo de Qualidade, a próxima subseção apresenta as discussões e a definição do termo Categoria de Qualidade.

5.3.5 Categoria de qualidade

O termo “categoria de qualidade” é fundamental para a compreensão da estrutura por meio da qual se organiza hierarquicamente o processo de avaliação de qualidade. Apesar de sua importância, poucas definições completas para o termo foram identificadas na literatura analisada.

A maior parte dos autores mencionam as categorias de qualidade visando apenas apresentar a estrutura de quatro categorias de Wang e Strong (1996).

Os autores mencionam que as categorias foram previamente elaboradas para agrupar as dimensões de qualidade em famílias. O processo de elaboração dessas categorias partiu, inicialmente, de uma análise da literatura de qualidade de dados, visando a criação de categorias iniciais para agrupar por semelhança os diferentes aspectos da qualidade de dados (Wang; Strong, 1996).

Os autores estabeleceram um modelo inicial de organização para as dimensões, baseado em quatro categorias. Esse modelo foi testado com dois grupos distintos de participantes, que receberam as dimensões, e buscaram organizá-las em categorias, rotulando essas categorias. Com base nos resultados obtidos Wang e Strong realizaram ajuste “Como resultado desse reexame, renomeamos duas das quatro categorias. As categorias resultantes, portanto, são: DQ intrínseca, DQ contextual, DQ representacional, DQ acessibilidade”. (Wang; Strong, 1996, p. 19, tradução nossa).

Essa estrutura de organização de dimensões segue sendo amplamente adotada pela literatura de qualidade de dados. Cada uma dessas categorias será analisada individualmente.

Para construir a definição termo, buscou-se realizar uma análise de como os autores se refere a essas categorias. O quadro 17 apresenta os resultados dessa análise.

Quadro 17 - Análise das abordagens de categorias de qualidade

Termo	Autores
Classificação de dimensões, critérios ou métricas	Cherix *et al.* (2014); Langer *et al.* (2018); Acosta *et al.* (2018); Nayak *et al.* (2022)
Estrutura conceitual	Wang e Strong (1996)
Grupos de dimensões	Nahari *et al.* (2017); Ahmed (2017)
Categorias que organizam dimensões de qualidade	Elbattah e Ryan (2019)
Coleção selecionada de métricas	Zhang *et al.* (2023
Entidades abstratas	W3C (2017)

Fonte: Autora (2025)

Como é possível observar no quadro 17, as categorias de qualidade são frequentemente definidas por sua relação com seus termos subordinados, sendo eles dimensões, critérios e métricas. Com base nos termos utilizados pelos autores tornam-se relevantes a busca de definições para os termos categoria, classe e classificação.

Com base no dicionário de biblioteconomia e arquivologia (Cunha, Cavalcante, 2008, p. 74) o termo categoria pode ser definido como: “Classe fundamental (ou básica) que resulta da divisão do universo de conhecimentos, de acordo com as características intrínsecas, ou fundamentais, de cada conceito”. Enquanto classe é definido pelos autores como “conjunto de itens que possuem pelo menos uma característica em comum” (Cunha; Cavalcante, 2008, p. 83).

Os autores definem ainda a classificação como:

> Agrupamento real, ou ideal, daquilo que é semelhante e a separação do que é diferente. Em geral, a classificação é o ato da divisão, em várias classes, de um conjunto de objetos. É também o produto que resulta da operação precedente, quando esta dá como resultado um sistema coerente e estruturado (Cunha; Cavalcante, 2008, p. 74).

Em relação às três definições levantadas, destacam-se alguns aspectos: 1) categorias podem ser consideradas como classes fundamentais ou basilares, 2)

classes agrupam itens que possuem características em comum, e 3) o ato de classificar permite agrupar pela semelhança e individualizar pelas diferenças.

Em uma definição mais completa de categoria de qualidade, Zhang *et al.* (2023, p. 4, tradução nossa, grifo nosso) apontam que “uma categoria de qualidade é definida como uma **coleção selecionada** de métricas de qualidade objetivas que **refletem questões de qualidade semelhantes**, com um escopo mais restrito do que uma dimensão de qualidade”.

O W3C (2017, não paginado, tradução nossa, grifo nosso) indica ainda que as categorias de qualidade “Representam um grupo de dimensões de qualidade nas quais um **tipo comum de informação** é usado como **indicador de qualidade**.”

Nesse sentido, pode-se definir que as categorias são as classes basilares que permitem organizar a avaliação de qualidade. O que está sendo agrupado com base em semelhança são questões de qualidade, ou os diferentes aspectos de qualidade que podem ser analisados e melhorados em conjuntos de dados.

As categorias são abstrações, mais genéricas e abrangentes que dimensões, que por sua vez possuem certo nível de abstração em relação aos critérios de qualidade. A questão do caráter abstrato das categorias é apontada pelo W3C (2017, não paginado, tradução nossa): “Dimensão e categoria são entidades abstratas”.

Essas entidades abstratas representam os dois níveis possíveis de organização de aspectos de qualidade, sendo os critérios representantes mais concretos, focados em aspectos específicos e mensuráveis da qualidade.

Com base na fragmentação das definições e nas discussões realizadas elaborou-se a sistematização apresentada na figura 22.

Figura 22 - Sistematização do termo “categoria de qualidade”

Fonte: Autora (2025)

Com base nas definições analisadas e na sistematização apresentada na figura 22, elaborou-se a seguinte definição para compor o glossário:

> Uma categoria de qualidade é uma classe que permite a organização e o agrupamento de distintos aspectos de qualidade com base em sua semelhança. É considerado o nível mais abrangente na hierarquia da avaliação de qualidade. As categorias são compostas por dimensões, que agrupam critérios, que podem ser mensurados com base em métricas de qualidade. Como categorias e dimensões possuem um caráter abstrato, para se avaliar os níveis de qualidade dos dados em relação a determinada categoria torna-se necessário a aplicação de um conjunto de métricas, indicadores que permitem mensurar quantitativa e qualitativamente a qualidade dos dados. A literatura tradicionalmente organiza o processo de avaliação de qualidade e quatro categorias, sendo elas: Intrínseca, Contextual, Acessibilidade e Representacional.

Apresentada a definição do termo Categoria de Qualidade, a próxima subseção apresenta as discussões e a definição do termo Categoria Contextual.

5.3.9 Categoria contextual

É um consenso na literatura que a qualidade contextual se define por meio de sua dependência em relação a tarefa na qual serão empregados os dados. (Ahmed,

2017; Farber *et al.*, 2017; Acosta *et al.*, 2018; Zaveri *et al.*, 2015; Debattista *et al.*, 2018; Assaf; Troncy; Senart, 2015; Nahari *et al.*, 2017; Gurdur; El-khoury; Nyberg, 2019; Abian et al, 2018; Nayak; Bozic; Longo, 2021; Wang; Strong, 1996).

Como discutido durante a análise do termo qualidade de dados, a literatura atual da área costuma definir qualidade de dados por meio do conceito de "*fitness for use*" ou adequado ao uso. Entretanto, essa definição limita a qualidade de dados a uma perspectiva contextual.

Na perspectiva contextual, a qualidade dos dados está relacionada com os objetivos e com as tarefas que o usuário pretende executar com esses dados. Nessa perspectiva, um conjunto de dados pode ser perfeitamente adequado para determinada tarefa e não atuarem satisfatoriamente em outro contexto (Juran, 1988; Wang; Strong, 1996; Zaveri *et al.*, 2012).

Em uma perspectiva contextual, discutir, avaliar, mensurar e melhorar a qualidade de dados são ações que dependem das características do domínio e das necessidades do contexto em que serão empregados esses dados. Um conjunto de dados é considerado de boa qualidade quando atende satisfatoriamente às necessidades do domínio ou de seu contexto de aplicação.

Debattista *et al.* (2018, p. 877, tradução nossa) destacam que "a categoria contextual agrupa as dimensões e métricas que são altamente dependentes da tarefa em questão". Abian *et al.* (2018) acrescentam ainda 3 fatores que impactam no estabelecimento de dimensões contextuais: 1) o usuário que irá realizar a tarefa, 2) o tempo no qual essa tarefa tem que ser realizadas e 3) o ambiente no qual ela será realizada. Essa dependência pode agregar um alto nível de subjetividade a análise das dimensões contextuais.

Para viabilizar a análise da qualidade dos dados com base em aspectos contextuais, os autores indicam que "uma abordagem é parametrizar dimensões contextuais para cada tarefa, de modo que um consumidor de dados possa especificar o tipo de tarefa que está sendo executada e os parâmetros contextuais apropriados para essa tarefa" (Wang; Strong, 1996, p. 17).

No contexto do *Linked Data*, os domínios de origem e os contextos de aplicação em potencial dos dados são diversos, com dados que abordam desde a criação de páginas *Web* até dados a serem empregados na área da saúde, que exigem alto rigor em relação a aspectos de qualidade.

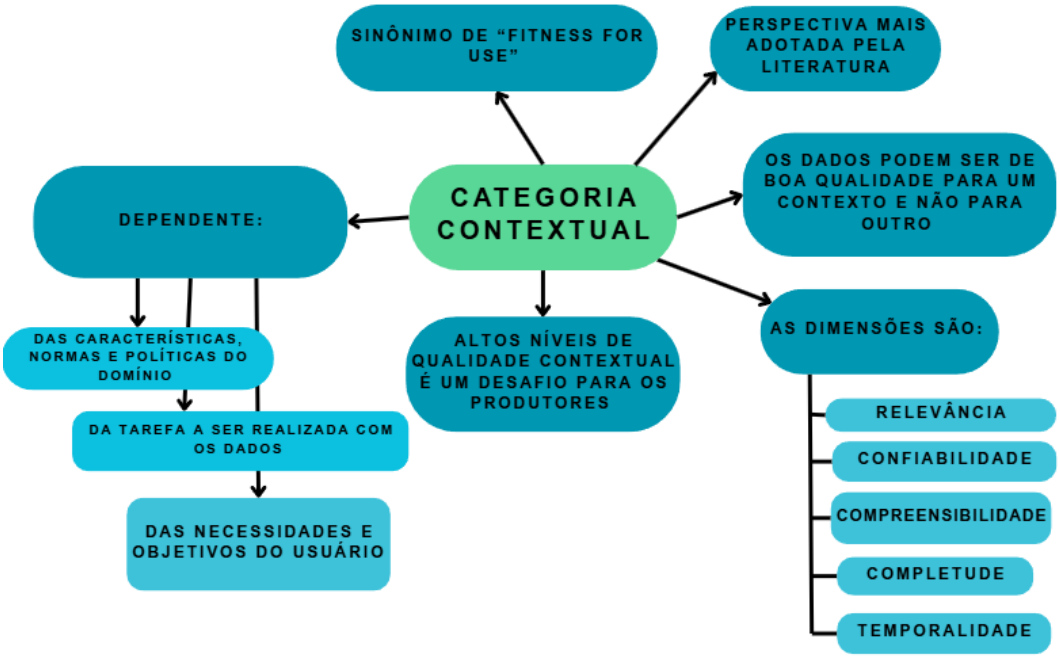
Além disso, considerando que os dados são criados em um paradigma de compartilhamento, em muitas situações se torna complexo prever em que contexto esses dados serão empregados no futuro (Helth; Bizer, 2011).

Esse aspecto reforça o desafio da qualidade contextual para os produtores de dados. “Como as tarefas e seus contextos variam ao longo do tempo e dos consumidores de dados, atingir alta qualidade de dados contextuais é um desafio de pesquisa” (Wang; Strong, 1996, p. 17, tradução nossa).

No contexto do *Linked Data*, as principais dimensões contextuais são: relevância, confiabilidade, compreensibilidade, completude e temporalidade (Zaveri, 2012; Farber *et al.*, 2016; Debattista *et al.*, 2018).

Com base na fragmentação das definições e na análise da literatura, realizou-se a sistematização apresentada na figura 23.

Figura 23 - Sistematização do termo “Categoria Contextual”



Fonte: autora (2025)

Com base na fragmentação das definições e na sistematização apresentada na figura 23, elaborou-se a seguinte definição:

A categoria contextual permite a organização e o agrupamento de dimensões cuja avaliação depende das características, políticas e boas práticas do domínio de

criação/uso dos dados; da tarefa que será realizada com esses dados; e/ou das necessidades e objetivos do usuário. É a perspectiva mais adotada pela literatura, sinônimo da expressão "*fitness for use*". Nessa perspectiva um conjunto de dados pode possuir boa qualidade para um contexto e não ser adequado para outro. Por sua dependência de fatores externos, é um desafio para os produtores atingirem altos níveis de qualidade contextual. No contexto do *Linked Data*, os domínios de origem e os contextos de aplicação em potencial dos dados são diversos e o paradigma de compartilhamento que circunda a publicação desses dados torna ainda mais complexo prever em que cenário esses dados serão empregados no futuro, ampliando esse desafio. As principais dimensões contextuais de qualidade de dados são: relevância, confiabilidade, compreensibilidade, completude e temporalidade.

Apresentada a definição do termo Categoria Contextual, a próxima subseção apresenta as discussões e a definição do termo Categoria Intrínseca.

5.3.10 Categoria intrínseca

As definições da categoria intrínseca, em geral, são baseadas em sua oposição em relação a categoria contextual.

Ahmed (2017) aponta que na categoria intrínseca, a qualidade dos dados é definida de maneira independente de fatores externos. *Acosta et al.* (2018), Zaveri *et al.* (2015) e Abian *et al.* (2018) apontam que a qualidade nessa categoria é independente do contexto do usuário.

Abian *et al.* (2018) acrescenta ainda que nessa categoria a qualidade dos dados não depende da tarefa em que serão empregados. Ahmed (2017, p. 3, tradução nossa) aponta que "As dimensões intrínsecas permitem definir a qualidade dos dados do ponto de vista do provedor de dados".

Buscou-se então, com base nas definições identificadas e analisadas, compreender a qualidade intrínseca de maneira independe da qualidade contextual.

De acordo com Wang e Strong (1996), corroborado por Farber *et al.* (2017) a qualidade intrínseca está relacionada a dados que possuem qualidade por si só. Para Nahari *et al.* (2017) a qualidade intrínseca está relacionada a características inerentes aos dados.

Farber *et al.* (2017, p. 3, tradução nossa), indicam que as dimensões intrínsecas buscam verificar se os dados "[...] refletem a realidade e são logicamente consistentes". Zaveri *et al.* (2015) apontam que as dimensões agrupadas na categoria intrínseca se concentram em verificar se os dados estão sintática e semanticamente corretos e são logicamente consistentes.

Debattista *et al.* (2018) apontam que as dimensões intrínsecas buscam verificar se os dados estão corretos e coerentes.

Gurdur; El-khoury; Nyberg (2019) ressaltam que a categoria intrínseca busca verificar se os dados estão corretos e acrescentam ainda necessidade de verificar se os dados estão compactos e completos.

Nayak, Bozic e Longo (2021), Abian *et al.* (2018) e Zaveri *et al.* (2015) mencionam que as dimensões da qualidade intrínseca buscam avaliar se os dados representam de maneira adequada, correta, consistente e completa o mundo real.

Wang e Strong (1996) agrupam nessa categoria as dimensões: credibilidade, objetividade, acurácia e reputação. Zaveri *et al.* (2015) descreve como sendo as dimensões intrínsecas: validade sintática, precisão semântica, consistência, concisão e completude.

A qualidade, em uma perspectiva intrínseca, estaria relacionada então a acepção de qualidade como sinônimo de “livre de erros que exigem fazer o trabalho repetidamente (retrabalho) ou que resultam em falhas de campo, insatisfação do cliente, reclamações do cliente, e assim por diante” (Juran *et al*., 1998, p. 2.1).

A qualidade pode ser então discutida através de um conjunto de características inerentes a esses dados, com uma abordagem mais genérica, de uma maneira independente do domínio e do contexto de futura aplicação. Nessa perspectiva, um conjunto de dados pode ser considerado de qualidade quando está livre de erros estruturais e de más formações.

No contexto do *Linked Data*, a qualidade intrínseca está relacionada com a criação de bons URIs, a aplicação correta da estrutura do RDF e com a seleção e aplicação de vocabulários e propriedades. Nesse contexto as principais dimensões intrínsecas são: acurácia, confiabilidade, validade sintática, validade semântica, precisão semântica, consistência, concisão e completude (Zaveri, 2012; Farber *et al.*, 2016; Debattista *et al.*, 2018).

Com base na fragmentação das definições e na análise da literatura, realizou-se a sistematização apresentada na figura 24.

Figura 24 - Sistematização do termo "categoria intrínseca"

Fonte: Autora (2025)

Com base na fragmentação das definições e na sistematização apresentada na figura 24, elaborou-se a seguinte definição:

A categoria intrínseca permite a organização e o agrupamento de dimensões relacionadas às características inerentes dos dados, que podem ser mensuradas de maneira independente da tarefa a ser realizada ou das necessidades e objetivos do usuário, embora sua avaliação seja, em determinada medida, influenciada pelo domínio de criação dos dados. Geralmente a qualidade intrínseca é definida do ponto de vista do produtor. Nessa perspectiva, para que os dados sejam considerados de boa qualidade, eles precisam ser logicamente consistentes, sintaticamente corretos, coerentes, compactos e corretos, estarem livres de anomalias e representarem adequadamente a realidade a qual estão relacionados. No contexto do *Linked Data*, a qualidade intrínseca está relacionada com a criação de bons URIs, a aplicação correta da estrutura do RDF e com a seleção e aplicação de vocabulários e propriedades. Nesse contexto as principais dimensões intrínsecas são: acurácia, confiabilidade, validade sintática, validade semântica, precisão semântica, consistência, concisão, completude.

5.3.11 Qualidade representacional

A qualidade representacional, quando comparada as qualidades intrínseca e contextual, é significativamente menos abordada pela literatura.

Acosta *et al.* (2018, p. 4) define os aspectos representacionais da qualidade de dados como “[...] aqueles que capturam aspectos relacionados ao design dos dados”. Essa definição, que associa a qualidade representacional ao processo de design dos dados é mencionada ainda por Zaveri *et al.* (2015), Debattista *et al.* (2018), Abian *et al.* (2018) e Nayak, Bozic e Longo (2021).

Ahmed (2017, p. 3, tradução nossa) aponta ainda que “As dimensões representacionais estão relacionadas à concepção de dados”.

Outro aspecto mencionado pelas definições analisadas está relacionado ao formato utilizado para registrar os dados. Farber *et al.* (2017, p. 3-4, tradução nossa) indica que as dimensões representacionais da qualidade “[...] contém aspectos relacionados ao formato dos dados e [...] ao significado dos dados". Gurdur, El-khoury e Nyberg (2019) apontam que as dimensões representacionais têm como foco os aspectos relacionados ao formato com o qual os dados são codificados e armazenados de forma persistente.

Wang e Strong (1996) reforçam a relação entre a qualidade representacional e o formato dos dados, indicando que a qualidade da representação inclui a preocupação com a representação concisa e consistente dos dados e de seu significado, estando relacionado a aspectos como interpretabilidade e facilidade de compreensão.

Wang e Strong (1996, p. 20, tradução nossa) abordam ainda como ocorre a percepção de qualidade representacional pelo usuário dos dados: “para que os consumidores de dados concluam que os dados estão bem representados, eles devem ser não apenas concisos e consistentemente representados, mas também interpretáveis e fáceis de entender”.

No contexto do *Linked Data* a qualidade de dados representacional está relacionada com os formatos de serialização dos dados e com a presença de comentários e anotações que facilitem o seu entendimento.

Está relacionada ainda com a análise estrutural dos URIs, que não devem ser demasiadamente longos ou de difícil compreensão, com a análise dos literais, que devem seguir os padrões de escrita e forma estabelecidos pelo domínio. Está

relacionado ainda com a interoperabilidade dos dados, fortemente impactada pela aplicação dos vocabulários. Nesse domínio, as principais dimensões são: concisão representacional, interoperabilidade, interpretabilidade, versatilidade e facilidade de compreensão (Zaveri, 2012; Farber *et al.,* 2016; Debattista *et al.*, 2018).

Com base na análise das citações, é possível inferir que, no contexto da qualidade de dados, os aspectos representacionais não estão relacionados às descrições dos conjuntos de dados, como o termo poderia ser conceituado no âmbito da Ciência da Informação. Nesse domínio, a representação está relaciona ao registro do dado enquanto unidade de conteúdo abstrata em determinado suporte, que permite o seu armazenamento, uso e compartilhamento.

Com base na análise dos modelos de qualidade de dados *Linked Data* e fragmentação das definições, realizou-se a sistematização apresentada na figura 25.

Figura 25 - Sistematização do termo “qualidade representacional”

Fonte: Autora (2025)

Com base na análise dos modelos de qualidade de dados *Linked Data*, na fragmentação das definições e na sistematização apresentada na figura 25, elaborou-se a seguinte definição:

A categoria representacional agrupa e organiza dimensões relacionadas com a qualidade da estrutura responsável pela materialização dos dados, que possibilita o seu registro, a sua codificação e armazenamento, partindo do entendimento de dados como “unidades de registro abstratas que necessitam de um suporte”. Nessa perspectiva, para serem considerados de boa qualidade, precisam ter clareza estrutural, estarem livres de redundâncias, coerentes em relação a sua sintaxe e semântica, e serem fáceis de interpretar para usuários humanos e para agentes computacionais. No contexto do *Linked Data* a qualidade de dados representacional está relacionada com os formatos de serialização dos dados, com a presença de comentários e anotações que facilitem o seu entendimento, com a análise estrutural dos URIs, que não devem ser demasiadamente longos ou de difícil compreensão, e com a análise dos literais, que devem seguir os padrões de escrita e forma estabelecidos pelo domínio. Está relacionado ainda com a interoperabilidade dos dados, fortemente impactada pela aplicação dos vocabulários. As principais dimensões de qualidade representacional são: concisão representacional, interoperabilidade, interpretabilidade, versatilidade, facilidade de compreensão.

Apresentada a definição do termo Categoria Contextual, a próxima subseção apresenta as discussões e a definição do termo Categoria De Acessibilidade

5.3.12 Categoria acessibilidade

A categoria acessibilidade busca avaliar “[...] até que ponto os dados estão disponíveis e podem ser recuperados (Ahmed, 2017, p. 3)”. Nessa categoria também busca-se avaliar o grau de “funcionamento adequado dos métodos de acesso para que [os dados] sejam obtidos por um consumidor (Mihindukulasooriya; Garcia-Castro; Gomez-Perez, 2017, p. 150, tradução nossa).

Ao se referirem aos consumidores de dados, Debattista *et al.* (2018, p. 888, tradução nossa) acrescentam ainda que “As dimensões na categoria acessibilidade abordam a facilidade com que máquinas e humanos podem (re)utilizar recursos Linked Data.”

Dessa definição, destaca-se a importância de garantir a reutilização dos dados *Linked Data,* bem como de seu potencial de utilização tanto por usuários humanos como por agentes computacionais.

Além da capacidade de acesso e recuperação dos dados, diversos autores acrescentam ainda um terceiro aspecto: a autenticidade dos dados. (Zaveri *et al.*, 2015; Acosta *et al.*, 2018; Nayak; Bozic; Longo, 2021). Färber *et al.* (2016) inclui na categoria acessibilidade a disponibilização de metadados que permitam a recuperação do conjunto de dados.

A perspectiva de acessibilidade está relacionada, portanto, à capacidade dos usuários de obterem acesso aos dados de maneira eficaz e eficiente. Nessa perspectiva, um dado é considerado de qualidade quando pode ser localizado e obtido sempre que necessário.

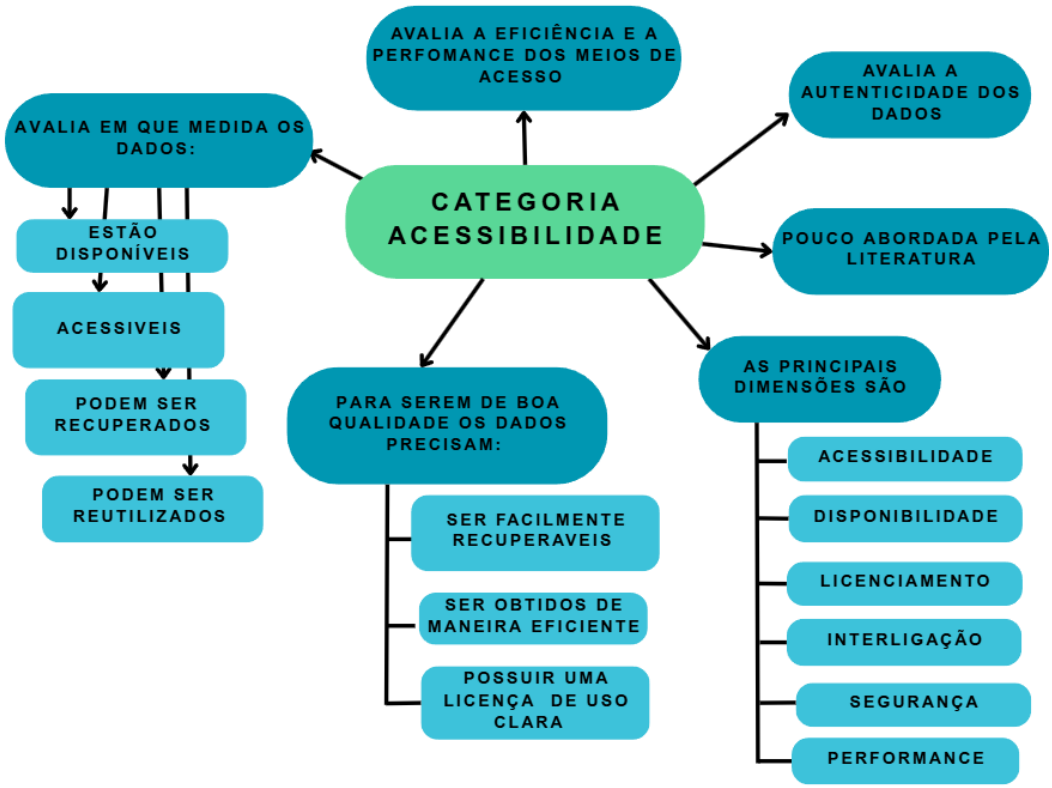
Como no *Linked Data* uma forma de acesso aos dados é por meio das ligações entre diversas fontes, a categoria acessibilidade se refere ainda a identificar [...] “em que medida os dados estão suficientemente interligados a outros recursos.” (Abian *et al.*, 2018, p. 145, tradução nossa).

Para os dados publicados como *Linked Data*, a qualidade de acessibilidade está relacionada com a disponibilidade e o funcionamento dos *endpoints* SPARQL, com a disponibilidade de *download* do conjunto de dados em diferentes formatos de serialização e com a presença de licença legível por humanos e máquinas.

As principais dimensões relacionadas a acessibilidade são: disponibilidade, licenciamento, interligação, segurança, performance e acessibilidade.

Com base na análise dos modelos de qualidade de dados *Linked Data* e na fragmentação das definições, realizou-se a sistematização apresentada na figura 26.

Figura 26 - Tematização do termo “categoria acessibilidade”



Fonte: Autora (2025)

Com base na fragmentação das definições e na sistematização apresentada na figura 26, elaborou-se a seguinte definição:

A categoria acessibilidade permite a organização e o agrupamento de dimensões que buscam avaliar em que medida os dados podem ser recuperados, estão disponíveis e acessíveis, permitindo seu uso e reuso. Também reúne dimensões que buscam avaliar a performance e a eficiência dos meios de acesso e a autenticidade dos dados. Em uma perspectiva de acessibilidade são considerados de boa qualidade dados que podem ser facilmente recuperados, obtidos e que possuem uma licença de uso clara e disponível tanto para usuários humanos como para agentes computacionais. Nesse contexto do *Linked Data*, a qualidade de acessibilidade está relacionada com a disponibilidade e o funcionamento dos *Endpoints* SPARQL, com a disponibilidade de *download* do conjunto de dados em diferentes formatos de serialização, com a presença de licença legível por humanos e máquinas. Nesse contexto, também busca verificar em que medida os dados estão conectados com fontes externas relevantes. As principais dimensões de acessibilidade são: disponibilidade, licenciamento, interligação, segurança, performance e acessibilidade.

Apresentada a definição do termo Categoria De Acessibilidade, a próxima subseção apresenta as discussões e a definição do termo Dimensão.

5.3.13 Dimensão

A análise da literatura permite observar que as dimensões de qualidade fazem parte da organização hierárquica da avaliação de qualidade. Os autores partem do entendimento de que as dimensões podem ser agrupadas em categorias de qualidade, sendo, portanto, o segundo nível na hierarquia do processo de avaliação de qualidade de dados (Farber *et al.*, 2017; Melo, 2017; Arruda *et al.*, 2019; Candela *et al.*, 2020).

A dimensão pode ser definida como “[...] um aspecto principal de como a qualidade dos dados pode ser visualizada (Farber *et al.*, 2017, p.3, tradução nossa).

Zhang; Benis; Cornet (2023, p. tradução nossa) definem dimensões como as características de um conjunto de dados “necessárias para atingir determinados objetivos em um aspecto de qualidade específico”. Nesse sentido, os objetivos a serem atingido podem variar, mas as dimensões seguem sendo relacionadas ao agrupamento intencional de característica semelhantes.

Melo (2017, p. 48) apresenta a seguinte definição para o termo dimensão: “Os problemas de qualidade são classificados em diferentes categorias, chamadas de dimensões, nas quais problemas do mesmo tipo são classificados de acordo com uma categoria específica”.

Nessa definição, destaca-se a preocupação em agrupar aspectos de qualidade com base em sua semelhança. Entretanto, definir os aspectos de qualidade como problemas de qualidade torna-se limitador, considerando que os dados podem ou não apresentar problemas em relação a determinados aspectos de qualidade.

Uma dimensão de qualidade pode ser entendida então como uma categoria que reúne "[...] um conjunto de **atributos de qualidade** de dados que **representa um único aspecto ou construto** da qualidade de dados" (Wang; Strong, 1996, p. 6, tradução nossa, grifo nosso). Nessa perspectiva, "Cada dimensão captura um aspecto específico incluído no âmbito geral da qualidade dos dados" (Batini; Scannapieco, 2016, p. 21, tradução nossa).

Os autores discutem ainda a capacidade de organização e sistematização das dimensões. Batini e Scannapieco (2016, p. 98, tradução nossa) apontam que:

> As dimensões [...] podem ser caracterizadas como uma estrutura de classificação comum que nos permite comparar dimensões entre diferentes tipos de informação. A estrutura baseia-se em uma classificação em grupos de dimensões [...], onde as dimensões são incluídas no mesmo grupo de acordo com sua similaridade.

As dimensões têm, portanto, o objetivo de organizar sistematicamente a análise da qualidade dos dados (W3C, 2017). Visando a sua organização e sistematização, as dimensões são agrupadas em categorias, como discutido anteriormente.

As dimensões precisam poder ser mensuráveis como aponta o DAMA INTERNATIONAL (2015, p. 454):

> Uma dimensão de Qualidade de Dados é um recurso ou característica mensurável dos dados. O termo dimensão é usado para fazer a conexão com dimensões na medição de objetos físicos (por exemplo, comprimento, largura, altura). As dimensões de qualidade de dados fornecem um vocabulário para definir os requisitos de qualidade de dados. A partir daí, elas podem ser usadas para definir os resultados da avaliação inicial da qualidade de dados, bem como da medição contínua. Para mensurar a qualidade dos dados, uma organização precisa estabelecer características que sejam importantes para os processos de negócios (valem a pena mensurar) e mensuráveis. As dimensões fornecem uma base para regras mensuráveis, que devem estar diretamente conectadas a riscos potenciais em processos críticos.

Para além da necessidade de poderem ser mensuradas, essa definição também ressalta outros aspectos relevantes para a definição do termo, como a sua

capacidade para fornecer vocabulário e estabelecer os requisitos de qualidade de dados em determinado contexto. A definição aborda ainda um aspecto importante da qualidade de dados: a de que ela pode ser um processo pontual ou contínuo.

Embora seja necessário poder mensurar as diferentes dimensões, elas são abstrações dos aspectos de qualidade de dados existentes. Por esse aspecto abstrato, as dimensões necessitam de critérios e métricas para viabilizar a sua avaliação.

A inclusão dos critérios não é um consenso na literatura, embora tenham sido incluídos na estruturação original de Wang e Strong. Alguns autores, inclusive o W3C (2017), apenas abordam a necessidade de métricas para avaliar as dimensões de qualidade. Para essa pesquisa, considera-se que os critérios são importantes pois "Uma dimensão de qualidade de dados compreende um ou vários critérios de qualidade de dados" (Farber *et al.*, 2017, p. 3, tradução nossa).

Os critérios são aspectos concretos da qualidade, passíveis de serem mensurados, e uma dimensão pode possuir mais de um critério. Da mesma forma, "Cada dimensão de qualidade deve ter uma ou mais métricas para medi-la" (W3C, 2017).

Com base na análise das citações, observa-se, portanto, que as dimensões de qualidade podem ser divididas em critérios que permitem uma maior especificidade dos aspectos a serem avaliados, mas devem necessariamente possuir uma ou mais métricas que permitam mensurar a qualidade dos dados em relação ao aspecto específico da qualidade por ela representado.

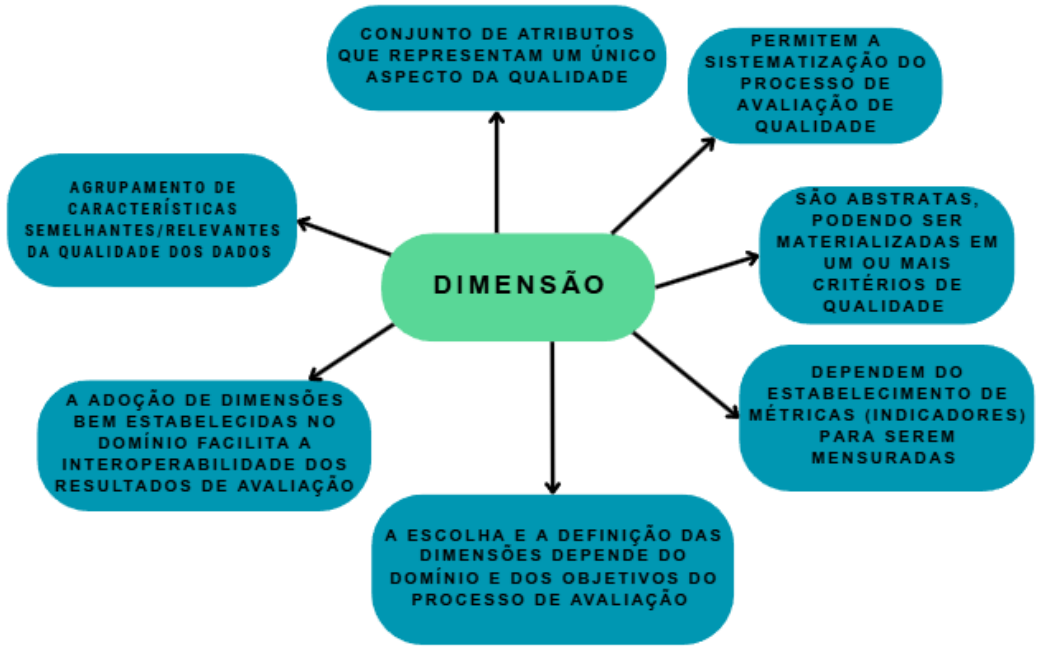
Melo (2017) aponta que existe um conjunto relativamente bem estabelecido de dimensões disponíveis para avaliar a qualidade dos dados. Essas dimensões possuem "[...] definições similares, porém a aplicação e como funciona cada dimensão possuem características distintas, de acordo com o domínio no qual estão sendo aplicadas" (Melo, 2017, p. 48).

Isso ocorre porque, mesmo em uma perspectiva intrínseca, os dados de distintos domínios possuem caraterísticas únicas. Quando falamos de *Linked Data*, por exemplo, é necessário considerar a adequação aos princípios, a aplicação do modelo RDF e das propriedades dos vocabulários. "A definição das dimensões de qualidade de dados *Linked Data* apresenta uma série de desafios únicos" (Batini; Scannapieco, 2016, p. 98, tradução nossa).

As dimensões já estabelecidas podem ser adaptadas para atenderem a distintos domínios, podendo ser desenvolvidos novos critérios e métricas. Entretanto, Albertoni, Martino e Quarati (2021) ressaltam que reutilizar classificações e métricas já estabelecidas no domínio aumenta a interoperabilidade dos resultados.

Com base nas discussões e na fragmentação das definições, elaborou-se uma sistematização, apresentada na figura 27.

Figura 27 - Sistematização do termo “dimensão”



Fonte: Autora (2025)

Com base na sistematização das definições na Grade e nos aspectos apresentados na figura 27 elaborou-se a seguinte definição:

> Uma dimensão de qualidade é uma categoria que reúne características semelhantes dos dados, relevantes para a avaliação da qualidade. Consiste em um conjunto de atributos abstratos que representam um aspecto único no âmbito geral da qualidade de dados, fornecendo assim vocabulário para estabelecer os requisitos de qualidade esperados para um conjunto de dados. Por seu caráter abstrato, podem ser materializadas em um ou mais critérios mensuráveis. Para avaliar a qualidade dos dados em relação a uma dimensão é necessário o estabelecimento de uma ou mais métricas, indicadores que permitem a avaliação quantitativa e qualitativa da qualidade em relação a determinada dimensão. A escolha e a definição das dimensões e dos critérios e métricas que a compõe

> depende do domínio no qual estão inseridos os dados e dos objetivos do processo de avaliação de qualidade. Existe um conjunto de dimensões estabelecido na comunidade de avaliação de qualidade, que podem ser adaptadas para atender a necessidades específicas. Também podem ser criadas dimensões para atender aos propósitos do processo de avaliação, entretanto, a adoção de dimensões bem estabelecidas facilita o uso de ferramentas automáticas e a interoperabilidade dos resultados do processo de avaliação.

Apresentada a definição do termo Categoria Dimensão, a próxima subseção apresenta as discussões e a definição do termo Critério.

5.3.14 Critério

Na elaboração do *framework* do processo de avaliação de qualidade, Wang e Strong (1996) estabeleceram uma série de atributos de qualidade que foram utilizados pelos usuários de dados para categorizar aspectos semelhantes de qualidade em dimensões. Os autores apontam que:

> Referimo-nos às características da qualidade dos dados como atributos ou mensurações da qualidade dos dados para distingui-los da dimensão da qualidade dos dados que resulta da análise fatorial ao longo desta seção.

Esses atributos ou mensurações, consistem em diferentes aspectos concretos de qualidade, mas não necessariamente em indicadores e fórmulas para avaliação da qualidade.

Uma parte significativa dos modelos de qualidade apresenta esse último nível de detalhamento da qualidade antes da apresentação das métricas que permitem a sua avaliação, e se referem a esse aspecto mais específico da qualidade como critérios de qualidade, como os modelos propostos por Rula (2011), Candela, *et al.* (2017), Farber *et al.* (2017) e Candela *et al.* (2020).

Alguns modelos de qualidade, como o de Zaveri *et al.*, apresentam um subnível de dimensões, mas não fazem diferenciação da nomenclatura entre dimensões e subdimensões, referindo-se a ambas como dimensões.

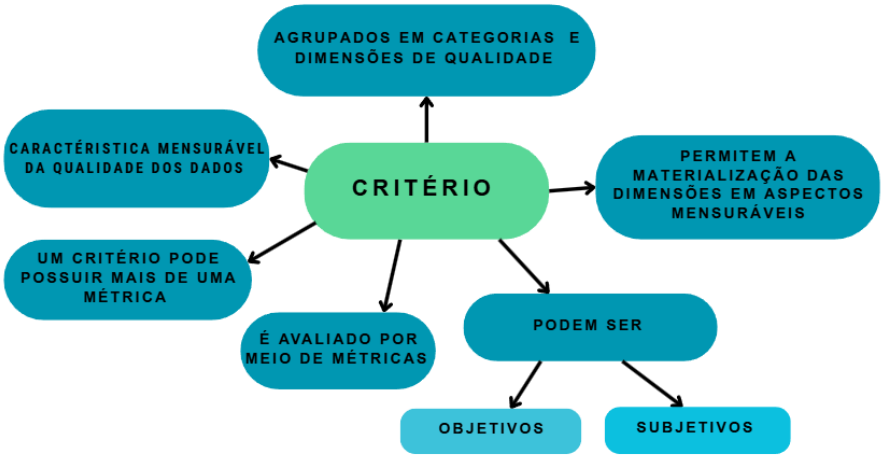
Candela, *et al.* (2017, p. 3, tradução nossa) definem os critérios de qualidade como sendo “uma característica articulada dos dados em relação à sua qualidade e pode ser subjetiva ou objetiva”.

Com base na sistematização do processo de avaliação de qualidade, entende-se que os critérios são importantes pois especificam o aspecto de cada dimensão que está sendo mensurado.

Como categorias e dimensões são abstratas, os critérios permitem a sua materialização, sendo aspectos mensuráveis de qualidade. Uma dimensão pode ser composta por mais de um critério e um critério pode ser mensurado por meio de distintas métricas.

Considerando a limitação em relação a definições para o termo critério de qualidade, elaborou-se a sua definição com base na análise dos modelos de qualidade de dados identificados e na sistematização do processo de avaliação de qualidade. Com base nas discussões e nessa análise, elaborou-se a sistematização apresentada na figura 28.

Figura 28 - Sistematização do termo “critério”



Fonte: Autora (2025)

Com base na análise dos modelos de qualidade e na sistematização apresentada na figura 28, elaborou-se a seguinte definição:

> Um critério de qualidade representa uma característica mensurável dos dados. Os critérios de qualidade são agrupados, com base em sua semelhança, em categorias e dimensões de qualidade. Podem possuir um caráter subjetivo ou objetivo, o que irá impactar na forma como serão avaliados. Para que a qualidade dos dados em relação a um critério possa ser avaliada são necessárias métricas de qualidade, podendo um critério ser avaliado com base em mais de uma métrica.

Apresentada a definição do termo Critério, a próxima subseção apresenta as discussões e a definição do termo Métrica.

5.3.15 Métrica

O termo métrica representa o nível mais granular de especificidade na hierarquização do processo de avaliação de qualidade.

Ao analisar as definições fragmentadas com base na Grade, observa-se que a maior parte dos autores definem métrica como um procedimento (Behkamal *et al.*, 2014; Zaveri *et al.*, 2015; Radulovic *et al.*, 2017; Nooghabi; Dastgerdi, 2016).

As métricas são estabelecidas em função dos indicadores de qualidade, e permitem calcular uma pontuação com base em uma função. Podem ser definidas como medidas de qualidade ou como sinônimo de função ou indicador (Zaveri *et al.*, 2015; Gurdur; El-khoury; Nyberg, 2019; Zhang; Benis; Cornet, 2023).

Com base na análise dos modelos de qualidade de dados, observa-se que a definição de métrica como um indicador pode causar ambiguidades, considerando que a maioria dos modelos adota uma estrutura que apresenta indicadores e métricas como partes distintas do modelo. Observa-se ainda uma limitação em relação a sua definição como função, uma vez que boa parte dos modelos adota também escalas como forma de mensurar a qualidade.

Como um procedimento, as métricas têm o objetivo de:

> [...] medir o grau em que um determinado critério de qualidade de dados é atendido para um determinado conjunto de dados, cada critério é formalizado e expresso em termos de uma função com o intervalo de valores. Chamamos essa função de métrica de qualidade de dados do respectivo critério de qualidade de dados (Radulovic *et al.*, 2017, p. 3).

Quanto a sua organização

> Essas métricas podem ser classificadas em duas categorias: métricas quantitativas (QN) e métricas qualitativas (QL). Métricas são medidas quantitativamente (QN), quando são quantificadas ou para as quais um valor concreto (pontuação) pode ser calculado. Métricas são medidas qualitativamente (QL), quando não podem ser quantificadas e dependem da percepção dos usuários (Almeida *et al.*, 2016, p. 3, tradução nossa).

Além de quantitativas e qualitativas as métricas também podem ser organizadas em subjetivas e objetivas:

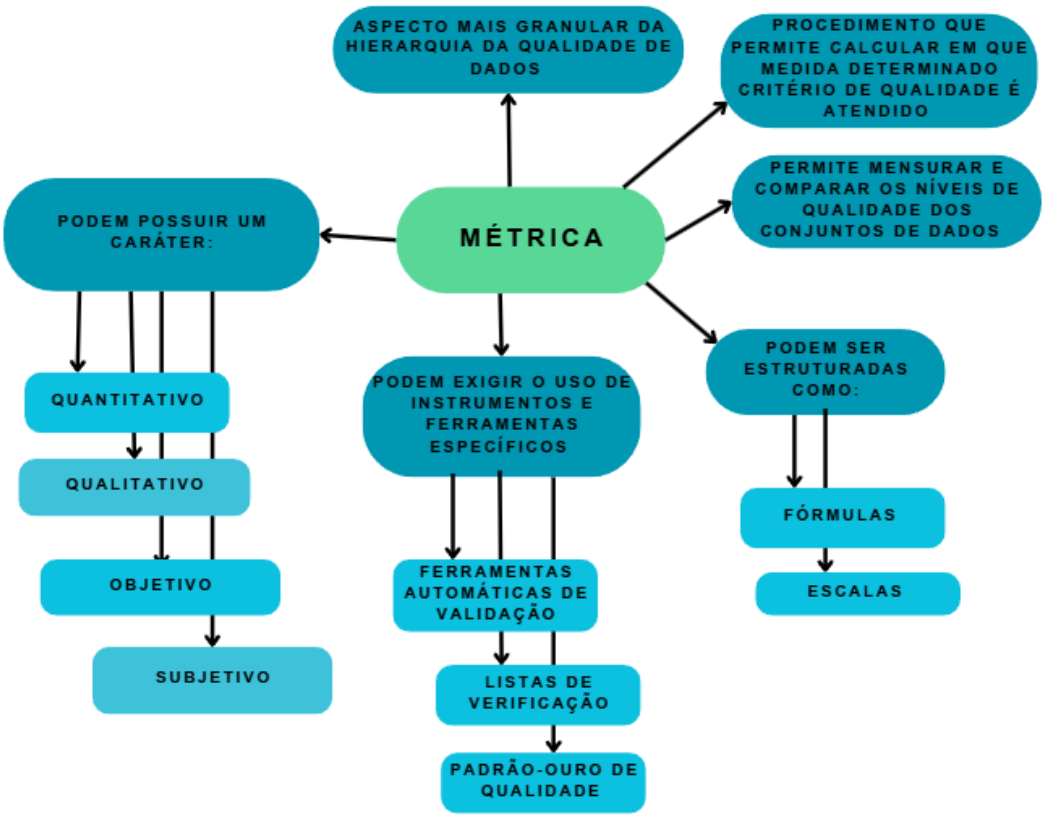
> Métricas podem ser subjetivas e, portanto, depender do julgamento humano (por exemplo, reputação); métricas podem ser objetivas, mas dependentes do contexto (por exemplo, completude); métricas podem ser fáceis de implementar (por exemplo, hora da última modificação); ou mais demoradas para avaliar (por exemplo, anotação precisa).

Dessa citação, observa-se que as métricas também variam em relação ao nível de complexidade observado em sua avaliação. Nesse sentido, a avaliação de determinadas métricas pode requerer o uso de ferramentas adicionais, como validadores de estrutura RDF ou validadores gramaticais de literais, por exemplo. Pode exigir ainda a criação de instrumentos para amparar a sua avaliação “Métricas de qualidade de dados precisam ser calculadas e frequentemente requerem fontes de dados adicionais para isso” (Homburg, 2020, p. 5, tradução nossa).

Zaveri *et al.* (2015) acrescentam ainda que pode existir mais de uma métrica associada a uma mesma dimensão e critérios, mas que uma métrica só pode ser associada a um único critério.

Arruda *et al.* (2019) apontam que as métricas geralmente precisam ser criadas ou adaptadas para cada domínio de aplicação. Com base na fragmentação das definições e na análise da literatura, realizou-se a sistematização apresentada na figura 29.

Figura 29 - Sistematização do termo “Métrica”



Fonte: Autora (2025)

Com base na fragmentação das definições e na sistematização apresentada na figura 29, elaborou-se a seguinte definição:

> Métricas de qualidade são procedimentos que permite calcular em que medida determinada critério de qualidade é atendido e os níveis de qualidade dos dados em relação a uma dimensão. Uma métrica é considerada o aspecto mais granular na hierarquia da qualidade de dados, permitindo mensurar e comparar os níveis de qualidade de um ou mais conjuntos de dados em relação a um aspecto específico. As métricas podem possuir um caráter quantitativo ou qualitativo, subjetivo ou objetivo. Podem ser estruturadas como escalas ou como fórmulas. Variam em relação a complexidade do processo de avaliação e podem exigir o uso de instrumentos e ferramentas específicos como validadores de estrutura e conteúdo, listas de verificação e padrões-ouro de qualidade.

Apresentada estrutura hierárquica da qualidade de dados, os termos e as definições necessárias para a compreensão da avaliação de qualidade, a próxima seção apresenta a análise processual da qualidade de dados *Linked Data.*

6 ANÁLISE PROCESSUAL DA QUALIDADE DE DADOS *LINKED DATA*

Essa seção foi construída com o objetivo de apresentar os resultados da discussão processual da qualidade de dados *Linked Data*, abordando processos, atividades, instrumentos e ferramentas que permeiam a avaliação de qualidade, com foco na sua aplicação na seleção de fontes de dados *Linked Data* para a criação de *links* entre fontes distintas.

A análise processual foi dívida em duas etapas principais: 1) análise de ciclo de vida dos dados e processos e 2) análise dos instrumentos e ferramentas de avaliação de qualidade.

A primeira parte da seção é composta pelas subseções 6.1 a 6.3, elaboradas com o objetivo de apresentar o contexto, as etapas e atividades de diferentes processos direta e indiretamente conectados com a seleção de fontes de dados *Linked Data*, visando mapear o fluxo do processo de seleção de fontes.

Partiu-se de uma abordagem ampla do contexto dos dados, inserindo o consumo de dados e a avaliação de qualidade no ciclo de vida dos dados, apresentado na subseção 6.1. Buscou-se então discutir as etapas da interligação no *Linked Data*, abordando a necessidade de seleção de fontes e avaliação de qualidade nesse processo, apresentado na subseção 6.2. Por fim, a subseção 6.3 apresenta os processos relacionados com a qualidade, sendo eles: planejamento de qualidade, controle de qualidade, avaliação de qualidade e melhoria de qualidade. Nessa subseção foram discutidas as etapas e as atividades que compõe esses processos, abordando como elas se relacionam ou podem ser adaptadas para serem aplicadas à seleção de fontes de dados.

A segunda parte da seção é composta pelas subseções 6.4 a 6.6 e foi construída com o propósito de apresentar os instrumentos, metodologias e ferramentas criados para apoiar a avaliação de qualidade de dados *Linked Data* e que podem auxiliar no processo de seleção de fontes.

A seção 6.4 apresenta os modelos de qualidade para dados *Linked Data*, analisando a sua estrutura e as principais particularidade de cada modelo. A seção 6.5 discute as possíveis abordagens para a criação e aplicação de modelos de qualidade. A seção 6.5 discute o papel dos vocabulários e ontologias na avaliação de qualidade, tendo a sua condução guiada pelo uso desses instrumentos nos modelos de qualidade.

6.1 A qualidade e a seleção de fontes no ciclo de vida de dados *Linked Data*

A condução da análise processual da qualidade de dados *Linked Data* foi realizada, nessa pesquisa, com o objetivo de identificar procedimentos, etapas e tarefas relacionadas com a seleção de dados para criação de *links* com fontes externas.

A presente subseção busca identificar nos ciclos de vida dos dados (CVD), etapas, atividades é tarefas relacionadas com a qualidade de dados e com o processo de seleção de fontes de dados, considerando que:

> [...] entender o ciclo de vida permite uma melhor compreensão sobre a natureza dos dados, além de fornecer um vocabulário compartilhado que permite a diferentes profissionais discutirem questões essenciais relacionadas à publicação e consumo de dados na *Web*. Além disso, um ciclo de vida de dados ajuda a explicar mudanças de paradigma, a comparar a funcionalidade de diferentes plataformas e a auxiliar na integração de esforços de implementação previamente díspares (Santos *et al.*, 2018, p. 2, tradução nossa).

Para essa análise foram selecionados modelos de CVD gerais e focados em dados *Linked Data*, nos quais puderam ser identificadas etapas e tarefas relacionadas com o consumo de dados, seleção de fontes e avalição de qualidade. O quadro 18 apresenta os CVDs analisados e os motivos para sua inclusão na discussão.

Quadro 18 - Modelos de ciclo de vida analisados

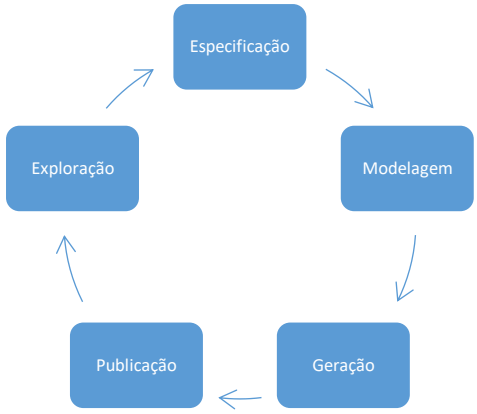
Modelo	Autor	Justificativa
Ciclo de vida da publicação de dados governamentais como *Linked Data*	Villazón-Terrazas *et al.* (2011)	Modelo focado em dados *Linked Data*. Apresenta, na etapa de publicação, a tarefa de criação de *links*.
Ciclo de vida de dados *Linked Data*	Auer *et al.* (2012)	Modelo de ciclo de vida para dados *Linked Data*. Incluí uma etapa chamada "*interlinkig*" que aborda a criação de *links.* Inclui também uma etapa de avaliação de qualidade.
Ciclo de vida de dados na *Web*	Lóscio, Oliveira e Bittencourt (2015) (2015)	Modelo baseado nas Melhores Práticas para a Publicação de Sados na Web (W3C, 2017), abordando aspectos relacionados ao *Linked Data*. Inclui a etapa de consumo, que pode ser relacionada com a seleção de fontes.

Ciclo de Vida dos Dados para Ciência da Informação	Sant'Ana (2016)	Modelo abrangente, pautado nas contribuições da Ciência da Informação e que tem como um dos pilares a qualidade de dados. Apresenta como uma das etapas a coleta, com a qual é possível traçar aproximações com os processos e seleção de fontes.
Ciclo de vida da qualidade de dados *Linked Data*	Mihindukulasooriya (2020)	Apresenta um ciclo de vida de dados *Linked Data* focado na qualidade de dados.

Fonte: Autora (2025)

Villazón-Terrazas *et al.* (2011) apresentam um modelo de ciclo de vida focado na publicação de dados governamentais como *Linked Data*. Embora seja focado em dados governamentais, o ciclo apresentado é bastante abrangente e as etapas podem ser facilmente adaptadas a outros domínios. O ciclo de vida proposto pode ser visualizado na figura 30.

Figura 30 - Ciclo de vida da publicação de dados governamentais como *Linked Data*



Fonte: Traduzido de Villazón-Terrazas *et al.* (2011)

O modelo não possui uma etapa de coleta ou consumo de dados, mas inclui a criação de *links* com fontes externas como uma tarefa da etapa de publicação, relacionada a atividade de ligação com fonte externas.

Ao abordar a criação de *links* os autores apontam que: "Essa tarefa envolve a descoberta de relacionamentos entre itens de dados. Podemos criar esses *links* manualmente, o que é uma tarefa demorada, ou podemos contar com ferramentas automáticas ou supervisionadas" (Villazón-Terrazas *et al.,* 2011, p. 36, tradução nossa).

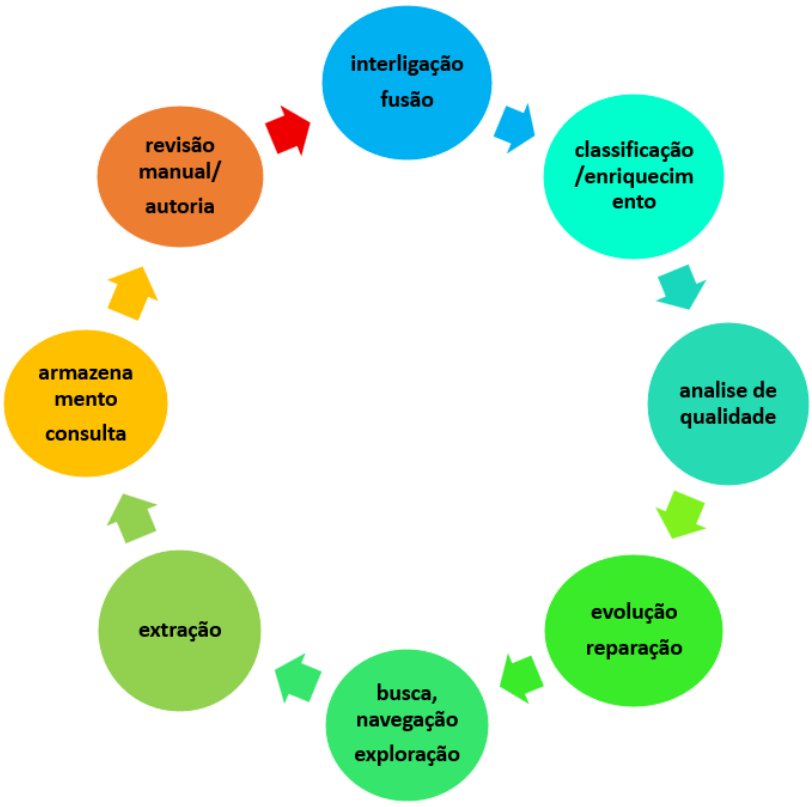
No modelo a tarefa de ligação com fontes externas é dividida 3 atividades principais, sendo elas:

- **Identificação de conjuntos de dados** – consiste na busca por conjuntos com temas semelhantes, que pode ser realizado em catálogos de dados;
- **Identificação de relacionamentos** – busca pelos relacionamentos que possam existir entre os dados do conjunto de dados e os identificados na etapa anterior, que pode ser realizada utilizando ferramentas automáticas;
- **Validação dos relacionamentos** – consiste na validação dos relacionamentos identificados, os autores mencionam que essa etapa geralmente é realizada por especialistas do domínio ou pela aplicação de validadores de *links*.

Nesse ciclo de vida, a busca por dados baseia-se em uma abordagem exploratória, que parte de fontes reconhecidas e relevantes ou de catálogos de dados *Linked Data* para identificar dados com potencial de ligação, levantando, inclusive, a possibilidade de uso de ferramentas semiautomáticas para a identificação de *links* que depois podem passar por um processo de validação.

Auer *et al.* (2012) apresentam um ciclo de vida organizado em 8 etapas, apresentado na figura 31:

Figura 31 - CVD proposto por Auer *et al.* (2012)



Fonte: (traduzido de Auer *et al.*, 2012)

Esse CVD considera aspectos relacionados à publicação e ao consumo de dados *Linked Data*, além de inclui a avaliação de qualidade como uma das etapas do processo, sendo apresentada em relação direta com a identificação/correção de problemas de qualidade.

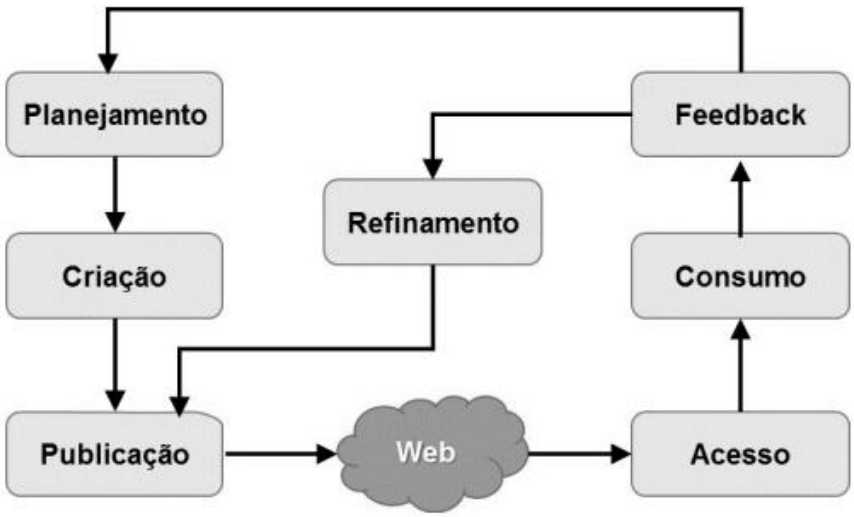
Em relação a etapa de interligação os autores apontam que:

> Criar e manter *links* de forma (semi)automatizada ainda é um grande desafio, crucial para estabelecer coerência e facilitar a integração de dados. Buscamos abordagens de linkagem que gerem alta precisão e *recall*, que se configurem automaticamente ou com o *feedback* do usuário final (Auer *et al.*, 2012, p. 3, tradução nossa).

Para a criação de *links* semiautomáticos, torna-se necessária a identificação previa de fontes em potencial. Nesse modelo, a seleção de fontes pode ser inserida como uma tarefa do processo de interligação e fusão, necessária como uma etapa para descoberta de fontes para esse processo, a etapa de identificação das fontes não é aprofundada nesse modelo.

Lóscio, Oliveira e Bittencourt (2015) apresentam um CVD voltado para a publicação e o consumo de dados na *Web*. O modelo intitulado “Ciclo de vida de dados na Web” é apresentado na Figura 32.

Figura 32 - Ciclo de vida de dados na *Web*



Fonte: Lóscio, Oliveira e Bittencourt (2015, p. 51)

Nesse CVD as etapas de planejamento, criação, publicação e refinamento são atividades que tem como agente principal o publicador de dados. O consumidor começa a figurar a partir do acesso, entretanto é na etapa consumo onde pode ser inserida a atividade de seleção de fontes. Ao discutirem o modelo os autores apontam que o consumo:

> Implica o momento em que os dados são usados para a criação de visualizações, como gráficos e mapas de calor, bem como para aplicações que permitem o cruzamento e a realização de análises sobre os dados. Esta etapa do ciclo de vida está diretamente relacionada ao consumidor de dados, que pode ser desde uma grande empresa interessada em usar os dados disponíveis na *Web* para a melhoria de seus produtos e serviços, até um único desenvolvedor interessado em usar os dados para criar uma aplicação que irá melhorar a qualidade de vida na sua cidade (Lóscio; Oliveira; Bittencourt, 2015, p. 52).

Nesse modelo de ciclo de vida as etapas são discutidas com base em sua relação com os principais desafios para publicação e consumos de dados na *Web*, mencionados pelo W3C (2017).

Em relação a etapa de consumo, os principais desafios mencionados são:

- A necessidade de metadados;
- Necessidade de informações a respeito da proveniência dos dados;
- Necessidade de informações a respeito versionamento;
- Aspectos relacionados com a qualidade dos dados,
- Questões relacionadas com a identificação baseada em URIs;
- Questões relacionadas aos mecanismos de acesso;
- Questões relacionadas aos formatos dos dados;
- Questões relacionadas aos vocabulários; e
- Questões relacionadas a possibilidade de fornecimento de *feedback*.

Como pode ser observado, destaca-se novamente a importância dos metadados, especialmente dos aspectos relacionados à proveniência dos dados. Destacam-se ainda aspectos relacionado ao acesso aos dados, a aderência ao formato e ao uso correto de propriedades e vocabulários.

O modelo apresentado não posiciona os aspectos de qualidade em relação a nenhuma etapa específica do modelo, entretanto, os autores ressaltam a relevância da qualidade e os desafios de se estabelecer um modelo que atenda às necessidades

heterogenias dos dados publicados na *web*. Ressaltam ainda a importância da representação e disponibilização de informação a respeito da qualidade:

> [...] para a avaliação dos dados publicados na *Web*, é preciso levar em consideração a diversidade dos modelos e formatos de dados usados para a representação dos dados, bem como a criação de novos critérios ou dimensões de qualidade. Além disso, as informações de qualidade também precisam ser armazenadas de forma estruturada para permitir o monitoramento e verificação automática por máquinas. Apesar das diversas propostas para descrição de dimensões e critérios de qualidade, não há solução amplamente aceita pela comunidade e, principalmente, não há consenso sobre quais dimensões e métricas são fundamentais para avaliação de dados e conjuntos de dados publicados na *Web*. (Lóscio; Oliveira; Bittencourt, 2015, p. 57).

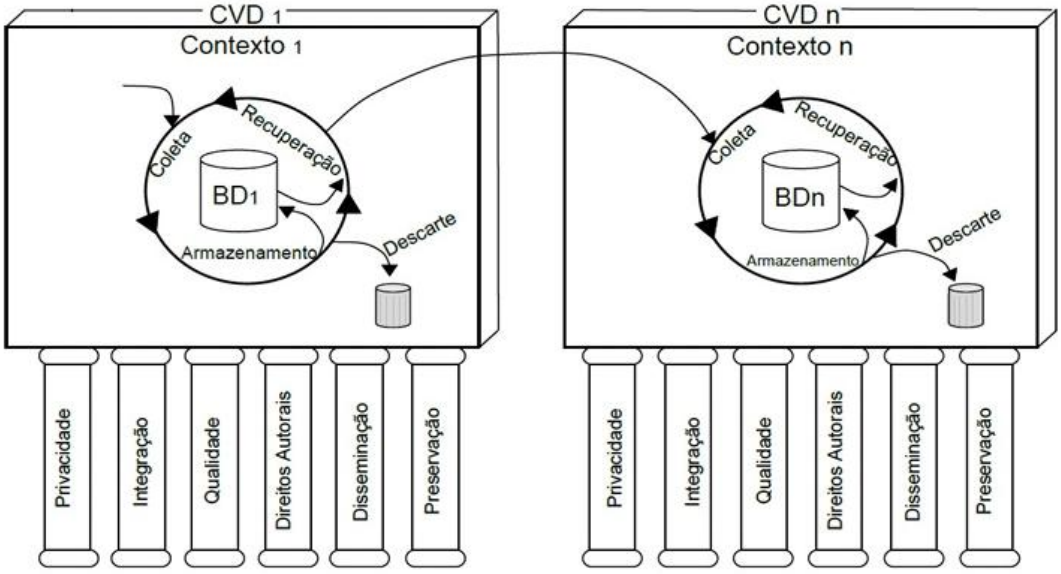
Os autores também ressaltam a importância dos vocabulários para o consumo de dados, especialmente no contexto do *Linked Data*, já que:

> Vocabulários são usados para especificar os termos comuns de domínios específicos, bem como os relacionamentos entre esses termos e possíveis restrições que podem ser aplicadas aos termos. A partir dessa especificação, é possível estabelecer um canal de comunicação entre provedores e consumidores de dados, no qual ambos compartilham a mesma visão da realidade (Lóscio; Oliveira; Bittencourt 2015, p. 59).

Sant'Ana (2016, p. 116) apresenta uma proposta de ciclo de vida que tem "[...] como elemento central, os próprios dados, amparando-se nos conceitos e contribuições que a Ciência da Informação pode proporcionar, sem abrir mão da reflexão sobre o papel de outras áreas chave como a Ciência da Computação".

Embora não seja focado em dados *Linked Data*, o modelo apresenta uma versão aprofundada das etapas que compõe o CVD, discutindo as competências necessárias para a condução de cada fase e identificando os pilares que as permeiam. A figura 33 apresenta o modelo, intitulado Ciclo de Vida dos Dados para Ciência da Informação (CVD–CI):

Figura 33 - Ciclo de Vida dos Dados para Ciência da Informação



Fonte: Sant'Ana (2016, p. 123)

Como é possível observar o CVD apresentado na figura 33 é composto por 4 etapas principais: coleta, recuperação, armazenamento e o descarte dos dados.

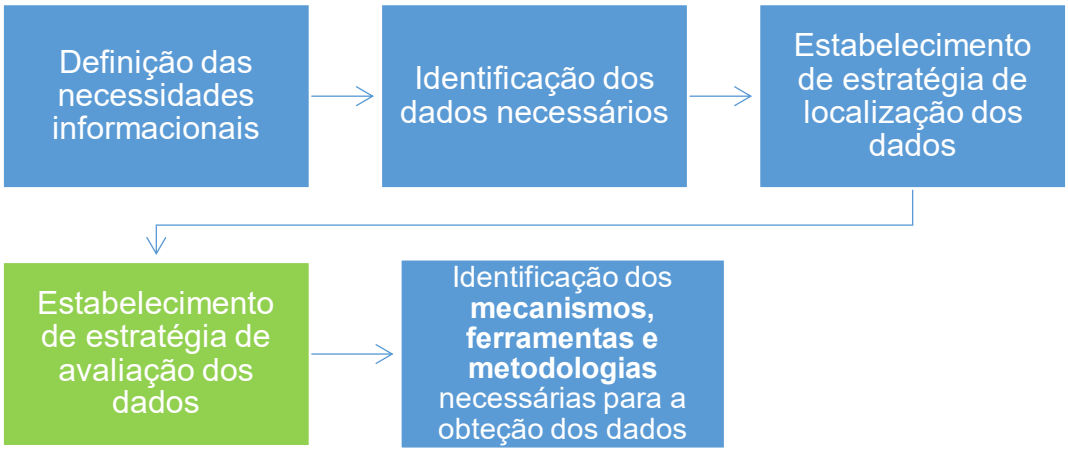
A qualidade dos dados aparece nesse ciclo como um dos pilares que influenciam a condução satisfatória de todas as etapas. Nesse cenário, a avaliação de qualidade permeia diferentes momentos da vida dos dados, sendo realizada com objetivos distintos.

Ao analisar o modelo, entende-se a seleção de fontes como uma tarefa da etapa de coleta de dados. Em relação a etapa de coleta o autor ressalta que:

> A coleta pode ser caracterizada como um projeto ou como um processo. Existem casos em que a coleta ocorre a partir de fonte de dados que permite a aquisição constante dos mesmos, o que geralmente configura situação em que estes dados correspondem a informações que representam situações dinâmicas constituindo um processo, que pode ser contínuo, de fornecimento destes dados (Sant'Ana, 2016, p. 122).

Essas diferentes caracterizações têm impacto direto na seleção das fontes, influenciando em como serão estabelecidos os critérios de seleção utilizados e na necessidade ou não de avaliação contínua de fontes. Para uma melhor compreensão da etapa de coleta, elaborou-se a sistematização apresentada na figura 34:

Figura 34 - Sistematização da etapa de coleta do ciclo de vida dos dados.



Fonte: adaptado de Sant'Ana (2016)

Embora sejam pensadas para representar a etapa de coleta, as atividades apresentadas na figura 34 podem ser diretamente adaptas e aplicadas na seleção de fontes baseada na avaliação de qualidade, uma vez que a seleção possui um caráter fortemente contextual. Mesmo em uma abordagem intrínseca, em que se pretende avaliar se os dados estão livres de erros e más formações, faz-se necessário entender que tipo de dados está sendo avaliado, a quais erros podem estar sujeitos, qual a sua estrutura sintática e se semântica.

Destaca-se a necessidade de atuação ativa do usuário nessa etapa, compostas por atividades de planejamento que exigem discussão aprofundada do projeto/processo de coleta dos dados que se pretende realizar.

Em relação a etapa de coleta, observa-se a importância da "percepção de qualidade", relacionada a aspectos como proveniência, a confiabilidade, e as garantias de integridade dos dados:

> Uma característica essencial de repositórios de dados é a definição e garantia de elementos que permitam a **percepção da qualidade dos dados coletados**, portanto, elementos como a **procedência, mecanismos de coleta e garantias de integridade física e lógica** representam apenas alguns dos aspectos a serem considerados**. A confiabilidade dos dados é condição *sine qua non* para que um dado seja útil.** (Sant'Ana, 2016, p. 119, grifo nosso)

Em relação a presença da qualidade nos processos relacionados a publicação de dados *Linked Data*, Mihindukulasooriya (2020), destaca que os CDVs de dados

Linked Data não apresentam uma etapa de qualidade e/ou preparação dos dados. O autor aponta que o processo de publicação de dados *Linked Data:*

> [...] inclui a transformação de dados a partir de fontes originais, mapeamento de dados para vários vocabulários ou ontologias, e fusão e interligação de dados de diferentes fontes de dados. Cada um desses passos pode abrir a porta para possíveis problemas de qualidade de dados em muitos aspectos diferentes (Mihindukulasooriya, 2020, p. 3, tradução nossa).

Mihindukulasooriya propõe um modelo de CVD focado na qualidade de dados *Linked Data.* A figura 35 apresenta o ciclo proposto:

Figura 35 - Ciclo de vida da qualidade de dados *Linked Data*

Fonte: Mihindukulasooriya (2020)

O autor aponta que "´[...] a avaliação e reparo de qualidade podem ser realizados em todas as diferentes fases do ciclo de vida. Além disso, propomos que a análise de qualidade pode ser considerada como tendo um ciclo de vida próprio" (Mihindukulasooriya, 2020, p. 7).

O ciclo de vida apresentado se aplica tanto aos dados publicados como triplas RDF, como aos vocabulários/ontologias utilizados na definição das classes e propriedades, e aos metadados criados para descrição dos conjuntos de dados.

Nesse CVD, a primeira etapa é totalmente baseada no estabelecimento do modelo de qualidade, que irá orientar a condução do processo de avaliação, esse

modelo será composto pelas dimensões e pelas métricas que serão adotadas para avaliar a qualidade dos dados. O modelo “representa uma especificação de informações relacionadas à qualidade” (Mihindukulasooriya, 2020, p. 6, tradução nossa).

Nesse sentido, o modelo de qualidade de dados se estabelece como uma representação das necessidades informacionais do projeto em questão, refletindo o que se espera em aspectos de qualidade dos dados que serão publicados ou consumidos.

Em seguida, os dados passam pelo processo de avaliação de qualidade. Os problemas de qualidade são então mapeados e tem início a etapa de correção dos problemas identificados.

> Uma vez realizadas as tarefas de avaliação de qualidade e reparo nos dados, os dados limpos podem ser armazenados e utilizados para seus fins pretendidos. Não apenas os dados limpos, mas também os resultados da avaliação de qualidade, bem como os dados intermediários produzidos pelos procedimentos de avaliação de qualidade, podem ser úteis para futuras avaliações de qualidade. Assim, todas essas informações úteis poderiam ser armazenadas e exploradas efetivamente (Mihindukulasooriya, 2020, p. 8, tradução nossa).

Em relação ao CVD proposto, entende-se que em nem todas as etapas do ciclo de vida de dados *Linked Data* a aplicação da etapa de reparação é necessária ou possível. Quando abordamos a questão da seleção de fontes, muitas vezes a etapa de reparo não ocorre, tendo em vista que a identificação de problemas de qualidade pode levar a exclusão da fonte.

Além desse aspecto, concorda-se com Mihindukulasooriya (2020) e com Sant’ana (2016) em relação a aplicabilidade da qualidade de dados em todas as etapas do ciclo de vida dos dados, tendo um impacto importante na qualidade de dados *Linked Data*, cuja origem muitas vezes deriva de processos de conversão automática, de fontes estruturadas e não estruturadas. Como discutido, esses dados possuem níveis distintos de curadorias e são criados com propósitos e para domínios muito distintos, tornando o processo de avaliação complexo, mais indispensável.

Como pode ocorrer em diferentes etapas, a avaliação de qualidade também possuí objetivos distintos. Ela pode ser realizada pelos publicadores como parte do projeto de criação dos dados, ocorrendo pontualmente e visando a identificação de

problemas de qualidade que possam ser reparados. Também pode ser realizada pelos publicadores, visando garantir a manutenção dos níveis de qualidade de seus conjuntos de dados, como um processo contínuo.

Observa-se ainda que a avaliação de qualidade pode ser desempenhada pelos consumidores dos dados, e que nesse caso os objetivos do processo também podem variar, podendo ser relacionados com a comparação de diferentes conjuntos, a avaliação e amadurecimento de um determinado catálogo de dados, ou com o processo de seleção de fontes para interligação.

Quando relacionado com a seleção de fontes, a avaliação pode ser precedida do estabelecimento dos dados necessários, por meio de uma análise das necessidades informacionais ou ocorrer como um processo de exploração, onde se buscam fontes que abordem dados semelhantes que possam ser interligados, sem uma especificação previa do que se está buscando.

Com base na análise da relação entre qualidade de dados e ciclo de vida de dados *Linked Data*, observou-se a importância da interligação para a etapa de consumo de dados. Nesse sentido, a próxima subseção discute a criação de *links* em dados *Linked Data*.

6.1 O processo de interligação no *Linked Data*

Considerando que a criação de *Links* é uma das etapas do ciclo de vida de dados *Linked Data* que possui mais relações com o consumo de dados, sendo muitas vezes o objetivo final da seleção de dados *Linked Data*, e que os ciclos de vida apresentados não se aprofundam no fornecimento de informações sobre como conduzir essa etapa, a presente subseção busca apresentar um panorama geral a respeito da criação de *links* no *Linked Data.*

> Um dos principais obstáculos na publicação de *Linked Data* é conectar o conjunto de dados publicado externamente com fontes de dados relacionadas na nuvem, conhecido como interligação de dados (Kettouch; Luca, 2022, p. 2685, tradução nossa).

A atividade é denominada pela literatura como “*interlinking*”, “*data interlinking*” ou interligação e “[...] representa todo o processo e as etapas necessárias para estabelecer *links* de similaridade entre dois recursos” (Kettouch, 2017, p. 56).

O *Linked Data*, como o próprio nome sugere “significa que os dados estão conectados a outras coisas. Dados isolados raramente são valiosos, no entanto, dados interligados tornam-se repentinamente muito valiosos” (W3C, 2014, não paginado, tradução nossa).

> A interligação LD descreve a tarefa de criar um relacionamento entre uma entidade em um conjunto de dados LD e uma entidade em outro conjunto de dados LD. As interligações podem ser usadas como uma forma de representar que ambas as entidades descrevem a mesma coisa ou como uma forma de indicar que elas são semelhantes ou relacionadas entre si de alguma forma (McKenna; Debruyne; O’Sullivan, 2020, p. 2, tradução nossa).

Os *links* criados no contexto do *Linked Data* podem ser chamados também de *links* RDF uma vez que se baseiam na estrutura de triplas do modelo RDF e tem como característica a presença de um URI na posição de valor da declaração.

> *Links* RDF descrevem o relacionamento entre dois recursos. Os *links* RDF consistem em três referências URI. Os URIs nas posições de sujeito e objeto do *link* identificam os recursos relacionados. O URI na posição de predicado define o tipo de relacionamento entre os recursos. (Heath; Bizer, 2011, não paginado, tradução nossa).

O URI que aparece na posição de predicado/valor nos *links RDF* pode fazer parte do conjunto de dados, criando ligações entre distintos recursos de uma mesma fonte, permitindo a navegação nesses dados. Pode também pertencer a uma fonte externa, permitindo a ligação de *links* externos.

> [...] um *link* RDF externo é um triplo RDF em que o sujeito do triplo é uma referência URI no *namespace* de um conjunto de dados, enquanto o predicado e/ou objeto do triplo são referências URI que apontam para os *namespaces* de outros conjuntos de dados. A desreferenciação desses URIs produz uma descrição do recurso vinculado fornecido pelo servidor remoto. Essa descrição geralmente contém *links* RDF adicionais que apontam para outros URIs que, por sua vez, também podem ser desreferenciados, e assim por diante. (Heath; Bizer, 2011, não paginado, tradução nossa).

Os *links* RDF podem ser categorizados em: 1) *links* de identidade; 2) *links* de relacionamento; 3) *links* de vocabulários (Heath; Bizer, 2011; Kettouch; Luca, 2022).

Os *links* de relacionamento permitem indicar e rotular relacionamentos existentes entre duas entidades de fontes distintas. Para que o tipo de relação

existente entre as entidades possa ser explicitado são utilizadas propriedades de um vocabulário RDF.

Os *links* de identidade são considerados o tipo de *link* mais comum do *Linked Data* (Kettouch; Luca, 2022). Esses *links* são necessários pela característica dos identificadores criados no contexto da *Web*, já que:

> Muitas linguagens têm a chamada suposição de "nomes únicos": nomes diferentes referem-se a coisas diferentes no mundo. Na *web*, tal suposição não é possível. Por exemplo, a mesma pessoa pode ser referida de muitas maneiras diferentes (ou seja, com referências URI diferentes) (W3C, 2004, não paginado, tradução nossa).

Como seria inviável estabelecer um único URI para cada entidade descrita na *Web* em uma escala global, torna-se necessário estabelecer uma forma de conectar as diferentes fontes que dizem respeito a uma mesma entidade.

Esse tipo de *link* é criado geralmente por meio da utilização da propriedade *owl:same*. A propriedade "*owl:sameAs*" é usada para afirmar que duas referências de URI se referem ao mesmo indivíduo." (W3C, 2004, não paginado, tradução nossa).

Já os *links* de vocabulários conectam as propriedades dos vocabulários às suas respectivas descrições e permitem o mapeamento entre vocabulários, possibilitando a identificação de propriedade que possuem significados semelhantes ou idênticos. A interligação pode ser feita de maneira manual ou automática/semiautomática, sendo a abordagem manual normalmente empregada em conjuntos de dados menores (Villazón-Terrazas *et al*., 2011).

Em relação a abordagem manual, os dados podem:

> [...] ser pesquisados manualmente para encontrar os URIs dos recursos de destino para interligação. Se uma fonte de dados não fornecer uma interface de pesquisa, como um *endpoint* SPARQL ou um formulário da *Web* em HTML, um navegador de *Linked Data* pode ser usado para explorar o conjunto de dados e encontrar os URIs relevantes (Heath; Bizer, 2011, não paginado, tradução nossa).

O Sindice [8] é uma ferramenta que pode auxiliar na identificação de fontes de URIs. "O Sindice coleta dados da *Web* de várias maneiras, seguindo os padrões da *Web* existentes, e oferece pesquisa e consulta nesses dados, atualizados ao vivo a cada poucos minutos" (SINDICE, 2025). A *Linked Open Data Cloud* (LOD-Cloud)

[8] https://sindice.com/index.html

também é uma boa fonte de dados *Linked Data*, e pode ser usada tanto em abordagens manuais como automáticas[9].

Ao estabelecer *links* manualmente, deve-se levar em consideração que as declarações a respeito de pessoas, lugares e demais objetos do mundo real devem apontar para o URI que as representam e não para as URLs de documentos que fornecem informações a respeito dessas entidades, o que pode comprometer a semântica das declarações (Heath; Bizer, 2011).

Em relação às abordagens automáticas, elas podem ser baseadas em diferentes tipos de ferramentas, como *frameworks* para identificação de similaridade, abordagens baseadas em chave, algoritmos de *Machine Learning*, e ferramentas com abordagem baseadas em similaridade de esquema e ontologias.

Ao final do processo de criação dos *links* podem ser aplicadas diferentes técnicas de avaliação de qualidade, visando verificar a precisão dos *links* criados. Também podem ser empregadas avaliações manuais ou baseadas em técnicas de *crowdsourcing* e amostragem estatística.

As abordagens baseadas em chaves se aproveitam de identificadores únicos estabelecidos no domínio em questão, como por exemplo *International Standard Book* (ISBN), identificador único internacional que pode ser atribuído a publicações bibliográficas.

> Se um conjunto de dados contiver tais identificadores, estes devem ser expostos como parte dos URIs ou como valores de propriedade. Tais propriedades são chamadas de *propriedades funcionais inversas*, pois seu valor identifica exclusivamente o sujeito da tripla e deve ser definido como tal no vocabulário correspondente (Heath; Bizer, 2011, não paginado, tradução nossa).

A inclusão dessas propriedades para explicitar a relação com identificadores torna possível a utilização de algoritmos que identificam URIs que representam uma mesma entidade, permitindo a criação automática ou semiautomática de *links.*

As abordagens baseadas em ontologias e esquemas buscam identificar relações com base na correspondência entre conceitos e propriedades de ontologias:

> Ontologias em interligação de dados são geralmente usadas para identificar e comparar instâncias que fazem parte das mesmas

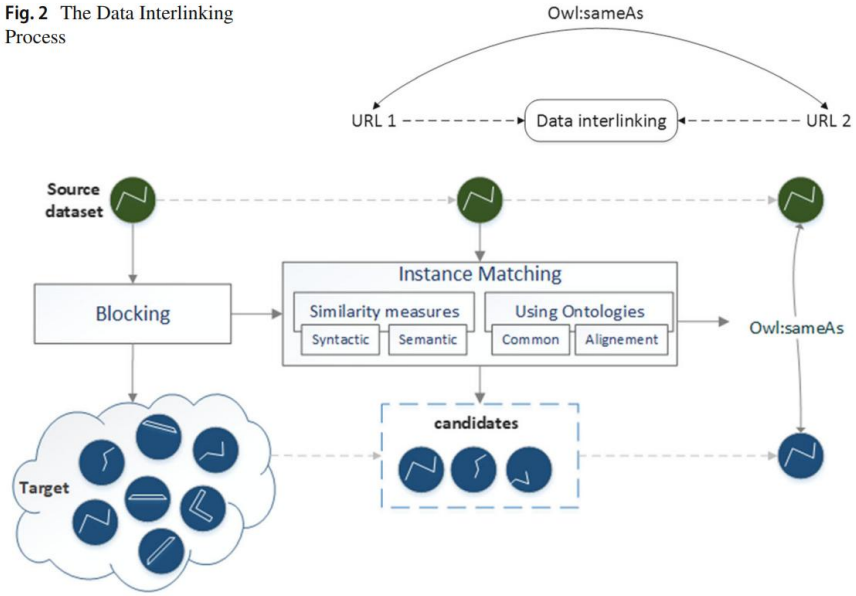
[9] https://lod-cloud.net/

> classes, com base no fato de possuírem as mesmas propriedades. O uso de ontologias não exclui a possibilidade de usar outras técnicas de similaridade dados (Kettouch; Luca, 2022, p. 2691, tradução nossa).

A outra forma de criar *links* automaticamente é adotando uma estratégia baseada em similaridade, utilizando ferramentas que permitem a identificação de porcentagens de similaridade entre duas fontes de dados.

Kettouch e Luca (2022) apresenta uma sistematização do fluxo de criação de *links* com abordagem de similaridade, focado na criação de *links* de identidade baseados na propriedade OWL. A sistematização pode ser vista na figura 36.

Figura 36 - processo de interligação baseado na propriedade *owl:sameAs*



Fonte: Kettouch e Luca (2022)

Nessa sistematização é possível observar que a ligação automática tem início com a identificação das fontes, e que o conjunto de dados passa então pelo processo de "*blocking*".

> É o estágio inicial do processo de interligação, por meio do qual o número de candidatos é reduzido. Como resultado, um bloco que consiste em um conjunto de pares de instâncias potencialmente idênticos é gerado. Esta é uma etapa importante, pois afeta o desempenho do sistema, considerando que as entradas das operações de processamento pesadas na etapa de correspondência de instâncias terão resultado do bloqueio (Kettouch; Luca, 2022, p. 55, tradução nossa).

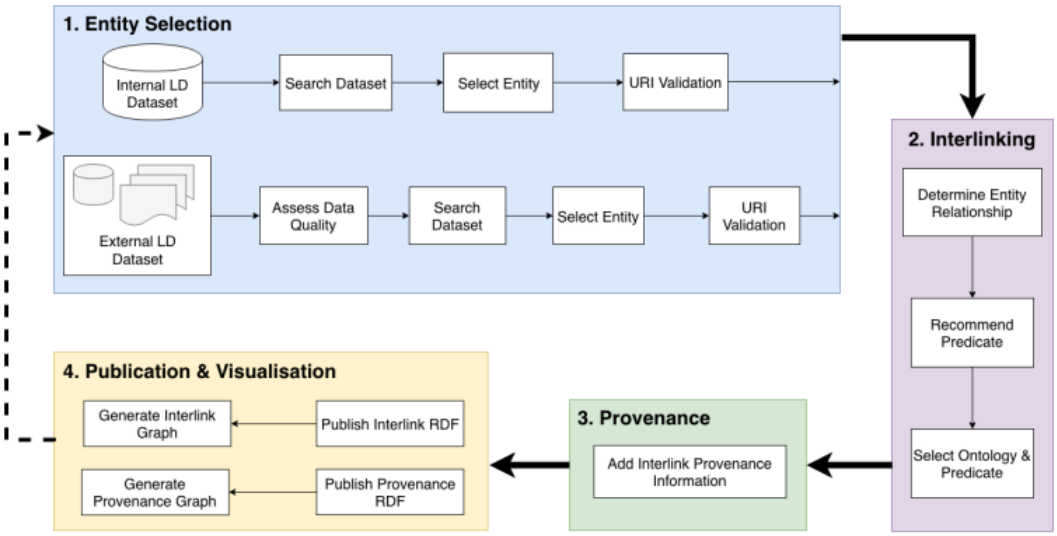
Em seguida algoritmos podem ser empregados para mensurar a similaridade semântica e sintática dos pares em potencial e/ou utilizada uma abordagem baseada em ontologias e esquemas e as ligações são construídas.

McKenna, Debruyne e O'Sullivan (2020) apresenta uma proposta e ferramenta para interligação, o NAISC, uma abordagem que permite a identificação do tipo de relacionamento existente entre as entidades alvo do processo de ligação. Os autores apontam que "uma exploração mais aprofundada da questão da interligação destacou que os processos de ontologia e seleção do tipo de ligação (determinação e descrição da relação entre duas entidades) eram áreas de particular dificuldade" (McKenna; Debruyne; O'Sullivan, 2020, p. 1).

Essa dificuldade está relacionada com os desafios da criação de *links* de relacionamento, que devem partir da identificação do tipo de relacionamento existente entre as entidades e da identificação de uma propriedade em um vocabulário bem estabelecido que represente essa relação.

Embora tenha sido pensada para mapear o fluxo da ferramenta proposta, os autores apresentam uma sistematização bastante completa que reflete bem as etapas e atividades da interligação. A figura 37 apresenta o *framework* do processo de interligação.

Figura 37 - framework do processo de interligação



Fonte: McKenna, Debruyne e O'Sullivan (2020)

Nesse *framework* observa-se que a ligação tem início com o processo de seleção de entidades, que se baseia no fornecimento de dados de entrada da fonte

interna de dados e dados da fonte externa. Os dados da fonte interna passam por um processo de filtragem e validação dos URIs, enquanto os dados da fonte externa possuem uma camada adicional, a de avaliação de qualidade.

A interligação ocorre com base no estabelecimento do tipo de *link* que se pretende estabelecer e da identificação de propriedades que possam representar essa relação.

A ferramenta NAISC aborda a questão da atribuição da proveniência para os *links* criados, que deve agregar aos dados, com base em um modelo informações (que na ferramenta é baseada no vocabulário PROV), e que representam quem, onde, quando, por que e como o *link* foi criado (McKenna, Debruyne e O'Sullivan, 2020).

As ferramentas Limes[10] e Silk [11]também podem ser utilizadas para identificar ligações em potencial com base na similaridade entre duas fontes (Rautenberg *et al*., 2018).

O limes é:

> [...] um framework de descoberta de *links* entre recursos na *web* [...] usa características matemáticas de espaços métricos durante o processo de mapeamento para filtrar e associar grandes quantidades de pares de instâncias (Rautenberg *et al*., 2018, p. 80).

Para a identificação dos *links* na ferramenta o usuário pode tanto configurar a relação ou configurar a ferramenta para que utilize algoritmos de *machine learning* no processo (Ngomo *et al*., 2021).

O Silk é um *framework* de código aberto para interligação, baseado nos princípios do *Linked Data*, que se utiliza do modelo de dados RDF e do protocolo SPARQL para acessar fontes de dados e identificar ligações em potencial, permitindo ainda especificar quais tipos de ligação se espera descobrir entre as fontes de dados indicadas (Rautenberg *et al.*, 2018).

O silk permite:

> [...] especificar quais tipos de *links* RDF devem ser descobertos entre fontes de dados, bem como quais condições os itens de dados devem atender para serem interligados. Essas condições de *link* podem combinar diversas métricas de similaridade e levar em consideração o

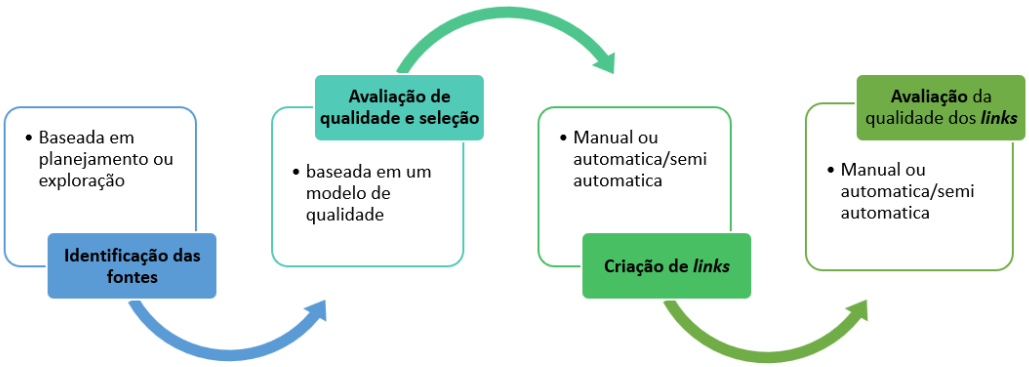
[10] https://aksw.org/Projects/LIMES

[11] http://silkframework.org/

> grafo em torno de um item de dados, o que é abordado por meio de uma linguagem de caminho RDF. O Silk acessa as fontes de dados que devem ser interligadas por meio do protocolo SPARQL e, portanto, pode ser usado em *endpoints* SPARQL locais e remotos (SILK, 2024, não paginado, tradução nossa).

Independentemente do tipo de *link* criado e da ferramenta adotada para interligação, observa-se a importância da avaliação de qualidade para a seleção de fontes de dados. Observa-se ainda a necessidade de uma etapa de planejamento da qualidade, onde se irá estabelecer como os dados serão avaliados e selecionados, quais os propósitos e a frequência da seleção e ainda quais as ferramentas que serão adotadas para a avaliação dos dados, criação dos *links* e avaliação da qualidade dos *links* gerados. A figura 38 apresenta uma síntese do processo de interligação.

Figura 38 -síntese do processo de interligação



Fonte: Autora (2025)

Compreendendo a importância dos processos que permeiam a qualidade para a seleção de fontes de dados *Linked Data*, a próxima seção apresenta e discute os principais processos e atividades relacionados à qualidade de dados.

6.2 Abordagens processuais da avaliação de qualidade

Visando compreender como selecionar dados com base em aspectos de qualidade, a presente subseção busca discutir os processos, atividades e etapas relacionados com a qualidade dos dados, com foco nas atividades de avaliação de qualidade.

Os aspectos processuais da qualidade podem ser abordados por meio do gerenciamento de qualidade, que reúne três grandes processos: 1) planejamento de qualidade, 2) controle de qualidade e 3) melhoria da qualidade (Juran; Godfrey, 1998).

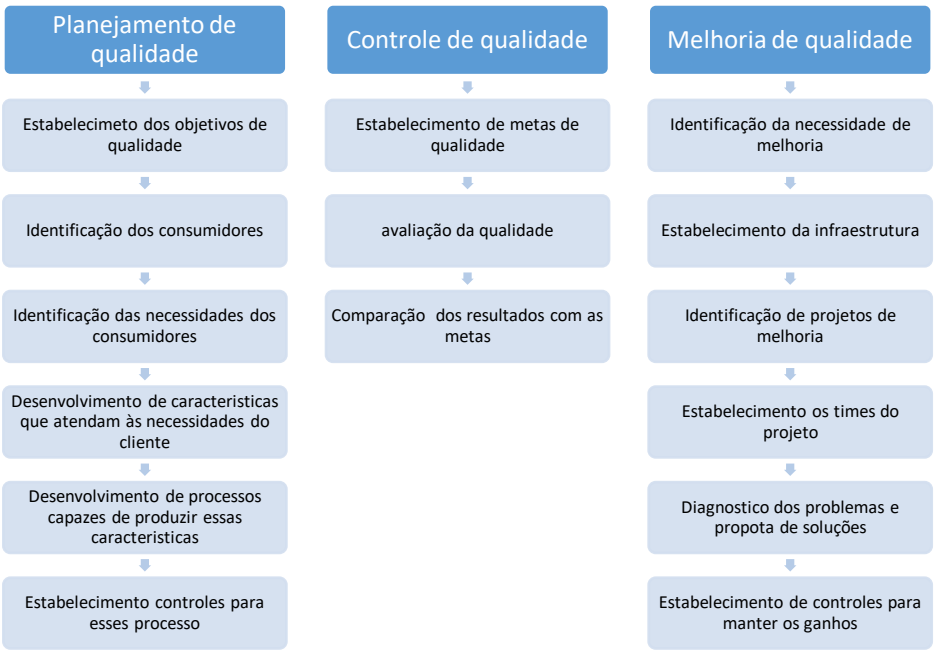
O planejamento de qualidade visa fornecer ferramentas, métodos e técnicas para que o produto atenda às necessidades de uso para a qual foi elaborado. O **controle de qualidade** é conduzido visando manter os níveis de qualidade estáveis, e consiste em avaliar o desempenho do produto e compará-lo com as metas de qualidade estabelecidas. O processo de **melhoria de qualidade** visa possibilitar uma mudança benéfica nos níveis de qualidade, reduzindo a diferença existente entre os níveis atuais e as metas de qualidade (Juran; Godfrey, 1998).

Esses processos são permeados pela avaliação de qualidade, que deve ser parte do planejamento e que fornece insumos para a aplicação de estratégias de controle e melhoria da qualidade.

Tanto os processos quanto a maior parte dos modelos e instrumentos relacionados com o a qualidade, inclusive no contexto da qualidade de dados, são pensados para orientar os produtores, o que se reflete também nas sistematizações desses processos.

Embora não sejam discutidos pela perspectiva do consumidor, entender esses processos é relevante pois reúnem os mesmos atores: consumidores, produtores, produtos, objetivos e indicadores de qualidade. Uma síntese do gerenciamento de qualidade organizado em três grandes processos é apresentada na figura 39.

Figura 39 - Síntese do gerenciamento de qualidade



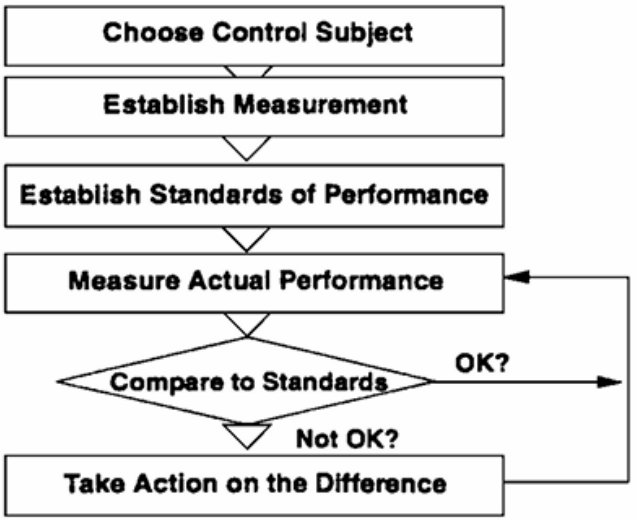
Fonte: baseado em Juran e Godfrey (1998)

Como observado na análise dos CVDs e da interligação no *Linked Data*, o planejamento de qualidade é uma etapa importante também para os consumidores, na medida em que precisam desse planejamento para que possam selecionar os dados.

É necessário estabelecer os objetivos de qualidade, estabelecer as necessidades informacionais, identificar os fornecedores em potencial e, a depender da abordagem adotada, identificar o grupo de entidades alvo do processo de ligação e os tipos de relacionamento que se pretende estabelecer com as fontes externas.

Pode-se traçar um paralelo entre o controle de qualidade e a avaliação de qualidade para seleção de fontes. Embora possuam objetivos distintos, os processos de controle de qualidade e avaliação de qualidade se concentram nas mesmas atividades: o diagnóstico dos níveis de qualidade dos dados e identificação de potenciais espaços para melhoria da qualidade. A figura 40 apresenta a sistematização do controle de qualidade.

Figura 40 - sistematização do controle de qualidade.



Fonte: Juran e Godfrey (1989)

Basili, Caldiera e Rombach (1994) apontam que o processo eficiente de avaliação precisa ter como base:

1. O foco em objetivos específicos;
2. Aplicação a todo o ciclo de vida;
3. Ser interpretável com base em caracterização e compreensão do contexto e ambiente de aplicação.

Nesse sentido, um aspecto importante para a condução da avaliação de qualidade é o estabelecimento dos objetos e objetivos do processo. O estabelecimento claro desse desses aspectos terá impacto na formulação do modelo de qualidade a ser utilizado, no estabelecimento das dimensões e métricas adotas e no processo de análise posterior dos resultados, uma vez que “quais métricas usar e como interpretá-las não fica claro sem os modelos e objetivos apropriados para definir o contexto” (Basili; Caldiera; Rombach, 1994, p. 2, tradução nossa).

Batinni e Scanapieco (2016) apresentam uma lista com os principais objetivos que podem ser relacionados ao processo de avaliação de avaliação de qualidade:

1) **Aquisição de novos dados –** para atualizar e completar o conjunto de dados;
2) **Padronização ou normalização –** para adequar o conjunto de dados a uma política ou norma;
3) **Identificação de objetos idênticos/semelhantes –** identificar registros que se referem a um mesmo objeto do mundo real, presente em fontes diversas;

4) **Integração de dados –** apresentar uma versão unificada de dados pertencentes a fontes heterogenias para o fornecimento de resultados de consulta mais precisos e para a resolução de conflitos de valores que se referem a um mesmo objeto do mundo real;
5) **Estabelecer a confiabilidade das fontes –** classificar e ranquear fontes com base nas informações que elas fornecem;
6) **Composição de qualidade –** avaliar os níveis de qualidade após a combinação de duas ou mais fontes;
7) **Localização de erros –** encontrar registros que não respeitam as regras sintáticas e semânticas aos quais estão sujeitos;
8) **Correção de erros –** identificar e corrigir erros relacionados a violação de regras semânticas e a presença de valores incorretos;
9) **Otimização de custos –** identificar entre fornecedores de dados, qual oferece a melhor relação custo/qualidade;
10) **Correspondência de esquema –** produzir o mapeamento entre diferentes esquemas (ou vocabulários no contexto do *Linked Data*);
11) **Limpeza de esquema –** identificar alterações no esquema que possam melhorar os níveis de qualidade dos dados; e
12) **Perfilamento –** explorar e compreender características dos conjuntos de dados disponíveis.

Definidos os objetivos da condução da avaliação de qualidade, podem ser estabelecidos os dados que serão objetos dessa avaliação. Existem diversas formas de selecionar os dados, a identificação pode derivar de uma necessidade declarada dos consumidores, da aplicação de tecnologias para identificar necessidades não declaradas, da identificação de aspectos intrínsecos do produto que afetam seus níveis de qualidade, da comparação com padrões de qualidade, como normas, políticas e melhores práticas (Juran; Godfrey, 1989).

A próxima etapa do processo é estabelecer como será realizada a avaliação da qualidade:

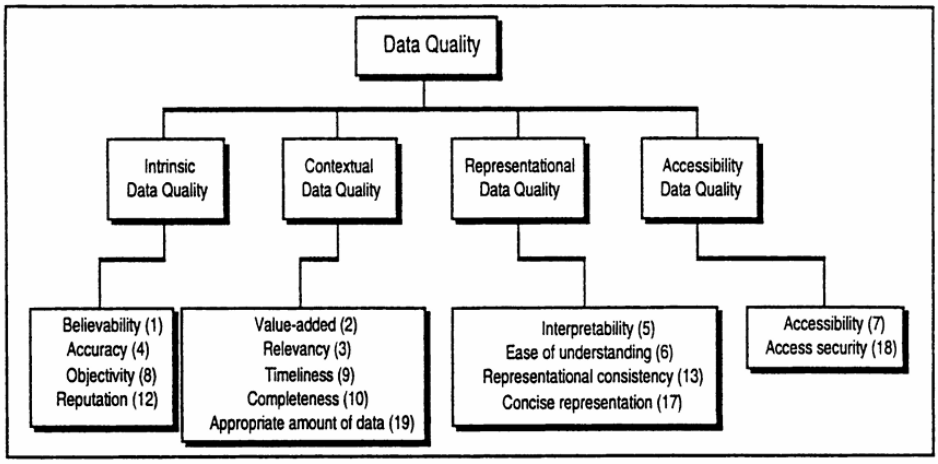
> Ao estabelecer a medição, precisamos especificar claramente os meios de medição (o sensor), a frequência da medição, a forma como os dados serão registrados, o formato para relatar os dados, a análise que será feita nos dados para converter os dados em informações

> utilizáveis, e quem fará a medição (Juran; Godfrey,1989, p. 4.6, tradução nossa).

Essa etapa está relacionada ao planejamento do processo de avaliação de qualidade. Como discutido tanto na análise teórica como na análise terminológica da qualidade de dados, a sistematização do processo de avaliação de qualidade de dados proposta por Wang e Strong (1996) segue sendo relevante para os processos atuais de avaliação de qualidade.

Com base no seu estudo a respeito da qualidade na perspectiva de consumidores de dados, Wang e Strong (1996), apresentam então um *framework* de qualidade de dados, estruturado em formato de arvore de domínio. A figura 41 apresenta esse *framework:*

Figura 41 - *Framework* da qualidade de dados proposto por Wang e Strong



Fonte: Wang e Strong (1996)

Nesse modelo, são organizados aspectos de avaliação de qualidade em categorias e dimensões, que podem então ser mensurados com base em critérios e métricas. Esse modelo pode ser adaptado às necessidades e objetivos do processo de avaliação de qualidade. Para reunir todas as informações sobre como será conduzida a avaliação de qualidade, quais as categorias, dimensões, critérios e métricas adotados, são criados e aplicados os modelos de qualidade de dados.

Compreendendo a importância do modelo de qualidade para a avaliação de qualidade, a próxima subseção apresenta e discute os modelos de qualidade de dados relacionados ao processo de avaliação de dados *Linked Data.*

6.3 Os modelos de qualidade para dados *Linked Data*

Foram identificados na literatura 10 modelos de qualidade de dados criados especificamente para auxiliar na avaliação de dados *Linked Data*, também foi considerado nas discussões o modelo de qualidade estabelecido pela ISO, levando em consideração a sua relevância no contexto da avaliação de qualidade de dados e sua influência em outros modelos e ferramentas de avaliação.

Os modelos foram analisados quanto ao seu contexto de criação, enfoque, estrutura, dimensões, critérios e métricas. Foram analisados ainda as metodologias utilizadas para eleger as amostras a serem avaliadas e as ferramentas utilizadas para auxiliar na avaliação das métricas.

Os modelos identificados apresentam diferentes níveis de granularidade, de formalização das métricas apresentadas e de detalhamento sobre como essas métricas podem ser aplicadas.

Também se observou uma variação em relação a como são estruturados e apresentados os modelos, possuindo diferentes níveis de sistematização. Enquanto alguns modelos são apresentados em formato de quadros correlacionando categorias, dimensões, critérios e métricas, outros apresentam uma lista organizada em seções e subseções, onde as métricas adotadas são apresentadas e discutidas em profundidade.

Os modelos também diferem em relação a terminologia e aos níveis de hierarquia adotados, podendo adotar a estrutura de categoria, dimensões, critérios e métricas, apenas alguns desses níveis ou ainda a nomenclatura proposta pela ISO 25012.

Visando facilitar a análise, comparação e apresentação dos modelos, buscou-se sistematizar todos em formato de quadro resumo, apresentando, sempre que possível, dimensões, critérios e métricas, bem como uma breve explicação do que se busca analisar com as métricas propostas. Considerando a extensão dos modelos e, em muitos casos, a necessidade da inserção de fórmulas para apresentar as métricas utilizadas, optou-se por apresentar a sistematização dos modelos como dados de pesquisa. As próximas subseções discutem os modelos analisados. [12]

[12] Ressalta-se que, considerando a diferença de extensão, detalhamento das fontes e profundidade das explicações dos modelos, será observada variação em relação a profundidade e extensão das discussões dos

6.3.1 Modelo de qualidade proposto pela norma ISO 25012

Considerando seu impacto na estrutura e na elaboração de outros modelos de qualidade, torna-se relevante mencionar o modelo proposto na norma ISO 25012 de 2008, intitulada "*Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*" (ISO, 2008).

O modelo tem como escopo "dados retidos em um formato estruturado dentro de um sistema de computador" (ISO, 2008, p. 1, tradução nossa). Foi pensado para:

> [...] definir e avaliar os requisitos de qualidade dos dados em processos de produção, aquisição e integração de dados; – identificar critérios de garantia de qualidade dos dados, também úteis para reengenharia, avaliação e melhoria dos dados; – avaliar a conformidade dos dados com a legislação e/ou requisitos. A detecção de erros ou ineficiências devido a dados gera intervenções de aprimoramento e correção referente aos dados e outros componentes do sistema em que os dados residem (ISO, 2008, p. 1, tradução nossa).

Com base no escopo estabelecido, o modelo tem como foco dados derivados/armazenados em sistemas computacionais, e as atividades de produção, aquisição, integração e detecção de erros visando a sua correção.

O modelo apresenta 15 características de qualidade, que equivalem as dimensões de qualidade.

Essas características são divididas entre inerentes e dependentes dos sistemas, sendo algumas delas simultaneamente inerentes e dependentes do sistema.

"Do ponto de vista inerente, a qualidade dos dados refere-se aos próprios dados[...]" (ISO, 2008, p. 3, tradução nossa) enquanto as características dependentes do sistema referem-se "[...] ao grau em que a qualidade dos dados é alcançada e preservada dentro de um sistema computacional quando os dados são utilizados em condições especificadas" (ISO, 2008, p. 4, tradução nossa).

---

modelos. As análises dos modelos podem ser acessadas no repositório da UNESP e pelo link: https://docs.google.com/spreadsheets/d/1vnsXcUQFSG2TTl3FeMh8cQMaRdEqeNio2DWUcnsbhlg/edit?usp=sharing

São apresentadas então as características de qualidade organizadas em uma das três categorias propostas, seguidas de uma definição e de exemplos de possíveis formas de avaliação. Esses exemplos podem ou não conter métricas e possuem um aspecto ilustrativo e não prescritivo.

Essa estrutura geral e o aspecto não normativo dos exemplos, torna o modelo abrangente a distintos domínios, mas também dificulta a sua aplicação pratica, já que para aplicá-lo torna-se necessária a formalização de métricas e escalas para cada uma das suas dimensões.

6.3.2 Modelo de qualidade proposto por Zaveri *et al.* (2012)

Zaveri *et al.* (2012) realizaram o primeiro trabalho sistemático de criação de um modelo para dados *Linked Data*. Os autores reuniram uma série de dimensões e métricas criadas para o contexto da *Web*, da *Web* semântica e do *Linked Data* por meio de uma Revisão Sistemática da Literatura.

Com base nos resultados da RSL, os autores organizaram os aspectos de qualidade em categorias e dimensões. Embora o conceito de critério não seja formalmente apresentado, as dimensões são compostas por subdimensões que podem ser avaliadas por diferentes métricas, de maneira equivalente a um critério de qualidade.

Essa questão se torna mais evidente considerando-se que em muitas situações são apresentadas mais de uma forma de se avaliar cada critério, como nos casos da dimensão “Valores sintaticamente precisos” da dimensão validade sintática, onde são apresentadas 4 formas possíveis de avaliação, ou no caso de “Ausência de valores imprecisos” da dimensão precisão semântica, onde são apresentadas 3 formas de avaliação.

O modelo é dividido nas categorias Intrínseca, contextual, representacional e acessibilidade. Foram identificadas 18 dimensões que reúnem 69 critérios de qualidade, tendo cada critério ao menos uma forma de avaliação possível. Ao longo do artigo, os autores apresentam ainda diversos exemplos sobre como cada da dimensão pode ser avaliada.

O modelo é de aspecto geral, não sendo necessariamente direcionado a consumidores ou publicadores, a um domínio específico ou uma fase específica do processo de avaliação de qualidade. Segue uma linha prescritiva, a sentando em

muitos casos uma série de verificações e as formas como essas verificações podem ser realizadas.

Nota-se ainda, nesse modelo, a relevância de vocabulários e de suas propriedades. Propriedades de vocabulários como OWL e RDFs são exploradas como forma de realizar as verificações de qualidade em diversos critérios. Um exemplo é a dimensão “interligação”, que pode ser avaliada com base na presença da propriedade “*owl:sameAs*”. A OWL também é explorada para verifica o uso incorreto de classes e propriedades e o uso de propriedades obsoletas.

Os vocabulários são explorados para avaliar a qualidade da estrutura dos dados, verificando as relações de dependência entre as propriedades, a adoção e vocabulários bem estabelecidos no domínio e nível de reutilização de vocabulários em relação ao estabelecimento de vocabulários novos.

Um outro aspecto em que são explorados os vocabulários é a obtenção de metadados a respeito do conjunto de dados, como para a obtenção de informações de licença.

Em relação as formas de análise dos dados, o modelo menciona para alguns critérios a possibilidade de uso de técnicas de *crowdsourcing*, de amostragem e de identificação de problemas de qualidade baseados na identificação de *outliers*. Em casos pontuais o modelo indica, inclusive, a necessidade de verificação manual.

Muitos dos aspectos de confiabilidade indicam a utilização de listas de confiança, a consulta a avaliação de outros usuários ou a consulta a lista de dados com baixa reputação, entretanto esse tipo de lista no contexto do *Linked Data* nem sempre está disponível, as que existem estão sujeitas a indisponibilidade e obsolescência.

6.3.3 Modelo de qualidade proposto por Behkamal *et al.* (2014)

Behkamal *et al.* (2014) propuseram um modelo de qualidade focado na acurácia de dados *Linked Data*. Os autores justificam a necessidade do modelo considerando que:

> O elevado número de fatores de qualidade e sua inter-relação tornam a avaliação da qualidade um problema complexo e difícil de considerar todos os fatores de qualidade simultaneamente. Para estudar a qualidade dos dados em profundidade, é necessário estudar cada

> fator de qualidade separadamente, bem como as propriedades do ambiente que o afetam (Behkamal *et al*., 2014, p. 80, tradução nossa).

O foco na acurácia é justificado pelos autores pois a maioria dos problemas de qualidade por eles observados estava relacionada a falta de precisão das fontes. É um modelo criado para publicadores de dados, já que seu foco é a avaliação para a melhoria, garantindo a qualidade dos dados no momento da sua publicação.

Outro aspecto relevante em relação a esse modelo é que foi adotada uma metodologia para o estabelecimento e seleção das métricas que iriam compor o modelo. A metodologia adotada foi a *Goad-Question-Metric* (GQM), onde “objetivos são gradualmente refinados em várias perguntas e cada pergunta é então refinada em métricas. Além disso, uma métrica pode ser usada para responder a múltiplas perguntas” (Behkamal *et al*., 2014, p. 81, tradução nossa).

O modelo possui duas dimensões: Acurácia Sintática e Acurácia Semântica. Foram estabelecidos 6 critérios de acurácia semântica e 9 critérios para acurácia sintática. São apresentadas métricas e explicações para cada métrica.

Em relação as formas de avaliação, o foco da maior parte do modelo está no uso adequado de classes e propriedades. São apresentadas formas de verificar valores ausentes em propriedades, propriedades e classes com erros ortográficos, propriedades sendo aplicadas fora do seu alcance, propriedades cujo valor é de um tipo diferente do especificado, uso incorreto de propriedade e classes mutuamente excludentes, ou o uso de classes e propriedades deslocadas, como uso de propriedades como classes e de classes como propriedades.

Nesse sentido, ressalta-se o papel fundamental dos vocabulários na avaliação da acurácia sintática e semântica dos dados *Linked Data* e a importância da análise da documentação que especifica esses vocabulários para aplicação correta de classes e propriedade.

6.3.4 Modelo de qualidade proposto por Debattista *et al.* (2015)

Debattista *et al.* (2015) apresentam um modelo simplificado de avaliação de qualidade que combina métricas tradicionais com técnicas de análise probabilística para dados *Big Data*, considerando que esse muitas vezes é o contexto no qual os dados *Linked Data* estão inseridos.

As técnicas empregadas pelos autores são*: Reservoir Sampling, Bloom Filters and Clustering Coefficient estimation.*

Reservoir sampling é uma técnica baseada em estatísticas da família dos algoritmos aleatórios que facilita a criação de amostragem de itens uniformemente distribuídos. *Bloom Filter* é uma técnica adotada para consultar se um elemento é parte de um conjunto. (Debattista *et al.*, 2015). Já o *Clustering Coefficient estimation* "[...] mede a densidade das vizinhanças de um nó. O coeficiente de agrupamento é medido dividindo o número de arestas de um nó pelo número de conexões possíveis que os nós vizinhos podem ter" (Debattista *et al.*, 2015, p. 4, tradução nossa).

Os autores trabalham com 4 critérios de qualidade, sendo eles: desreferenciabilidade, concisão extensional, coeficiente de agrupamento de uma rede e existência de *links* para fornecedores de dados externos. O quadro 19 apresenta a relação entre os critérios e as técnicas que podem ser empregadas na sua avaliação.

Quadro 19 - Relação entre critérios de qualidade e técnicas para avaliação

Critério	Descrição	Técnica empregada
Desreferenciabilidade	A métrica de desreferenciabilidade avalia um conjunto de dados contando o número de URIs desreferenciáveis válidos. Não sendo muitas vezes possível desreferenciar todos os recursos URI que aparecem no sujeito e no objeto de todas as triplas.	*Reservoir Sampling*
***Links* para Fornecedores de Dados Externos**	Essa métrica mede o grau em que um recurso está vinculado a provedores de dados externos, podendo ser calculada por meio de estimativa em casos de grandes bancos de dados	*Reservoir Sampling*
Concisão Extensional	Essa métrica mede o número de instâncias únicas a partir de suas propriedades e valores. A abordagem tradicional é comparar cada recurso com todos os outros recursos no conjunto, o que pode ser computacionalmente inviável.	*Bloom Filters*
Coeficiente de Agrupamento de uma Rede	Está métrica tem como objetivo identificar quão bem os recursos são conectados,	*Clustering Coefficient Estimation*

	medindo a densidade da vizinhança de recursos. Calcular essa medida pode ser complexo pois isso acontece porque cada vértice na rede precisa ser considerado.	

Adaptado de Debattista *et al.* (2015)

Como é possível observar, O modelo se concentra, portanto, em apresentar técnicas para auxiliar a avaliação de critérios de qualidade específicos.

6.3.5 Modelo de qualidade proposto por Färber *et al.* (2016)

Färber *et al.* (2016) propõe um modelo que visa permitir a avaliação e comparação de resultados de qualidade, focado na avaliação de dados *Linked Data* de conhecimentos gerais (domínio cruzado ou enciclopédico), publicados em formato aberto. Se organiza em dimensões, critérios e métricas, seguindo as categorias propostas por Wang e Strong (1996).

Consiste em um modelo geral, organizado nas categorias intrínseca, contextual, representacional e acessibilidade, cobrindo a maior parte das dimensões propostas por Zeveri *et al.* (2012). Entretanto, apresenta diferenças significativas em relação aos critérios de qualidade abordados e a sua forma de avaliação.

O modelo foi elaborado de maneira a ser diretamente aplicável, destacando-se, portanto, pelo rigor em relação à apresentação de métricas para todos os critérios apresentados, divididas entre fórmulas e escalas, além da apresentação de informações relevantes para que esses critérios possam ser avaliados, como a indicação de necessidade de ferramentas complementares.

Em relação as formas de avaliação dos critérios e métricas, são mencionados o uso de ferramentas para auxiliar nesse processo, como a ferramenta RDF *validator*, do W3C, que permite checar a Validade Sintática de documentos RDF.

Para a análise de alguns critérios, como “Validade sintática de Literais”, é mencionada a checagem de aderência a políticas e normas externas, no caso exemplificado, os autores propõe a verificação da aderência ao padrão ISO 8601.

O modelo trabalha a questão da confiabilidade dos dados, um aspecto bastante relevante para o processo de seleção de fontes. Para essa análise se considera a ligação com fontes reconhecidamente confiáveis no domínio e a adoção de

vocabulários de proveniência, como o *DublinCoreMetadata* (*dcterms:provenance* e *dcterms:source*) e o W3CPROV-O (*prov:wasDerivedFrom*).

Outro aspecto a ser considerado no modelo para estabelecer a confiabilidade dos dados é a forma como esses dados foram obtidos, sendo considerados mais confiáveis os dados convertidos de fontes estruturadas que os de fontes não estruturadas. A escala também determina como mais confiáveis os dados que passam por curadoria manual, e menos confiáveis os dados convertidos automaticamente que não passaram por processo de curadoria.

Observa-se que esse é um dos modelos que mais explora o uso de vocabulários para a avaliação dos critérios e métricas. Eles podem ser utilizados para a checagem de informações de proveniência, de acesso e de licença. Podem ser aplicados na verificação de aspectos estruturais relacionados à acurácia e para a verificação de diferentes aspectos de completude dos dados, como a completude de conjunto e de esquema. Podem ainda ser utilizados para avaliação dos níveis de interoperabilidade, por meio do uso de propriedades da OWL que indicam equivalência entre vocabulários ou dos níveis de interligação, por meio da *owl:sameAs*.

O modelo considera ainda os vocabulários para verificar aspectos representacionais de qualidade, considerando um indicativo de boa qualidade a presença de descrição de maneira compreensível a humanos, por meio presença das propriedades *rdfs:label* e *rdfs:comment*.

O modelo aborda a presença de aspectos descritivos. A presença de metadados é abordada na categoria acessibilidade, onde o provisionamento de metadados é entendido como um indicativo de boa qualidade. É discutido que o provisionamento de informação pode ser via *sitemaps* semânticos ou via vocabulário VoID. Os metadados podem então ser anexados ao conjunto de dados ou fornecidos como um arquivo VoID separado.

Por fim, outro aspecto a ser destacado em relação a esse modelo é a adoção de padrões-ouro para auxiliar na avaliação de aspectos subjetivos ou de difícil mensuração. Os padrões-ouro são modelos pré-estabelecidos que podem ser utilizados como base para avaliar diversos fatores, no caso desse modelo o uso do padrão-ouro foi sugerido para medir a completude de esquema e de população do conjunto de dados.

6.3.6 Modelo de qualidade proposto por Cappiello *et al.* (2016)

Cappiello *et al.* (2016) apresentam a proposta de um modelo de qualidade minimalista, elaborando para embasar a etapa exploração de dados *Linked Data,* necessária para identificar conjuntos de dados de interesse.

Esse modelo seria aplicado em uma etapa de pré-avaliação, visando diminuir a quantidade de fontes candidatas, que posteriormente seriam submetidas a um processo de avaliação mais rigoroso. Dentre os modelos identificados, é o que mais possui relação com o processo de seleção e com a etapa de consumo de dados do ciclo de vida.

O modelo é composto por 5 dimensões: quantidade de dados, concisão, completude, navegabilidade e interligação, que não são organizadas em categorias e critérios.

Não são apresentadas fórmulas ou escalas para a avaliação da qualidade e sim explicações sobre como podem ser avaliadas. Para a avaliação da completude, é indicada também a criação/adoção de um padrão-ouro.

Nesse modelo o único vocabulário diretamente explorado é o OWL, que é utilizado visando verificar os níveis de interligação por meio da propriedade *owl:sameAs*.

O modelo não discute como estabelecer valores ou porcentagens mínimas em relação a essas dimensões para que os dados possam ou não ser considerados selecionados.

6.3.7 Modelo de qualidade proposto por Melo (2017)

Melo (2017) apresenta um modelo construído com base em revisão de literatura, em melhores práticas e padrões do W3C e nos requisitos estabelecidos pelo projeto Linked *Open Data*. O projeto “estabelece princípios de qualidade, reúne *datasets*, os organiza em diferentes categorias e promove a visibilidade dos que atendem a tais princípios” (Melo, 2017, p. 68).

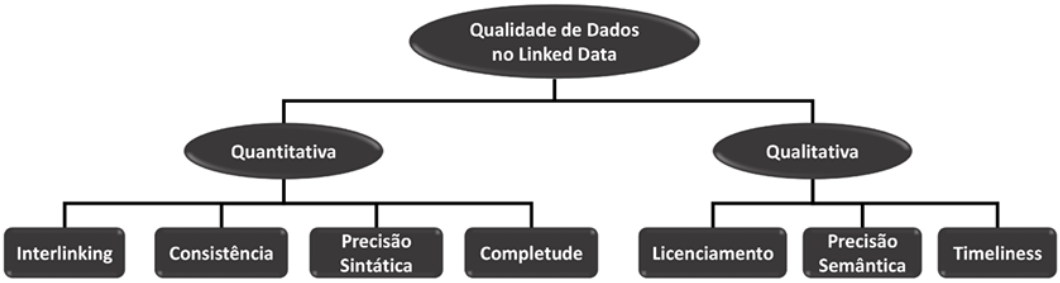
O modelo foi elaborado com uma proposta de abordagem interdisciplinar, para ser aplicado em diferentes contextos. Para a seleção das dimensões foram consideradas apenas dimensões relacionadas com os princípios de *Linked Data* e que foram adotadas de maneira unanime ou pela maioria dos estudos recuperados na revisão de literatura.

O modelo levou em consideração 6 dimensões de qualidade: *interlinking*, consistência, completude, licenciamento, avaliação temporal, precisão sintática e semântica.

Um aspecto relevante em relação a esse modelo é a apresentação de fórmulas para o cálculo dos índices locais (cálculo da média de qualidade de uma mesma dimensão) e gerais da qualidade, apontando ainda formas para estabelecer pesos para as métricas.

O modelo também divide as dimensões entre as que possuem aspectos quantitativos e qualitativos, como apresentado na figura 42.

Figura 42 - Relação de dimensões quantitativas e qualitativas



Fonte: Melo (2017, p. 71)

Nesse modelo a avaliação das métricas pode ser baseada na adoção de listas de verificação, que estabelece o que precisa ser verificado nos conjuntos de dados para que a qualidade possa ser mensurada.

O modelo destaca, na avaliação da dimensão consistência, a importância do conhecimento do avaliador em relação as propriedades e vocabulários adotados pelo domínio em questão.

Em relação a dimensão precisão sintática, o modelo indica o uso de padrões de dados bem estabelecidos, como DOI e ISSN, para serem utilizados como padrões-ouro para a checagem de consistência em relação ao tipo de dados.

No modelo os vocabulários são utilizados para a verificação da qualidade semântica e sintática dos dados e para a checagem da completude dos dados em relação ao seu esquema, permitindo avaliar aspectos como completude de propriedades e de população.

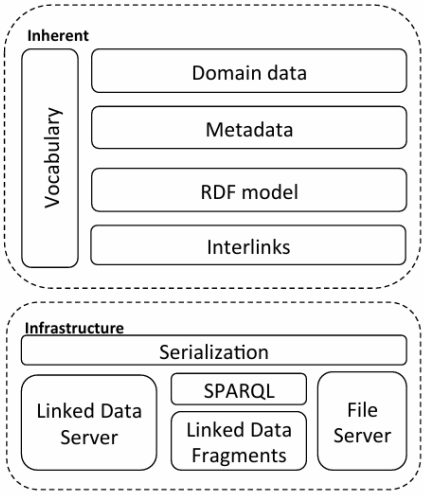
Também é discutido no modelo a presença de metadados como um indicativo de boa qualidade, ao avaliar a completude. A verificação desse aspecto é baseada em um modelo de níveis de Completude de Metadados.

6.3.8 Modelo de qualidade proposto por Radulovic *et al.* (2018)

Radulovic *et al.* (2018) apresentam um modelo bastante extenso e completo, baseado em revisão de literatura.

O modelo adota a nomenclatura e a estrutura proposta pela ISO 25012. Nela, como equivalente às categorias de qualidade têm-se duas grandes classes: aspectos inerentes aos dados e aspectos relacionados com a infraestrutura. Os autores organizam então as questões de qualidade de dados em torno desses dois aspectos. A figura 43 apresenta essa organização.

Figura 43 - Organização das questões de qualidade de dados *Linked Data* em aspectos inerentes e relacionados com a infraestrutura



Fonte: Radulovic *et al.* (2018)

Uma versão completa do modelo de qualidade deveria estar disponível via WIKI, entretanto o link disponibilizado não está em funcionamento, não sendo possível obter acesso ao modelo.

No artigo os autores apresentam como exemplo uma versão simplificada do modelo, sendo discutidos apenas aspectos chave para a sua compreensão, não sendo apresentadas as métricas, fórmulas e escalas necessárias para a efetiva aplicação do modelo.

6.3.9 Modelo de qualidade proposto por Ibanez *et al.* (2019)

Ibanez *et al.* (2019) estabelecem um modelo de qualidade simplificado para a análise do catálogo de dados do *European Data Portal* (EDP). A proposta é o estabelecimento de um modelo simplificado, organizado em sete dimensões, distribuídas nas categorias contextual, representacional e acessibilidade.

O modelo não apresenta critérios de qualidade e nem métricas formais, apenas explicações contextualizando como foram feitas as análises de cada dimensão.

Mesmo sendo um modelo simplificado, duas das dimensões são relacionadas a vocabulários, sendo elas uso de vocabulários conhecidos e uso de vocabulários pouco conhecidos. Em ambos os casos se considera o uso de vocabulários estabelecidos no domínio como um indicativo de boa qualidade.

O modelo explora ainda o DCAT para viabilizar a contagem do número de conjuntos de dados cadastrados no catálogo e como um caminho para verificar informações de proveniência, que nesse modelo são exploradas na dimensão contextual.

O modelo também avalia a presença de *links* para conjuntos de dados externos, sendo a interligação considerada um indicativo de boa qualidade.

6.3.10 Modelo de qualidade proposto por Candela *et al.* (2020)

O modelo proposto por Candela *et al.* (2020) foi adaptado para o contexto dos dados bibliográficos, produzidos no âmbito das bibliotecas digitais, utilizando como base parte da estrutura proposta por Farber (2016), o que faz com que os modelos guardem muitas semelhanças, embora tenham sido feitas adaptações para atender ao domínio proposto.

O modelo é organizado com base nas dimensões de Wang e Strong, apresenta 11 dimensões e 34 critérios de qualidade, para os quais são estabelecidas métricas em formato de fórmulas e escalas, bem como as explicações necessárias para a sua aplicação.

É possível observar no modelo o uso combinado de métricas quantitativas e qualitativas, tendo um foco em métricas de qualidade objetivas.

Em relação as formas de avaliação em relação a critérios e métricas, a avaliação muitas vezes se baseia no uso de padrão-ouro, de modelos de dados ou metadados que se estabelece o que é esperado dos dados em determinada situação avaliativa. Nesse modelo os padrões-ouro foram baseados em padrões do domínio bibliográfico, indicando quais as propriedades e classes mínimas e o tipo de relação esperado entre as propriedades. Observa-se que a adoção de padrões-ouro permite tornar mais objetiva e granular a avaliação em situações que poderiam ser baseadas apenas em análise subjetiva ou ainda binária e pouco representativa da avaliação.

O modelo discute o provisionamento de metadados do conjunto de dados, entretanto foca apenas identificar se existem ou não metadados legíveis por máquina, não apresentado uma forma de avaliar a completude desses metadados. Esse aspecto seria relevante, pois os metadados são amplamente explorados no modelo para obtenção de informações como frequência de atualização e licença.

Os aspectos de confiabilidade também são bastante explorados no modelo. A confiabilidade é avaliada pela forma de obtenção dos dados (com ou sem curadoria manual, provenientes de fontes estruturadas ou não estruturadas), e pela verificação da proveniência, por meio do vocabulário PROV ou do *DublinCore*.

Os vocabulários são explorados ainda para avaliar a adequação as restrições estruturais das propriedades e em relação a completude dos dados.

6.3.11 Modelo de qualidade proposto por Issa *et al.* (2021)

ISSA *et al.* (2021) apresentam um modelo focado apenas na dimensão completude, elaborado com base na condução de uma Revisão Sistemática da Literatura. Os autores discutem o aspecto subjetivo e contextual da completude apontando o desafio de se estabelecer níveis mínimos de completude. A completude:

> Pode ser medida como a porcentagem de dados disponíveis dividida pelos dados necessários, onde 100% é o melhor valor. A questão, no entanto, é se podemos considerar dados com 70% de completude como de alta qualidade? Essa quantidade de informações pode ser suficiente, por exemplo, para a descrição de um filme, mas não para um caso de uso médico.

O modelo divide a completude em 7 tipos: completude de população, completude de atualização, completude de esquema, completude de metadados,

completude de propriedades e completude de rotulagem. No modelo não são apresentadas fórmulas ou escalas para avaliar a qualidade, mas uma explicação sobre cada uma delas.

Como já apontado na discussão dos demais modelos, a análise da qualidade em relação a sua completude é fortemente dependente da exploração dos vocabulários, por isso os vocabulários são essenciais para a avaliação de todo o modelo proposto.

A análise dos vocabulários perpassa a compreensão a respeito da sua estrutura e de suas propriedades, visando identificar que tipos de dados são esperados como valores nas declarações que contêm as propriedades e quais são as propriedades e os valores obrigatórios para determinado contexto.

O modelo apresenta ainda um critério para avaliação de completude de metadados, onde se verifica a existência de metadados, a contagem de valores de metadados e proporção de valores ausentes em relação ao total de propriedades. Não é apresentado um modelo ou padrão ouro para a avalição da completude de metadados.

Apresentados os modelos de qualidade para dados *Linked Data*, a próxima subseção apresenta uma síntese desses modelos.

6.3.12 Síntese dos modelos discutidos

Na presente seção foram discutidos 11 modelos de qualidade, sendo 10 deles criados especificamente para dados *Linked Data* e um modelo de aspecto geral embasado na norma ISO de qualidade de dados. O quadro 44 apresenta um resumo dos modelos apresentados

Figura 44 - Resumo dos modelos de qualidade de dados apresentados

Autores	**Descrição**
ISOIEC25012 (2008)	É um modelo geral de qualidade cujo escopo são “dados retidos em um formato estruturado dentro de um sistema de computador”. O modelo é dividido em três categorias (inerente, inerente e dependente do sistema, e dependente do sistema), nas quais são categorizadas 15 características de qualidade que equivalem a dimensões em outros modelos. São apresentadas definições para as características, e em alguns

	casos exemplos de como essa poderia ser avaliada, não se aprofundando em estabelecer métricas para cada aspecto.
Zaveri et al. (2012)	O modelo foi criado com base em revisão sistemática da literatura. Em alguns critérios foram sugeridas metodologias para avaliação, tais como: *crowdsourcing*, uso de listas de verificação, consulta a lista de reputação e lista-negra de dados, avaliação manual por especialistas e análise probabilística.
Behkamal et al. (2014)	É um modelo focado em apresentar métricas para a avaliação da acurácia de dados *Linked Data*, dividido entre acurácia semântica (6 critérios) e acurácia sintática (9 critérios). Focado em publicadores de dados, visando a avaliação para a melhoria da qualidade, a ser aplicado antes da publicação dos dados. O modelo criado com base na metodologia Goad-Question-Metric (GQM).
Debattista et al. (2015)	É um modelo simplificado de avaliação de qualidade que combina métricas tradicionais com técnicas de análise probabilística para dados *Big Data*. Propõe a avaliação de quatro critérios aos quais as técnicas podem ser aplicadas.
Farber et al. (2016)	Modelo pensado para avaliação de dados *Linked Data* de conhecimentos gerais (domínio cruzado ou enciclopédico), publicados em formato aberto. Se organiza em dimensões, critérios e métricas, seguindo as categorias propostas por Wang e Strong. Apresenta escalas ou fórmulas para o cálculo de todos os critérios, além de informações complementares para auxiliar na avaliação.
Cappiello et al. (2016)	Modelo de qualidade minimalista, elaborando para embasar a etapa exploração de dados *Linked Data*, necessária para identificar conjuntos de dados de interesse, que posteriormente seriam submetidos a um processo de avaliação mais rigoroso. O modelo é composto por 5 dimensões (quantidade de dados, concisão, completude, navegabilidade e interligação), que não são organizadas em categorias e critérios.
Melo (2017)	Modelo interdisciplinar, baseado na literatura, nos princípios do *Linked Data* e nas recomendações do projeto *Linked Open Data*. É organizado em dimensões e métricas, sendo consideradas apenas as dimensões diretamente relacionadas com os princípios e com frequência significativa na literatura. Apresenta fórmulas para o cálculo dos índices locais e globais de qualidade.
Radulovic (2018)	Modelo extenso e abrangente organizado de acordo com a estrutura proposta na ISO, que

	reúne, baseado em revisão de literatura, dimensão e métricas aplicáveis a avaliação de qualidade de dados *Linked Data*. Não foi possível obter acesso ao modelo completo que contém as informações necessárias para a sua aplicação.
Ibanez et al. (2019)	Modelo de qualidade simplificado para a análise do catálogo de dados do *European Data Portal* (EDP). Apresenta sete dimensões, distribuídas nas categorias contextual, representacional e acessibilidade. Não apresenta critérios e nem fórmulas ou escalas para avaliação, apenas explicações contextualizando como foram feitas as análises de cada dimensão.
Candela et al. (2020)	Modelo elaborado para avaliação de dados bibliográficos, tendo sido estabelecido com base no modelo proposto por Farber *et al.* (2016), com adaptações em relação a forma de avaliação dos critérios e métricas. Fortemente baseado na exploração de metadados e no estabelecimento de padrões-ouro.
Issa et al. (2021)	Modelo de qualidade focado na dimensão completude, construído com base em Revisão Sistemática da Literatura. O modelo explora amplamente o uso de vocabulários para avaliação da qualidade.

Fonte: Autora (2025)

Como é possível observar, nenhum dos modelos identificados atendem completamente aos objetivos estabelecidos para a essa tese, considerando que não tem como foco auxiliar especificamente no processo de seleção, sendo o mais próximo desse objetivo o modelo proposto por Cappiello *et al.* (2016).

Para além de identificar e discutir os modelos existentes, a presente subseção teve como um dos seus propósitos analisar como são construídos e estruturados esses modelos de qualidade e como esses são adaptados para o contexto do *Linked Data*.

Com base em sua análise é possível observar que com frequência esses modelos são construídos com base em revisão de literatura. Parte dos modelos possuem uma abordagem generalista, focados em cobrir todos os aspectos de qualidade, e, em especial, foram pensados para auxiliar no processo de identificação e correção de erros, ou ainda no ranqueamento da qualidade dos conjuntos de dados, fornecendo métricas para o cálculo geral da qualidade que permita mensurar a qualidade e evolução de dados disponíveis em determinados catálogos de dados ou provenientes de domínios específicos.

Observa-se que a adoção ou não de uma abordagem geral na avaliação de qualidade é um aspecto importante do processo decisório de avaliação de qualidade, pois ao buscar cobrir todas as dimensões e analisar todos os aspectos possíveis da qualidade, tem-se uma visão abrangente, porém genérica da qualidade de desse conjunto de dados.

Quando se aborda a avaliação de qualidade para seleção de dados, adotar uma abordagem geral poderá levar a um processo de avaliação excessivamente complexo, que pode resultar em uma série de resultados que embora reveladores, não terão real impacto na seleção dos conjuntos de dados.

Embora de maneira isolada os modelos apresentados não possam ser aplicados para solucionar o problema de pesquisa, sua discussão e análise permitiu compreender qual a estrutura de um modelo de dados, identificar quais são as métricas disponíveis para o contexto do *Linked Data* e como essas podem ser adaptadas para atender a diferentes contextos.

Compreendendo a importância e a complexidade do processo de construção e estabelecimento de modelos de qualidade, a próxima subseção discute as abordagens que podem ser adotadas para a composição e aplicação dos modelos de qualidade.

6. 4 Construção e aplicação de modelos de qualidade de dados

Uma das etapas fundamentais para a avaliação de qualidade é elaboração e/ou seleção do modelo de qualidade que será utilizado para guiar o processo de avaliação e a identificação de instrumentos e ferramentas necessários para auxiliar na aplicação do modelo proposto.

Na avaliação de qualidade realizada em uma perspectiva contextual, um dos principais desafios é estabelecer as dimensões e métricas que irão compor o modelo de qualidade. Busca-se a composição de um modelo equilibrado, que não se torne desnecessariamente complexo a ponto de dificultar o processo de avaliação, nem demasiadamente simplificado, a ponto de fornecer resultados que não permitam atender aos objetivos que levaram a condução da avaliação.

No âmbito da seleção de fontes para criação de *links*, têm se ainda o desafio de estabelecer os parâmetros de qualidade esperados. Torna-se necessário identificar

os requisitos mínimos de qualidade aceitáveis e os critérios de exclusão que irão orientar o processo de seleção.

Considerando os aspectos apresentados, a presente subseção busca discutir as diferentes abordagens que auxiliam na construção de modelos de qualidade, bem como os instrumentos que podem auxiliar na aplicação desses modelos.

Ao discutir os modelos de qualidade de dados, Batini e Scannapieco (2016), apontam que a criação de modelos abrangentes de qualidade de dados pode ser baseada em 3 principais abordagens: teórica, empírica e intuitiva.

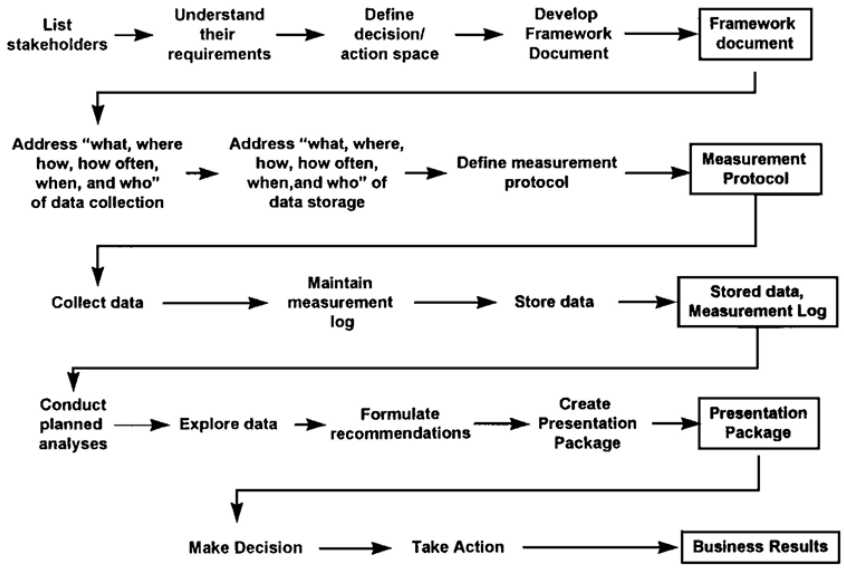
> A abordagem teórica adota um modelo formal para definir ou justificar as dimensões. A abordagem empírica constrói o conjunto de dimensões a partir de experimentos, entrevistas e questionários. A abordagem intuitiva simplesmente define dimensões de acordo com o bom senso e a experiência prática (Batini; Scannapieco, 2016, p. 37).

A abordagem teórica parte dos princípios de que os dados devem ser uma representação fiel dos objetos do mundo real que representam, e, portanto, o modelo de qualidade pode possuir uma base ontológica ou de modelagem de domínio.

A abordagem empírica, baseia a construção do modelo na participação direta do usuário no processo de estabelecimento das dimensões e métricas. Um exemplo de abordagem empírica é o *framework* proposto por Wang e Strong (1996), que foi construído com base na aplicação de entrevistas e experimentos com consumidores, que auxiliariam na seleção e categorização das dimensões relevantes.

Um fluxo para criação e aplicação de modelos de qualidade baseados em abordagem empírica é apresentado por Juran e Godfrey (1998). A figura 45 apresenta o fluxo mencionado.

Figura 45 - Fluxo da criação e aplicação de modelo de qualidade



Fonte: Juran e Godfrey (1998)

Como é possível observar na figura 45, a construção do modelo com abordagem empírica é pautada no conhecimento aprofundado tanto dos dados que serão objetos da avaliação, como pela análise de como, por quem e quando será conduzido o processo de avaliação.

Em relação às abordagens intuitivas, essas podem partir da experiencia dos consumidores com os dados do contexto em questão ou ainda do perfilamento de fontes, como catálogos de dados, relacionados ao domínio. Os dados podem ser explorados visando a identificação de problemas frequentes, e cada problema pode ser convertido em uma dimensão ou métrica de qualidade.

Os métodos intuitivos podem partir ainda da análise exploratória dos objetos e objetivos do processo de avaliação, onde se busca relacionar esses objetivos a dimensões e métricas.

Como observado nos modelos de qualidade, a análise dos vocabulários e da aplicação de suas propriedades demonstra-se fundamental para a análise exploratória dos dados no contexto do *Linked Data*. Os vocabulários permitem identificar inconsistências, más formações e níveis de ligação, como no caso da OWL e do RDFs, fornecem informações relevantes para o processo de avaliação, como no caso de vocabulários como o PROV, Dublin Core, SKOS e DCAT. A própria presença/ausência de vocabulários bem estabelecidos no domínio dos dados pode ser um indicativo importante dos seus níveis de qualidade e relevância.

A análise dos modelos de qualidade de dados *Linked Data* permitiu identificar outras abordagens, sendo elas: abordagem baseada em garantia literária; abordagem baseada em adequação a normas e melhores práticas; abordagem baseada na adaptação de metodologias da gestão de qualidade e abordagens mistas.

As abordagens baseadas na literatura partem de levantamentos bibliográficos para identificar dimensões e métricas, criando modelos abrangentes ou direcionados para um domínio específico. Os levantamentos também podem ser relevantes quando se busca analisar a qualidade de dados em relação uma categoria, dimensão ou métrica específica, permitindo identificar as dimensões e métricas aplicáveis a esse aspecto específico da qualidade. O modelo construído Zaveri *et al.* (2012) é um exemplo de modelo construído seguindo essa abordagem.

A abordagem baseada em adequação à normas e melhores práticas leva em consideração princípios, recomendações e normas que se aplicam ao domínio em questão. No contexto do *Linked Data*, podem ser consideradas as 35 melhores práticas para a publicação de dados na *Web*, as melhores práticas para publicação de dados *Linked Data*, os documentos e recomendações relacionados a criação de dados RDF, de bons URIs, para estruturação e aplicação de vocabulários e ontologias. São exemplos de modelos baseados em normas e melhores práticas os modelos propostos por Radulovic (2018) e Melo (2017).

Os modelos também podem ter como base uma metodologia que auxilie na seleção das dimensões e métricas pertinentes para determinado contexto, como no caso do modelo de Behkamal *et al.* (2014). O modelo proposto por Behkamal *et al.* (2014) foi construído com base na metodologia Goad-Question-Metric (GQM). “No GQM, os objetivos são gradualmente refinados em várias perguntas, cada pergunta é então refinada em métricas. Além disso, uma métrica pode ser usada para responder a várias perguntas” (Behkamal *et al.*, 2014, p. 81, tradução nossa).

A respeito dos propósitos e da origem da metodologia, Basili, Caldiera e Rombach (1994, p.2, tradução nossa) apontam que ela:

> baseia-se na premissa de que, para uma organização medir de forma proposital, ela deve primeiro especificar as metas para si mesma e para seus projetos, em seguida, deve rastrear essas metas até os dados que se destinam a defini-las operacionalmente e, por fim, fornece uma estrutura para interpretar os dados em relação às metas estabelecidas. Portanto, é importante deixar claro, pelo menos em termos gerais, quais são as necessidades informacionais da

> organização, para que essas necessidades de informação possam ser quantificadas sempre que possível, e as informações quantificadas possam ser analisadas para determinar se as metas foram alcançadas ou não. A abordagem foi originalmente definida para avaliar defeitos em um conjunto de projetos no ambiente do Centro de Voos Espaciais Goddard da NASA.

A metodologia pode ser dividia em 3 níveis: conceitual, operacional e quantitativo. No nível conceitual se estabelece o objetivo de qualidade; no nível operacional se estabelece um conjunto de perguntas que “tentam caracterizar o objeto de medição (produto, processo, recurso) com relação a um problema de qualidade selecionado e determinar sua qualidade a partir do ponto de vista selecionado” (Basili, Caldiera e Rombach, 1994, p. 2, tradução nossa).

No nível quantitativo, as questões são então convertidas em métricas, que podem ser objetivas ou subjetivas, que permitam mensurar os níveis de qualidade. Quando aplicado ao contexto da qualidade de dados, trata-se, portanto, da sistematização de um processo de tradução das necessidades informacionais em métricas mensuráveis de qualidade. O resultado da aplicação dessa metodologia é um modelo hierárquico onde um objetivo se relaciona a uma ou mais questões, que podem ser relacionadas a uma ou mais métricas.

Ao longo das discussões de qualidade de dados, algumas questões relacionadas a coleta de dados e avaliação de qualidade foram identificadas.

Juran e Godfrey (1998) apresentam questões norteadoras, para guiar os especialistas que devem ajudar na construção do modelo. São estabelecidas as perguntas “O quê”; “Onde”; “Quando”; “Como”; “Com que frequência” e “Quem” que podem ser direcionadas a avaliar como serão coletado, armazenados e analisados os dados. Os autores apontam que mesmo que os dados sejam de fontes conhecidas:

> [...] O planejador ainda é aconselhado a aprender sobre as complexidades da coleta de dados, pois dados coletados para uma finalidade podem não ser adequados para outra. Em seguida, o planejador especifica como, quando e com que frequência as medições devem ser feitas. Cada um deve ser detalhado. "Como" envolve não apenas como uma medição específica deve ser feita, mas também como o equipamento de medição deve ser calibrado e mantido e como os dados precisos devem ser obtidos. "Quando" e "com que frequência" devem ser abordados para garantir que dados amplos estejam disponíveis (Juran; Godfrey, 1998, p. 4.14, tradução nossa).

Sant'Ana (2016, p. 119, grifo nosso) apresenta uma série de questões norteadoras para a etapa de coleta, que podem também ser adaptadas para o processo de seleção de fontes de dados, sendo elas:

> **Qual é o escopo da necessidade informacional? Que tipo de resultado se espera? Com quais características? Quais são os dados necessários? Onde estão as fontes para estes dados? Como os dados podem ser coletados?** Em que formato estão? Quais são os tratamentos necessários para que fiquem adequados ao que se precisa? A coleta destes dados não proporciona risco de privacidade para os indivíduos ou entidades referenciadas por eles? Elementos, que em alguns casos poderiam ser considerados como secundários, que permitam a integração entre os diversos dados coletados estão sendo obtidos? **Como avaliar sua integridade física e lógica, além de outros elementos que garantam sua qualidade? Como identificar sua procedência? Têm-se o direito ou permissão de coletar estes dados? Estão sendo coletados dados que permitam que estes venham a ser identificáveis e recuperáveis em um momento futuro?** Estão sendo coletados dados que propiciem a manutenção e acesso a eles no futuro caso venham a ser armazenados? (Sant'Ana, 2016, p. 119, grifo nosso).

Analisando as questões apresentadas, observa-se que elas podem ser divididas em dois aspectos: questões relacionadas com o contexto de uso dos dados e questões relacionados com as características dos dados.

As questões relacionadas ao contexto de uso dos dados visam auxiliar na análise prévia do projeto/processo em que serão empregados os dados, identificando as informações necessárias para a busca e recuperação de dados e fontes pertinentes que possam atender às necessidades informacionais dos usuários. Já as questões relacionadas com as características dos dados buscam reunir informações relacionadas a sua estrutura, direitos de uso, disponibilidade e níveis de qualidade intrínsecos.

No contexto da avaliação de qualidade, essas questões podem ser aplicadas para auxiliar no estabelecimento do modelo de qualidade, identificando que aspectos relevantes e qual o peso desses aspectos para os usuários, auxiliando na escolha das dimensões e métricas que irão compor o processo de avaliação.

Ressalta-se ainda a importância dos metadados e das informações de proveniência para obtenção de informações relevantes, considerando que a disponibilização estruturada de informações a respeito dos dados e de seus fornecedores, especialmente em formato legível por máquina, facilita recuperação e

análise dos dados, bem como o uso de ferramentas automática e semiautomáticas para auxiliar nesse processo.

Heath e Bizer (2011), apresentam algumas questões norteadoras pensadas especificamente para seleção de dados para ligação, sendo elas:

> Qual é o valor dos dados no conjunto de dados de destino? Em que medida isso agrega valor ao novo conjunto de dados? O conjunto de dados de destino e seu *namespace* estão sob propriedade estável e manutenção ativa? Os URIs no conjunto de dados são estáveis e provavelmente não sofrerão alterações? Existem *links* contínuos para outros conjuntos de dados para que os aplicativos possam acessar uma rede de fontes de dados interconectadas? (Heath; Bizer, 2011, não paginado, tradução nossa).

As questões para o processo de seleção podem ser diretamente relacionadas a dimensões e critérios de qualidade, como apresenta o quadro 20.

Quadro 20 - Relação entre questões norteadoras e dimensões de qualidade

Questão	Dimensão	Critério
Qual é o valor dos dados no conjunto de dados de destino?	Relevância	Cobertura
Em que medida isso agrega valor ao novo conjunto de dados?	Relevância	Cobertura
O conjunto de dados de destino e seu *namespace* estão sob propriedade estável e manutenção ativa?	Confiabilidade	Confiabilidade do fornecedor de informações
Existem *links* contínuos para outros conjuntos de dados para que os aplicativos possam acessar uma rede de fontes de dados interconectadas?	Interligação	Existência de links para provedores de dados externos

Fonte: Autora (2025)

Outro aspecto relevante identificado nos modelos é o uso combinado de metodologias que auxiliam na aplicação de modelos de qualidade, especialmente nas avaliações de critérios e métricas de abordagem qualitativa e subjetiva. São exemplos desses instrumentos 1) As listas de verificação (*checklists*), melhores práticas e princípios; estabelecimento de padrões-ouro; aplicação de técnicas de análise probabilística e técnicas de *crowdsourcing*.

As listas de verificação podem ter como base normas, modelos, instrumentos de padronização e boas práticas estabelecidos no domínio de origem, como por exemplo os princípios *Findability*, *Accessibility*, I*nteroperability*, *and Reuse of Digital Assets* (FAIR), no contexto dos dados científicos e de pesquisa e o modelo *Resource Description and Access* (RDA) no contexto do domínio bibliográfico.

Os padrões-ouro estabelecem um modelo de dados com o qual os resultados podem ser comparados, esse modelo pode ser utilizado para avaliar aspectos como completude dos dados em relação a quantidade esperada de propriedades, a quantidade esperada de entidades relacionadas a uma propriedade e ao formato dos dados quando se adotam literais como valores.

Compreendendo a relevância dos vocabulários no processo de seleção, do uso combinado de metodologias e ainda a importância das melhores práticas e dos padrões-ouro para a avaliação da qualidade, as próximas subseções se aprofundam, respectivamente, nesses aspectos e em sua relação com a qualidade de dados.

6.5 Vocabulários no processo de avaliação de qualidade

Com a realização dos estudos teóricos e terminológicos a respeito da qualidade de dados *Linked Data* foi possível observar a importância dos vocabulários para o processo de avaliação de qualidade.

Embora a literatura reconheça amplamente o papel que os vocabulários exercem no processo de avaliação de dados *Linked Data*, na maior parte dos estudos os vocabulários são discutidos de maneira secundária, geralmente em uma perspectiva de aplicação, não existindo um grande aprofundamento a respeito dessa relação.

Esses vocabulários, ontologias e esquemas são a base para a estrutura dos dados publicados em RDF, e por isso influenciam na avaliação sintática e semântica da qualidade dos dados.

A análise dos modelos de qualidade ressaltou a importância e o impacto dos vocabulários no processo de avaliação, uma vez que são responsáveis, nesse contexto, por atuarem como padrões de metadados, que permitem a descrição formal dos conjuntos de dados, garantindo o acesso a informações indispensáveis para o processo de avaliação de qualidade, como licença, proveniência e conteúdo.

Nesse contexto, os vocabulários podem ser explorados para avaliação utilizando ferramentas automáticas e semiautomáticas, especialmente no que concerne aos aspectos de acessibilidade e confiabilidade dos conjuntos de dados, dois aspectos essenciais para o processo de seleção de fontes de dados *Linked Data.*

Os vocabulários se destacam ainda por sua aplicabilidade no registro e compartilhamento dos resultados do processo de avaliação, permitindo que os publicadores compartilhem junto com os dados, análises relacionadas a sua qualidade. Permitem ainda que consumidores compartilhem resultados de processos de avaliação. Esse tipo de vocabulário permite que as ferramentas de avaliação exportem os resultados em um formato legível por máquina. Os resultados descritos nesse formato também podem ser consumidos pelas ferramentas, gerando assim um ciclo virtuoso.

A análise do processo de interligação de dados permitiu ainda ressaltar a importância dos vocabulários para a criação de *links*, uma vez que esses vocabulários permitem explicitar o tipo de relação existente entre as entidades.

Com base nas análises realizadas, destacaram-se alguns vocabulários por seu impacto nos processos de avaliação, seleção de fontes, interligação e representação de níveis de qualidade.

A presente subseção busca apresentar os vocabulários explorados para a avaliação de qualidade pelos modelos apresentados na subseção 6.2, sendo selecionados apenas vocabulários que podem ser diretamente explorados pelo processo de Avaliação de Qualidade De Dados.

Os vocabulários identificados foram discutidos visando abordar o seu papel em relação aos seguintes aspectos: 1) Vocabulários relacionados a estruturação e interligação dos dados; 2) Vocabulários relacionados ao fornecimento de metadados; e 3) Vocabulários para o compartilhamento dos resultados do processo de avaliação. O quadro 21 apresenta uma lista dos vocabulários a serem discutidos.

Quadro 21 - Relação de vocabulários e aspectos discutidos

Vocabulário	Aspecto discutido
Simple Knowledge Organization System **(SKOS)**	Estrutura e interligação
Resource Description Framework Schema **(RDFs)**	Estrutura e interligação
Web Ontology Language **(OWL)**	Estrutura e interligação
Data Catalog Vocabulary **(DCAT)**	Fornecimento de metadados

***Vocabulary of Interlinked Datasets* (VOID)**	Fornecimento de metadados
***PROV Ontology* (PROV)**	Fornecimento de metadados
***Dublin Core* (DC)**	Fornecimento de metadados
***Data Quality Vocabulary* (DQV)**	Descrição da qualidade

Fonte: Autora (2025)

Os vocabulários para discussão foram selecionados com base na sua ocorrência nos modelos de qualidade de dados *Linked Data*, com exceção do modelo DQV, que foi incluído por sua relação intrínseca com a qualidade de dados *Linked Data*.

As próximas subseções discutem diversos vocabulários que podem ser diretamente relacionados ao processo de avaliação e foram organizadas em relação aos três principais aspectos identificados.

6.5.1 Vocabulários relacionados a interligação e estruturação de dados

Considerando que a criação dos *links* se baseia na identificação dos tipos de *link* possíveis e na seleção de propriedades que apresentem adequadamente essa relação, buscou-se relacionar os tipos de *link* mais comuns e os vocabulários/propriedades mais utilizados para representar esse tipo de conexão. McKenna, Debruyne e O'Sullivan (2022) apresentam uma lista dos possíveis tipos de *links* no *Linked Data*:

- **É idêntico a -** Ocorre quando a URI de uma entidade interna representa exatamente a mesma entidade da URI de uma fonte externa;
- **É idêntico em certo contexto a -** Ocorre quando, em determinado contexto, URIs de duas fontes representam a mesma entidade;
- **É quase idêntico a -** Ocorre quando URIs de duas fontes não representam exatamente a mesma entidade, mas entidades com características muito semelhantes e com pequenas variações de propriedades;
- **É similar a -** Ocorre quando URIs de duas fontes não representam exatamente a mesma entidade, mas representam entidades com características semelhantes;
- **É associado com -** Ocorre quando URIs de duas fontes possuem algum tipo de relação, que não a de semelhança. Pode ser usado para relações do tipo todo - parte.

- **É diferente de -** Ocorre quando é necessário indicar que duas URIs não representam uma mesma entidade.

Embora existam outros vocabulários que possam ser utilizados para esse propósito, para representar relações que indicam semelhança, igualde e diferença, destacam-se na literatura a ontologia OWL, o esquema RDFs e o vocabulário SKOS[13].

O RDF-s é considerado uma extensão do vocabulário básico do RDF, composto por classes e propriedades utilizadas para estruturar vocabulários RDF (W3C, 2023). Ele desempenha um papel importante na estruturação de outros vocabulários, definindo a hierarquia desses vocabulários e estabelecendo características importantes como o alcance das propriedades. Também desempenha um papel importante na ligação, principalmente na criação de *links* de vocabulário.

A OWL pode ser usada para representar explicitamente o significado de termos em vocabulários e as relações entre esses termos" (W3C, 2009, não paginado, tradução nossa). Ela amplia as possibilidades de estruturação, indo além da semântica básica fornecida pelo RDF-s. Em algumas situações a OWL reutiliza termos do RDF e do RDFs, essa reutilização pode ser identificada pela utilização prefixos rdf e rdfs.

O *Simple Knowledge Organization System* (SKOS) se caracteriza como uma:

> [...] linguagem leve e intuitiva para desenvolver e compartilhar novos sistemas de organização do conhecimento. Pode ser usado isoladamente ou em combinação com linguagens formais de representação do conhecimento, como a linguagem de ontologia da *Web* (W3C, 2009, não paginado, tradução nossa).

Quando comparado a OWL e RDFs, o SKOS é considerado um vocabulário mais simples, em razão a aspectos formais. Foi elaborado com o propósito de possibilitar a publicação de Sistemas de Organização do Conhecimento (SOCs) em um formato legível por máquina. O SKOS pode ser usado em conjunto com o RDFs e a OWL.

A documentação da OWL divide seus termos com base na função que devem exercer, sendo relevantes para interligação os termos que estruturam categorias e

[13] Em relação a nomenclatura adotada, com base no estudo terminológico realizado, todos esses instrumentos podem ser adequadamente denominados como vocabulários, sendo considerados de tipos diferentes. Nesse sentido, optou-se por utilizar a nomenclatura adotada pela documentação para se referir ao tipo específico de cada documento, sendo utilizado o termo vocabulários para se referir a eles enquanto um conjunto.

classes provendo a hierarquização e os termos que representam relações de equivalência, de igualdade e de desigualdade (W3C, 2009).

O quadro 22 apresenta as propriedades que podem ser utilizadas como predicados para a criação de *links* entre os recursos, seguindo a estrutura proposta na OWL.

Quadro 22 - Relação de propriedades e suas funções na criação de *links*

Termo	Descrição	Função	Tipo de relacionamento
rdfs:Resource	Indicar um recurso descrito em RDF (W3C, 2023).	Hierarquia	É associado com
owl:Individual	Indicar instancias de classes e propriedades (W3C, 2009)	Hierarquia	É associado com
rdfs:Classe owl:Class	Indicar uma classe (W3C, 2023).	Hierarquia	É associado com
owl:Class	Indicar uma classe (W3C, 2009)	Hierarquia	É associado com
rdfs:subClassOf	Indicar uma subclasse (W3C, 2023).	Hierarquia	É associado com
rdf:Property	Indicar uma propriedade (W3C, 2023).	Hierarquia	É associado com
rdfs:subPropertyOf	Indicar uma sub propriedade (W3C, 2023).	Hierarquia	É associado com
skos:related	Expressa a relação não hierárquica entre dois conceitos (W3C, 2009).	Hierarquia	É associado com
rdfs:seeAlso	Indicar um recurso que acrescenta informações sobre a entidade (W3C, 2023).	Relação	É associado com
owl:equivalentClass	Permite indicar que duas classes são equivalentes e possuem o mesmo conjunto de indivíduos (W3C, 2009).	Igualdade e desigualdade	É idêntico a
owl:equivalentProperty	Indicar duas propriedades	Igualdade e desigualdade	É idêntico a

	possuem as mesmas entidades (mesmo que tenham significados diferentes) (W3C, 2009).		
owl:sameAs	Indicar que dois URIs se referem a mesma entidade (W3C, 2009).	Igualdade e desigualdade	É idêntico a
owl:differentFrom	Indicar que dois URIs não se referem a mesma entidade (W3C, 2009).	Igualdade e desigualdade	É diferente de
owl:AllDifferent	Indicar que uma lista de indivíduos é diferente (W3C, 2009).	Igualdade e desigualdade	É diferente de
skos:broadMatch	Indicar a relação hierárquica entre dois conceitos, onde o primeiro é mais abrangente que o segundo (W3C, 2009).	Igualdade e desigualdade	É associado com
skos:narrowMatch	Indicar a relação hierárquica entre dois conceitos, onde o primeiro é mais específico que o segundo (W3C, 2009).	Igualdade e desigualdade	É associado com
skos:closeMatch	Indicar que dois conceitos são similares, mas não equivalentes (W3C, 2009).	Igualdade e desigualdade	É quase idêntico a
skos:mappingRelation	Indicar a relação entre conceitos diferentes que possuem semelhança em determinado contexto (W3C, 2009).	Igualdade e desigualdade	É idêntico em certo contexto a
skos:relatedMatch	Indicar a relação entre conceitos diferentes, onde o primeiro permite encontrar o	Igualdade e desigualdade	É associado com

	segundo (W3C, 2009).		
skos:broader	Indicar a relação hierárquica entre dois conceitos onde o primeiro engloba o segundo (W3C, 2009).	Igualdade e desigualdade	É associado com

Fonte: Autora (2025)

Esses vocabulários que permitem a interligação podem ser explorados na avaliação de qualidade para mensurar os níveis de interligação dos conjuntos de dados e para a identificação de novas fontes para exploração.

Nesse conjunto, *sameAs* é o termo mais explorado, aparecendo na maioria dos modelos de qualidade. Neles o termo é aplicado para avaliar principalmente o nível de interligação do conjunto de dados com dados externos, (Zaveri *et al.*, 2012; Färber *et al.*, 2016; Cappiello *et al.*, 2016; Candela *et al.*, 2020).

Ainda em relação a dimensão interligação, a owl:sameAs também pode ser explorada para verificar a validade dos *links* externos (Färber *et al.*, 2016; Candela *et al.* 2020) e para a detecção de *links* internos de boa qualidade (Zaveri *et al.*,2012).

Os termos relacionado a apresentar relações de igualdade e desigualdade podem ser explorados principalmente para avaliação dos dados em relação a dimensão interligação, permitindo verificar o quão conectados com outras fontes os conjuntos de dados estão e qual o nível de qualidade e funcionalidade dos *links* com fontes externas. Permitem ainda a avaliação da interoperabilidade com vocabulários externos, por meio da identificação da presença de relações de equivalência entre classes e propriedades.

Ainda em relação a OWL, destacam-se para a avaliação de qualidade outras propriedades que não são usadas para interligação, mas para estruturação das declarações. As propriedades *owl:ObjectProperty* e *owl:DatatypeProperty*, têm como função estabelecer o que se espera receber no campo valor quando determinada propriedade é aplicada, entretanto, o termo *ObjectProperty* concentra-se em declarações onde o valor é um outro individuo, ou seja é representado por um URI, enquanto o *DatatypeProperty* concentra-se em declarações que possuem como valor um literal.

Essas propriedades podem ser aplicadas na avaliação da consistência do conjunto de dados, ao verificar se as restrições para o tipo de valor esperado da propriedade foram respeitadas (Zaveri *et al.*, 2012).

Färber *et al.* (2016), acrescente ainda como critério de qualidade verificar se, na política de aquisição de dados do conjunto de dados, são realizadas checagens dessas restrições, garantindo que não existam problemas de consistência futuros.

Nesse sentido, observa-se que os termos criados para declarar características das propriedades podem ser utilizados principalmente na avaliação da consistência dos dados, verificando se as propriedades dos vocabulários estão sendo corretamente empregadas.

Também foi identificado nos modelos o uso de termos relacionados ao controle de versão, como *DeprecatedClass* e *DeprecatedProperty*, utilizados para identificar tanto classes como propriedades que estão em desuso. Zaveri *et al.* (2012) mencionam o uso dessas propriedades como meio para a avaliação da consistência dos dados, verificando se propriedade e classes em desuso não estão sendo adotadas.

Discutido o uso de vocabulários relacionados a interligação e a estrutura na avaliação de qualidade, a próxima seção discute o uso de vocabulários para o fornecimento de metadados.

6.5.2 Vocabulário no fornecimento de informações e metadados

Para que o processo de avaliação seja possível, primeiro os dados precisam ser recuperados, passando muitas vezes por um processo de análise exploratória dos conjuntos de dados e pela sua pré-avaliação, que atua como filtro de fontes em potencial. Uma vez recuperados, uma série de informações a respeito desses dados são necessárias para a sua avaliação: quem são seus publicadores? Quais são as suas licenças de uso? Que tipos de dados o conjunto contém? Quais são os vocabulários adotados?

Tanto a recuperação dos dados, como a resposta para essas questões, que perpassam a avaliação de distintas dimensões de qualidade, podem ser descritas pelos publicadores, ao fornecerem metadados a respeito do conjunto de dados, sendo os vocabulários importantes para o fornecimento dessas informações.

> Os vocabulários desempenham um papel importante na estrutura de um conjunto de dados, uma vez que um ou mais desses vocabulários descrevem os recursos do conjunto de dados. (Debattista *et al.*, 2018, p. 880, tradução nossa).

Esse papel possui ainda mais destaque quando se trata de dados *Linked Data*, tendo em vista que a própria estrutura desses dados, baseada no modelo RDF, implica no uso de propriedades, formalizadas e descritas em vocabulários.

> Ter metadados como parte de um conjunto de dados publicado é o primeiro passo para colocar um conjunto de dados no mapa de dados abertos (encorajando assim a descoberta), pois geralmente é o primeiro ponto de acesso para os consumidores que desejam usar os dados publicados. Os metadados garantem que eles estejam em conformidade com as melhores práticas, tornando-se auto descritivos. Portanto, 'fazer os metadados corretamente' é essencial para qualquer tipo de publicação de dados abertos (Debattista *et al.*, 2018, p. 862-863, tradução nossa).

Heath e Bizer (2011) ao discutirem a publicação de dados como *Linked Data*, mencionam a relevância do fornecimento de metadado, destacando os metadados de proveniência, de licenciamento e de descrição a nível conjunto de dados.

Ao longo do desenvolvimento do *Linked Data* foram elaborados e estabelecidos um conjunto de vocabulários com proposito de estruturar o fornecimento de metadados, com base na análise da literatura de qualidade de dados, observa-se como relevantes para o fornecimento de metadados os vocabulários: DCAT, VOID, PROV e DC.

O VOID é um vocabulário criado justamente para a disponibilização de metadados em RDF.

> VoID é um vocabulário de esquemas RDF para expressar metadados sobre conjuntos de dados RDF. Ele foi criado para servir como uma ponte entre os publicadores e os usuários de dados RDF, com aplicações que vão desde a descoberta de dados até a catalogação e arquivamento de conjuntos de dados (W3C, 2011, não paginado, tradução nossa).

De acordo com o W3C (2011) os metadados no VOID são divididos em 4 categorias principais, sendo elas: metadados gerais, metadados de acesso, metadados estruturais e metadados de descrição de *links* entre conjuntos de dados.

Os metadados gerais são relevantes para o processo de avaliação de qualidade pois:

> ajudam potenciais usuários de um conjunto de dados a decidir se ele é apropriado para seus propósitos. Eles incluem informações como título e descrição, a licença do conjunto de dados e informações sobre seu assunto" (W3C, 2011, não paginado, tradução nossa).

Os metadados gerais no VOID reaproveitam propriedades do Dublin core. "O Dublin Core", também conhecido como Conjunto de Elementos de Metadados Dublin Core, é um conjunto de quinze elementos "centrais" (propriedades) para descrever recursos" (Dublin Core, 2025, tradução nossa).

W3C (2011) sistematiza as propriedades utilizadas do DC para o fornecimento de metadados gerais em RDF, o quadro 23 apresenta essa sistematização

Quadro 23 - Propriedades do DC utilizadas no VOID para descrição de metadados gerais em RDF

Propriedades para descrição de metadados gerais	
cterms:title	O nome do conjunto de dados.
dcterms:description	Uma descrição textual do conjunto de dados.
dcterms:creator	Uma entidade, como uma pessoa, organização ou serviço, que é a principal responsável pela criação do conjunto de dados. O criador deve ser descrito como um recurso RDF, em vez de apenas fornecer o nome literal.
dcterms:publisher	Uma entidade, como uma pessoa, organização ou serviço, responsável por disponibilizar o conjunto de dados. O editor deve ser descrito como um recurso RDF, em vez de apenas fornecer o nome literal.
dcterms:contributor	Uma entidade, como uma pessoa, organização ou serviço, responsável por fazer contribuições ao conjunto de dados. O contribuidor deve ser descrito como um recurso RDF, em vez de apenas fornecer o nome literal.
dcterms:source	Um recurso relacionado do qual o conjunto de dados é derivado. A fonte deve ser descrita como um recurso RDF, e não como um recurso literal.
dcterms:date	Um ponto ou período de tempo associado a um evento no ciclo de vida do recurso. O valor deve ser formatado e tipado como xsd:date.
dcterms:created	Data de criação do conjunto de dados. O valor deve ser formatado e tipado como xsd:date.

dcterms:issued	Data de emissão formal (por exemplo, publicação) do conjunto de dados. O valor deve ser formatado e digitado como *xsd:date*.
dcterms:modified	Data em que o conjunto de dados foi alterado. O valor deve ser formatado e definido como xsd:date.

Fonte: Traduzido de W3C (2011)

Os metadados de acesso do VOID permitem incluir informações sobre as formas de acesso aos dados *Linked Data*, como os *endpoits* SPARQL e os conjuntos de RDF descarregáveis (*dump* RDF) em diferentes formatos de serialização, permite fornecer acesso as entidades principais em conjuntos de dados hierarquizados e permitem indicar pontos de acesso desreferenciáveis, disponíveis como URIs HTTPs (W3C, 2011).

Os metadados de acesso podem ser explorados para avaliação de dimensões relacionadas a categoria acessibilidade. O modelo proposto por Färber *et al.*, (2016), por exemplo, acrescenta à dimensão acessibilidade o critério provisionamento de metadados sobre o conjunto de dados, que avalia a presença de metadados VOID como um aspecto positivo da qualidade dos dados.

Os metadados estruturais são relevantes para a etapa de exploração das fontes, pois permitem representar informações a respeito da estrutura dos URIs, sobre o "[...] vocabulário usado no conjunto de dados, estatísticas sobre o tamanho do conjunto de dados e exemplos de recursos típicos no conjunto de dados" (W3C, 2011, não paginado, tradução nossa).

As informações estáticas são bastante relevantes para exploração de fontes e aplicação de critérios de exclusão, usando o vocabulário podem ser fornecidas informações como a quantidade de triplas, de entidades, de classes, de objetos e de documentos desreferenciáveis.

A última classe de metadados descrita no VOID são **os metadados para descrever conjuntos de triplas** e a relação existente entre diferentes conjuntos, aspecto que também é abordado pelo DCAT. "Esquemas como DCAT e VoID permitem a descrição de metadados em um formato semanticamente interoperável e podem ser trocados entre vários agentes" (Debattista *et al.*, 2018, p. 863, tradução nossa).

“O DCAT é um vocabulário RDF desenvolvido para facilitar a interoperabilidade entre catálogos de dados publicados na *Web*” (W3C, 2024, não paginado, tradução nossa).

O vocabulário fornece classes e propriedades que permitem aos publicadores de dados descreverem os conjuntos de dados possibilitando a criação de catálogos de dados, facilitando a busca, recuperação e reuso dos dados, podendo inclusive impactar na preservação digital dos mesmos (W3C, 2024).

No vocabulário, um conjunto de dados é entendido como:

> [...] uma "coleção de dados, publicada ou com curadoria de um único agente, e disponível para acesso ou *download* em uma ou mais serializações ou formatos". Um conjunto de dados é uma entidade conceitual e pode ser representado por uma ou mais **distribuições** que serializam o conjunto de dados para transferência. As distribuições de um conjunto de dados podem ser fornecidas por meio de **serviços de dados** (W3C, 2024, não paginado, tradução nossa).

O DCAT fornece uma estrutura, baseada no reaproveitamento de outros vocabulários já estabelecidos, para padronizar a descrição de informações relevantes a respeito dos catálogos de dados, tais como temas, conjuntos de dados e serviços abrangidos.

Possibilita a descrição de conjuntos de dados catalogados, com propriedades para a descrição de informações que possibilitem a recuperação e seleção desses conjuntos, tais como direitos de acesso, descrição, criador, data de lançamento, data de atualização/ modificação.

Também permite a descrição de informações relacionadas as formas de distribuição dos dados, contendo propriedades para descrever o volume dos dados, seus formatos de serialização e compressão, permitindo identificar como os dados podem ser acessados, baixados e reutilizados. O DCAT:

> [...] diminui problemas como ambiguidade e auxilia no processo de descoberta de dados e serviços de dados, uma vez que busca a padronização dos termos de descrição dos catálogos, conjuntos de dados e dos relacionamentos que podem ser estabelecidos. Sendo desenvolvido para o contexto dos catálogos e seus dados. (Tomoyose, 2021, p. 158).

O DCAT se relaciona ao processo de avaliação de qualidade especialmente quando se aborda a identificação, análise e pré-seleção de conjuntos de dados de

interesse. A criação de catálogos de dados com o DCAT facilita o uso de ferramentas automáticas e semiautomáticas no processo de exploração dos dados, permitindo assim a criação e aplicação de filtros temáticos, temporais, geográficos, relacionados a proveniência e aos direitos de uso desses conjuntos de dados.

As informações de proveniência dos dados são relevantes para o processo de seleção, permitindo avaliar aspectos como a confiabilidade e a relevância dos conjuntos de dados. A proveniência pode ser definida como “Informações sobre entidades, atividades e pessoas envolvidas na produção de um dado ou objeto, que podem ser usadas para formular avaliações sobre sua qualidade ou confiabilidade” (W3C, 2013, não paginado, tradução nossa).

As informações de proveniência no *Linked Data* podem ser fornecidas utilizando a estrutura da chamada família PROV. “A Família de Documentos PROV define um modelo, serializações correspondentes e outras definições de suporte para permitir o intercâmbio interoperável de informações de procedência em ambientes heterogêneos, como a *Web*” (W3C, 2013, não paginado, tradução nossa).

A família PROV:

> [...] oferece um conjunto de especificações que promovem a interoperabilidade e a troca de informações de proveniência na *Web*, assegurando que metadados abrangem não apenas descrições contextuais, mas também informações essenciais para o gerenciamento, uso e preservação dos dados. Ao permitir o rastreamento detalhado de informações sobre a origem, os padrões de metadados PROV fornecidos pela W3C auxiliam na promoção da garantia dos dados disponibilizados em repositórios de dados (Silva; Arakaki, 2025, p. 5).

Eles permitem identificar os responsáveis pelos dados, rastrear a sua origem e realizar o controle de versão, analisando aspectos relacionados com a temporalidade dos dados. Os aspectos de proveniência também são relevantes para a interligação, considerando a necessidade de atribuir informações de proveniência aos *links* gerados.

Embora seja indiscutível a relevância desses vocabulários e seu potencial de contribuição para facilitar a avaliação de qualidade e a seleção de fontes, eles ainda são pouco explorados pela literatura de qualidade de dados, aparecendo apenas pontualmente dos modelos de analisados.

Discutido o uso de vocabulários relacionados ao fornecimento de metadados relevantes para a avaliação de qualidade, a próxima seção discute o vocabulário de qualidade de dados proposto pelo W3C e seu papel na avaliação de qualidade.

6.5.3 O *Data Quality Vocabulary* (DQV) na representação da qualidade dos dados.

A descrição formal da qualidade dos conjuntos de dados tem impacto direto na sua reutilização, auxiliando consumidores de dados no processo de seleção de fontes adequadas para suas aplicações, permitindo o uso de agentes computacional nesse processo, facilitando a adoção de técnicas automáticas e semiautomáticas.

O W3C (2017, não paginado, tradução nossa) afirma que:

> A documentação da qualidade dos dados facilita significativamente o processo de seleção do conjunto de dados, aumentando as chances de reutilização. Independentemente das peculiaridades específicas do domínio, a qualidade dos dados deve ser documentada e os problemas de qualidade conhecidos devem ser explicitamente declarados nos metadados.

Para instrumentalizar essa descrição o W3C elaborou o *do Data Quality Vocabulary (DQV),* um vocabulário que prove termos e formaliza as relações existentes entre os conceitos utilizados para a descrição formal da qualidade de conjuntos de dados.

A descrição explícita das limitações dos conjuntos de dados tem impacto direto na sua confiabilidade. Conhecendo os potenciais problemas de qualidade os consumidores passam a ter maior autonomia para verificar se esses dados atendem ou não às necessidades da sua aplicação.

A entidade central e principal das declarações a serem feitas utilizando o DQV é o *dataset*, ou seja, o conjunto de dados objeto da avaliação de qualidade.

A indicação desse conjunto de dados pode ser realizada de duas maneiras. Pode ser feita por meio da instância do DCAT *dcat:Dataset*, que se refere a “Uma coleção de dados, publicada ou com curadoria de um único agente, e disponível para acesso ou *download* em uma ou mais representações” (W3C, 2020, não paginado, tradução nossa). Pode ser feita pela instância *dcat:Distribution*, que “representa uma forma acessível de um conjunto de dados, como um arquivo para *download*” (W3C, 2020, não paginada tradução nossa).

Para representar as características de qualidade dos conjuntos de dados são estabelecidas classes, propriedades e instâncias. Na explicação de cada um desses elementos é apresentada uma definição, a indicação de relação com outros elementos, apontando se o elemento é uma subclasse de outra classe do vocabulário. Também se indica a equivalência desse termo em outros vocabulários, quando essa equivalência se mostra pertinente. A estrutura geral do vocabulário é composta por oito classes e duas subclasses, apresentadas no quadro 24.

Quadro 24 - Classes e subclasses do DQV.

Rótulo	Definição
dqv:QualityMeasurement	Classe que representa os resultados da avaliação de qualidade de um *dataset* em relação a uma métrica específica. A classe é relacionada a propriedades que permitem indicar a métrica que está sendo observada, o valor da avaliação, a unidade de medida e o tipo de dados que se espera obter com essa avaliação
dqv:Metric	Representa os indicadores utilizados para mensurar as dimensões de qualidade
dqv:Dimension	Representa a característica que está em observação, cada dimensão obrigatoriamente precisa ser associada a uma ou mais métricas e agrupada em uma categoria
dqv:Category	Representa a organização das dimensões de acordo com a perspectiva de qualidade adotada
dqv:QualityMeasurementDataset	Categoria que permite representar um *dataset* de avaliação de qualidade, onde estariam armazenados os resultados de avaliação de um ou mais *datasets*
dqv:QualityPolicy	Permite representar uma política ou acordo, adotado pelo provedor dos dados, que tenha orientado a elaboração, manutenção e disponibilização dos dados
dqv:QualityAnnotation	Permite a representação de notas, como a indicação de selos de qualidade ou o registro de *feedbacks*. É necessário vincular essa propriedade com uma indicação de motivação para especificar o propósito dessa anotação
dqv:QualityCertificate (subclasse)	É uma subclasse de *dqv:QualityAnnotation*, focada especificamente na indicação de certificados ou selos de qualidade fornecidos por outras instituições
dqv:UserQualityFeedback (subclasse)	Também é uma subclasse de *dqv:QualityAnnotation* elaborada para indicar um *feedback* dos usuários em relação a percepção de qualidade. Além da necessidade de indicar a motivação da nota, é necessário incluir uma descrição do tipo de feedback (Ex: questionamentos, classificações, sugestões e etc)
dqv:QualityMetadata	Representa o agrupamento de metadados de qualidade, onde deve-se relacionar os

	certificados de qualidade, a política, as dimensões, as métricas e anotações de um determinado conjunto de dados. Esse registro pode ser feito em forma de triplas RDF

Fonte: Autora (2025)

Como é possível observar no quadro 24, o DQV não fornece dimensões e métricas para a avaliação de qualidade, mas sim os termos e propriedades necessárias para se referir aos resultados da avaliação de qualidade.

O vocabulário apresenta como sugestões não prescritivas as dimensões e métricas da ISO 25012 e o modelo de avaliação de qualidade de Zaveri *et al.* (2012). Caso as métricas e dimensões disponibilizadas como exemplo não atendam à necessidade do usuário, é possível ampliar o vocabulário para que represente o modelo de qualidade adotado. Também é possível mesclar novas categorias, dimensões e métricas às sugeridas no modelo (W3C, 2016).

A única ressalva feita no documento em relação ao uso de classificações, dimensões e métricas novas é a de que ela pode afetar a interoperabilidade com outras fontes, e ainda prejudicar a utilização de instrumentos voltados para a identificação e comparação automática de qualidade de dados.

O W3C (2016, não paginado, tradução nossa) indica, portanto que os usuários, ao estenderem o DQV, devem [...] “estar cientes de que confiar nas classificações e métricas existentes aumenta a interoperabilidade, ou seja, a chance de que agentes humanos e máquinas possam entender e explorar adequadamente suas avaliações de qualidade”.

O W3C (2016) enfatiza a importância de incentivar a participação da comunidade de usuários e de instituições responsáveis pelo fornecimento de certificações de qualidade no processo de avaliação. Essa participação faz com que os usuários não dependam exclusivamente dos publicadores para ter acesso a informações de qualidade dos conjuntos de dados. Faz ainda com que as avaliações realizadas por outros usuários de maneira orgânica possam ser reutilizadas, evitando retrabalho.

Outro aspecto importante sobre a estrutura do DQV, ressaltada pelo W3C, é o de que sua abrangência não se limita a descrição da qualidade dos dados, mas também a descrição da qualidade dos metadados.

> Os elementos DQV podem ser aplicados não apenas para expressar metadados sobre a qualidade dos conjuntos de dados; eles também podem ser usados para expressar declarações sobre a qualidade dos próprios metadados. Isso é especialmente verdadeiro quando se trata de representar a proveniência desses metadados ou sua conformidade com os padrões de metadados estabelecidos (W3C, 2017, não paginado, tradução nossa.

Mas uma vez destaca-se a importância dos metadados e seu papel tanto no processo de avaliação de qualidade como na recuperação e no reuso dos dados disponibilizados na *Web*.

Apresentados os principais vocabulários relacionados com a avaliação de qualidade de dados Linked Data, a próxima seção apresenta algumas ferramentas que podem auxiliar na condução da avaliação.

6.6 Ferramentas do processo de avaliação

Com a realização da RSL e a análise dos modelos de qualidade foram identificadas diversas ferramentas para auxiliar na realização da avaliação de qualidade. As ferramentas manuais, como modelos de qualidade e metodologia foram apresentados e discutidos nas seções 6.4 e 6.5. A presente seção busca, portanto, discutir as ferramentas de abordagem automática e semiautomática identificadas.

Localizar as ferramentas descritas nos artigos do *corpus* se mostrou um desafio, considerando que muitas dessas ferramentas estavam em estágios iniciais, de prototipagem, quando da publicação dos artigos e não avançaram para fases seguintes, ou avançaram com novos nomes e funcionalidades. Nem sempre os *links* disponibilizados nos artigos eram permanentes e em sua maioria já se encontram indisponíveis e não apresentam nenhum direcionamento para a ferramenta apresentada.

As ferramentas que puderam ser localizadas foram encontradas majoritariamente em *githubs* disponibilizados pelos autores, ainda assim, algumas possuem *links* necessários para o seu funcionamento que não estão mais disponíveis. Foram consideradas para discussão apenas as ferramentas atualmente disponíveis e com informações a respeito de como acessá-las e aplicá-las na avaliação.

A ferramenta LUZZU[14] é uma estrutura criada para permitir a avaliação de qualidade dos dados de maneira automática e customizável. “É uma estrutura genérica baseada na Ontologia de Qualidade de Conjuntos de Dados (DaQ), permitindo que os usuários definam suas próprias métricas de qualidade” (Debattista, 2019, não paginado, tradução nossa).

A ferramenta permite a aplicação das métricas, a geração de metadados customizáveis e a geração de relatórios detalhados a respeito dos resultados da avaliação de qualidade (Debattista; Auer; Lange, 2016).

O SemQuire[15],também é uma ferramenta para avaliação geral da qualidade, focada em aspectos de qualidade específicos do *Linked Data*, aproveitando-se das estruturas em RDF e SPARQl dos conjuntos de dados para a realização da avaliação de qualidade. A ferramenta também leva em consideração o DQV, que permite a exportação dos resultados de avaliação em formato legível por máquina (Langer *et al.*, 2018).

O Roomba[16] é uma ferramenta focada em metadados relacionados a conjuntos de dados *Linked Data*. As funcionalidades do *Roomba* são validação, correção e geração de metadados (Assaf; Troncy; Senart, 2015).

> O processo de validação de metadados identifica informações ausentes e a capacidade de corrigi-las automaticamente. Cada conjunto de metadados (geral, acesso, propriedade e procedência) é validado e corrigido automaticamente quando possível. Cada tarefa do *profiler* possui um conjunto de campos de metadados para verificação. O processo de validação verifica se cada campo está definido e se o valor atribuído é válido (Assaf, 2023, não paginado, tradução nossa).

A ferramenta permite a inspeção dos URIs, dos termos de metadados, e do modelo dos dados, permite ainda o enriquecimento dos metadados existentes. A fermenta permite verificar se o conjunto de metadados:

> Suporta múltiplas serializações; Possui diferentes pontos de acesso a dados; Utiliza vocabulários de descrição de conjuntos de dados; Existência de descrições sobre seu tamanho; Existência de descrições sobre sua estrutura (Tipo MIME, Formato); Existência de descrições sobre sua organização e categorização; Existência de *links* desreferenciáveis para o conjunto de dados e seus recursos;

[14] https://eis-bonn.github.io/Luzzu/about.html
[15] https://vsr.informatik.tu-chemnitz.de/demos/semquire
[16] https://github.com/ahmadassaf/roomba

> Existência de um *dump* RDF que pode ser baixado pelos usuários; Existência de pontos de extremidade consultáveis que respondem a consultas diretas; Existência de URLs desreferenciáveis válidas (responder à solicitação HTTP); Existência de informações de licença legíveis por humanos e máquinas; Existência de *links* desreferenciáveis para as informações completas da licença; Existência de carimbos de data/hora que possam acompanhar suas modificações; Inclui o tipo MIME correto para o conteúdo; Inclui o tamanho correto para o conteúdo; Ausência de erros sintáticos no nível de *links*; Uso do esquema HTTP URI; Existência de pelo menos um arquivo RDF exemplar; Existência de informações gerais (título, URL, descrição) para o conjunto de dados; Existência de lista de discussão, quadro de mensagens ou ponto de contato; Existência de metadados que descrevem suas informações autoritativas; Uso de controle de versão (Assaf, 2023, não paginado, tradução nossa).

O TripleCheckMate[17] é uma ferramenta customizável que cobre aspectos gerais da avaliação de qualidade de dados *Linked Data*, criada visando auxiliar na avaliação baseada em *crowdsourcing*. Nessa abordagem a tarefa de checagem é dívida entre especialista.

> Para usar a ferramenta, o usuário precisa se autenticar, o que não apenas previne *spam*, mas também ajuda a monitorar suas avaliações. Após a autenticação, ele prossegue com a seleção de um recurso (Etapa 1). Ele tem três opções: (i) por classe (ii) completamente aleatório e (iii) manual. [...] Após selecionar um recurso, o usuário é apresentado a uma tabela mostrando cada triplo pertencente a esse recurso em uma única linha (Kontokostas *et al.*, 2013, p. 6).

Uma vez analisada a tripla, o avaliador pode indicar a presença de aspectos incorretos e indicar o tipo de problema com base em uma lista pré-estabelecida fornecida pela própria ferramenta. A ferramenta permite ainda calcular a taxa de concordância entre diferentes avaliadores (Kontokostas *et al.*, 2013).

O Dacura[18] é um *framework* focado no processo de curadoria de dados *Linked Data* pensado para ser aplicado a todo o ciclo de vida dos dados, “fornece aos curadores de conjuntos de dados [...] ferramentas para coletar e realizar a curadoria de conjuntos de dados *Linked Data* em evolução mantendo a qualidade ao longo do tempo (Feeney *et al.*, 2014).

[17] TripleCheckMate
[18] https://github.com/kevinchekovfeeney/dacura-ld/tree/public-ld/js

A ferramenta prevê quatro funções, relacionadas a diferentes perfis de usuário, sendo elas: arquiteto de dados, especialista em domínio, coletor de dados e consumidor. Essa abordagem pretende permitir a divisão das tarefas de curadoria dos dados entre diferentes pessoas com níveis distintos de conhecimento técnico (Feeney *et al.*, 2014).

O LD *Sniffer*[19] é uma ferramenta de código aberto focada na avaliação de qualidade da acessibilidade de dados *Linked Data*. Ela permite:

> Realizar avaliações de qualidade sobre a acessibilidade de dados *Linked Data*. Gera resultados de avaliação inequívocos e comparáveis com semântica explícita, definindo métricas de qualidade e resultados de avaliação em RDF usando o vocabulário de qualidade de dados do W3C. O LD-Sniffer também é distribuído como uma imagem Docker, melhorando a facilidade de uso com zero configurações. (Mihindukulasooriya; García-Castro; Gómez-Pérez, 2017, p. 149, tradução nossa).

Como observado, a ferramenta também se utiliza do DQV para exportar os resultados do processo de avaliação.

O *Deep Fact Validation* (DeFacto[20]) é uma ferramenta para a checagem de fatos em triplas RDF que se baseia tanto na confirmação dos fatos com outras fontes como na pontuação da confiabilidade da fonte utilizada na checagem dos fatos (Esteves *et al*., 2018). Em relação a suas funcionalidades:

> O DeFacto visa fornecer uma maneira eficaz de validar fatos, fornecendo ao usuário trechos relevantes de páginas da *web*, bem como informações adicionais úteis, incluindo uma pontuação para a confiança que o DeFacto tem na exatidão do fato inserido. Para atingir esse objetivo, o DeFacto coleta e combina evidências de páginas da *web* escritas em vários idiomas. Além disso, o DeFacto fornece suporte para fatos com escopo temporal, ou seja, pode estimar em qual período de tempo um fato foi válido (Gerber *et al.*, 2015, p. 85, tradução nossa).

O RDF Validation Service[21] é uma ferramenta do W3C que permite a validação de triplas RDF. As triplas podem ser inseridas diretamente na ferramenta *Web* ou a checagem pode ser feita por meio da inserção de URI de acesso ao conjunto de dados (W3C, 2006).

---

[19] https://github.com/nandana/ld-sniffer?tab=readme-ov-file
[20] https://github.com/DeFacto/DeFacto?tab=readme-ov-file
[21] https://www.w3.org/RDF/Validator/

Os dados são inseridos e a ferramenta permiti validar a estrutura RDF dos dados. A ferramenta é amplamente adotada pelos modelos de qualidade para a checagem da estrutura sintática das declarações. O quadro 25 apresenta uma síntese das ferramentas discutidas.

Quadro 25 - Síntese das ferramentas discutidas

Ferramenta	Aplicação
Luzzu	Avaliação geral de qualidade de dados *Linked Data,* com possibilidade de customização de dimensões e métricas.
Roomba	Validação e enriquecimento de metadados dos conjuntos de dados
SemQuire	Avaliação geral da de qualidade de dados *Linked Data* com possibilidade de exportação dos resultados utilizando DQV
TripleCheckMate	Auxílio na avaliação realizada com base em *crowdsourcing*, permitindo a comparação das decisões de diferentes avaliadores
Dacura Quality Service	Auxílio na curadoria dos dados ao longo de todo o ciclo de vida dos dados
LD Sniffer	Avaliação de aspectos relacionados à acessibilidade dos dados *Linked Data*
DeFacto	Validação do conteúdo das triplas com base em fontes da *Web*, permitindo mensurar a confiabilidade da fonte consultada para validação
RDF Validator	Validação da estrutura RDF dos dados publicados como *Linked Data*

Fonte: Autora (2025)

Apresentadas as ferramentas disponíveis para auxiliar na avaliação automática e semiautomática dos dados *Linked Data*, a próxima seção apresenta o fluxo da seleção e interligação de dados *Linked Data.*

7 FLUXO DA SELEÇÃO DE FONTES E DA INTERLIGAÇÃO DE DADOS *LINKED DATA*

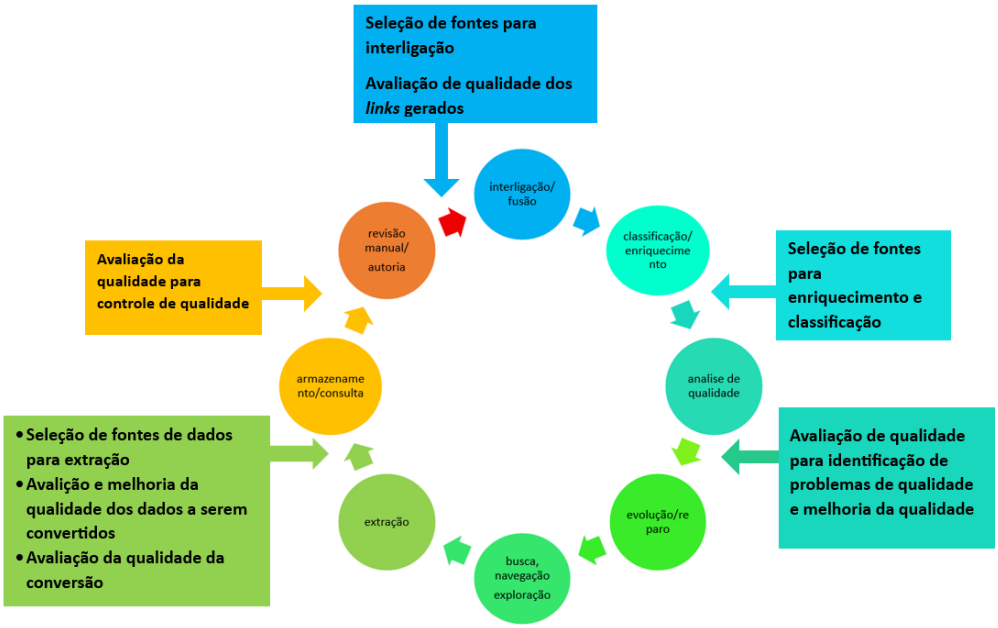
Considerando a análise teórica, terminológica e processual da qualidade dos dados *Linked Data*, a presente seção foi construída com o objetivo de mapear o fluxo da seleção de fontes para interligação de dados *Linked Data,* baseado em aspectos de qualidade de dados. Buscou-se também apresentar um modelo de protocolo e *checklist* para auxiliar na condução da seleção.

A análise do contexto e dos diferentes processos que permeiam a avaliação de qualidade de dados *Linked Data* permitiu observar que esse processo é centrado em dois agentes principal: consumidores e publicadores. Esse papeis não são fixos no *Linked Data*, consumidores muitas vezes também são publicadores de dados, por tanto o papel depende da atividade exercida no momento da avaliação de qualidade e da seleção de fontes.

Os publicadores realizam o processo de avaliação de qualidade visando garantir a melhoria e o controle da qualidade de seus dados antes e depois da publicação. Já os consumidores realizam a avaliação de qualidade visando principalmente a seleção de fontes e a avaliação dos resultados da interligação.

A figura 46 apresenta uma adaptação do ciclo de vida dos dados *Linked Data* proposto Auer *et al.* (2012), indicando os momentos em que ocorrem atividades de avaliação de qualidade e seleção de fontes.

Figura 46 - Ciclo de vida de dados *Linked Data* com identificação das atividades de avaliação e seleção de fontes.



Fonte: Traduzido e adaptado de Auer *et al.* (2012)

Com base na análise dos CVDs *Linked Data* e das discussões a respeito da interligação no *Linked Data* foi possível observar que a interligação pode ser realizada a partir de duas abordagens principais: Identificação pautada em análise exploratória de fontes e identificação pautada na explicitação de necessidades informacionais pré-estabelecidas.

A abordagem baseada em exploração busca identificar *links* entre o conjunto original e fontes bem estabelecidas ou relevantes para o domínio em questão, sem a definição prévia de um grupo entidades do conjunto que serão alvo da ligação e do relacionamento que se espera criar com os dados das fontes externas.

A abordagem baseada na explicitação de necessidade informacional previamente estabelecida parte da identificação do tipo de relacionamento que se pretende criar e do grupo de entidades do conjunto interno que será alvo dessa ligação com dados da fonte externa. As fontes são identificadas com base no seu potencial de contribuir para a necessidade informacional da fonte original.

Em ambas as abordagens a seleção de fontes pode ser mais bem compreendida por meio da discussão da etapa que a antecede (planejamento) e sucede (interligação). O fluxo da seleção de fontes pode ser organizado, portanto, em três etapas correlacionadas: planejamento, seleção de fontes e interligação.

As abordagens compartilham, portanto, os mesmos agentes (publicadores e consumidores) e etapas, mas se diferenciam pelos objetivos, que afetam as atividades das etapas de planejamento e a de seleção de fontes.

A etapa de planejamento parte da identificação do tipo de abordagem adotada, dos objetivos e de uma análise a respeito de como será conduzido o processo de avaliação e seleção de fontes para que esses objetivos sejam atingidos.

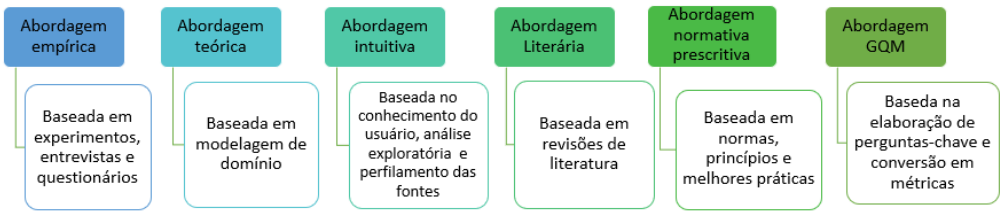
Os consumidores podem realizar a interligação como uma etapa pontual, que antecede a publicação dos dados no ciclo de vida do *Linked Data,* ou de maneira esporádica, quando novas fontes de dados relevantes são identificadas. Também podem realizar interligação quando se identifica uma necessidade de completar os dados em relação a um aspecto específico.

Uma vez estabelecidos os objetivos e metas, esses precisam ser convertidos em um modelo de qualidade, que irá definir como o processo de avaliação será conduzido, estabelecendo dimensões e métricas para apoiar a avaliação.

Observou-se que um dos principais desafios em relação a avaliação de qualidade é o estabelecimento do modelo de qualidade, como selecionar as categorias, dimensões e métricas a serem adotadas para compor esse modelo em abordagens contextuais da avaliação.

Foram identificadas diferentes abordagens para auxiliar no estabelecimento de modelos de qualidade, a figura 47 apresenta uma síntese dessas abordagens.

Figura 47 - Abordagens para construção de modelos de qualidade



Fonte: autora (2025)

As abordagens apresentadas na figura 47 não necessariamente precisam ser aplicadas de maneira isolada, podendo ser adotada uma abordagem mista que mescle duas ou mais dessas abordagens.

Como mencionado, para a seleção de fontes, além de se estabelecer as dimensões e métricas a serem adotadas, torna-se necessários estabelecer como serão interpretados os resultados obtidos por meio da avaliação.

Issa *et al.* (2021), exemplifica esse desafio mencionado que: se um conjunto de dados tem 70% de completude, ele deve ser aceito ou recusado como fonte de dados? Saber o nível de qualidade dos dados em relação a determinada dimensão não basta para que um conjunto de dados seja ou não considerado.

Portanto, para o processo de seleção de fontes, a elaboração do modelo de qualidade precisa ser contextual, e é necessário o estabelecimento prévio de valores mínimos esperados para cada critérios de qualidade.

O estabelecimento desses aspectos depende dos objetivos do processo de avaliação e da abordagem utilizada para a identificação de *links.* A explicitação das metas de qualidade pode ser representa como uma lista de características esperadas, contendo níveis ou requisitos mínimos de qualidade em relação a dimensões e

métricas, pode ser representada por meio do estabelecimento de padrões-ouro, pode ainda ser representada no formato de critérios de exclusão, que serão a base para a seleção das fontes. Deve-se estabelecer ainda se serão necessárias ferramentas específicas para auxiliar no processo de avaliação.

Ao final do planejamento, devem ser identificadas as fontes que serão alvo da avaliação e da seleção, a escolha das fontes depende também dos objetivos e da abordagem que será adotada. A identificação de fontes pode ocorrer a partir de revisão de literatura, identificando fontes confiáveis para o tipo de dados buscado. Pode ser feita por meio da consulta de catálogos reconhecidos de dados *Linked Data*, como o mantido pelo projeto LOD. Pode ainda ser feito explorando propriedades de vocabulários que sejam relevantes para o domínio em questão.

A etapa de seleção tem início com a análise exploratória dos dados que serão objeto do processo de avaliação, visando a obtenção de informações que serão relevantes para a aplicação da estratégia de avaliação estabelecida. Pode-se verificar como os dados estão estruturados, que vocabulários são utilizados, quais são as propriedades que devem estar presentes, que tipos de dados são esperados como valor para essas propriedades e a quais problemas de qualidade podem estar sujeitos a esses dados.

Na análise exploratória dos dados é pertinente verificar a existência de metadados que possam auxiliar no processo de avaliação com informações relacionadas à proveniência, ao acesso e ao conteúdo dos dados, que podem ser fornecidos por meio de propriedades dos vocabulários SKOS, DCAT, DC, VOID e PROV.

As Propriedade do vocabulário VOID podem ainda fornecer informações estatísticas importantes para a etapa de exploração, enquanto propriedades do vocabulário DQV podem fornecer informações relevantes a respeito da qualidade dos dados que podem economizar tempo do processo de avaliação de qualidade.

A análise exploratória dos dados também pode incluir a aplicação de uma ferramenta de identificação de *links* e análise de similaridade, uma vez que a qualidade e a quantidade de *links* identificados podem influenciar no processo de seleção, caso esse aspecto seja incluído no modelo de qualidade na etapa de planejamento.

Obtidas as informações relevantes, a próxima atividade é a avaliação de qualidade, que deve ser baseada no modelo de qualidade e na estratégia de avaliação

pré-estabelecida, utilizando as ferramentas necessárias para auxiliar nessa atividade. Com base na análise da qualidade das fontes e na aplicação dos critérios de exclusão é realizada a de seleção das fontes, que serão aplicadas então na criação dos *links*.

A etapa de interligação ocorrera com base nas fontes selecionadas**.** Como discutido, os *links* podem ser categorizados em: *links* de identidade; *links* de relacionamento e *links* de vocabulários. A figura 48 apresenta uma síntese de cada uma dessas categorias.

Figura 48 - Principais categorias de *Links RDF*

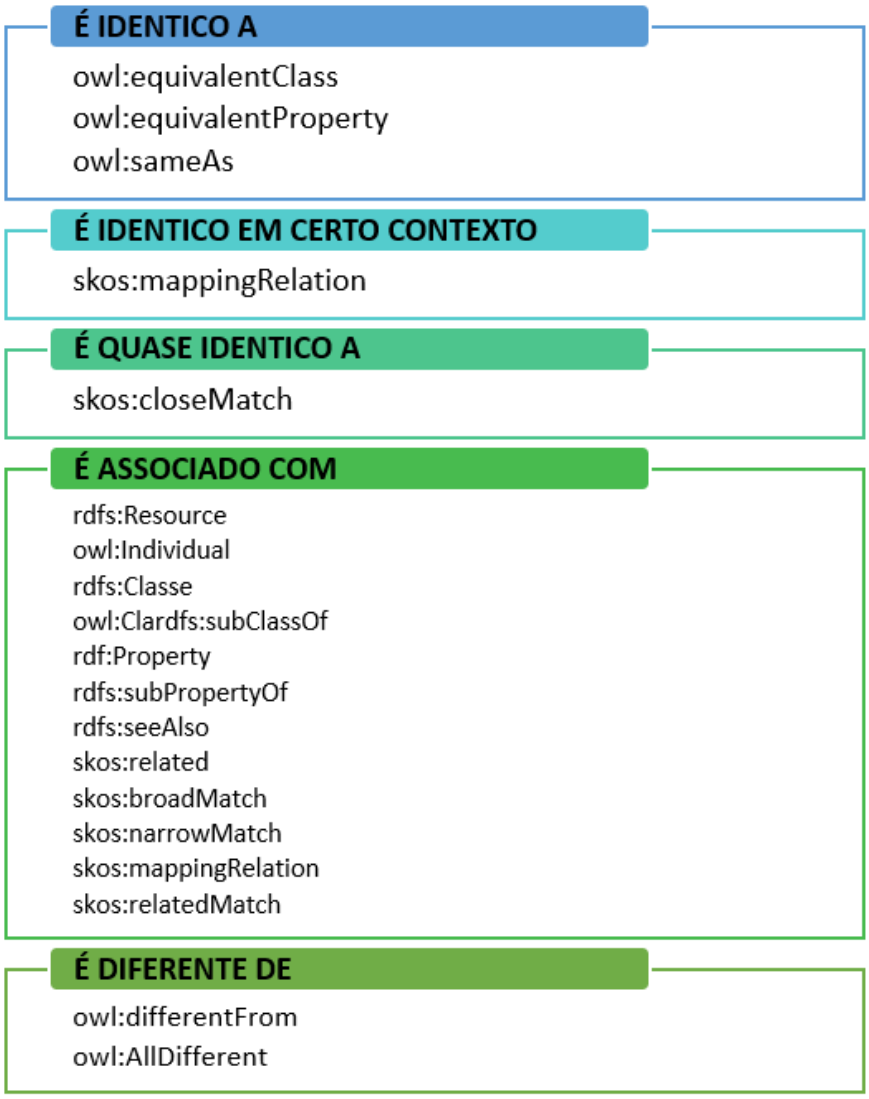
Fonte: Autora (2025)

Os *links* podem ser criados manualmente ou de maneira automática/semiautomática, embora, para bancos de dados maiores seja inviável a criação manual de *links*.

Existem diferentes relações possíveis entre as entidades interligadas, e a criação de cada *link* depende da identificação de uma propriedade apropriada, preferencialmente em um vocabulário bem estabelecido do domínio de origem, que represente o tipo de relação existente entre as entidades relacionadas ao objeto da interligação.

A análise dos vocabulários permitiu identificar as principais propriedades utilizadas na interligação. A figura 49 apresenta a relação entre os principais tipos de relações, as entidades e as propriedades para representar essas relações em OWL, SKOS e RDFs.

Figura 49 - Tipos de relação e propriedades relacionadas em OWL, SKOS e RDFs



Fonte: Autora (2025)

Na abordagem baseada em explicitação de necessidade informacional o planejamento dos *links* a serem criados é uma etapa de planejamento, uma vez que é necessário ter conhecimento sobre o tipo de *links* que se pretende estabelecer, de quais as relações que se pretende criar. Já na abordagem baseada em exploração das fontes, esses *links* são identificados apenas posteriormente, e sua relevância pode ser considerada um critério de seleção.

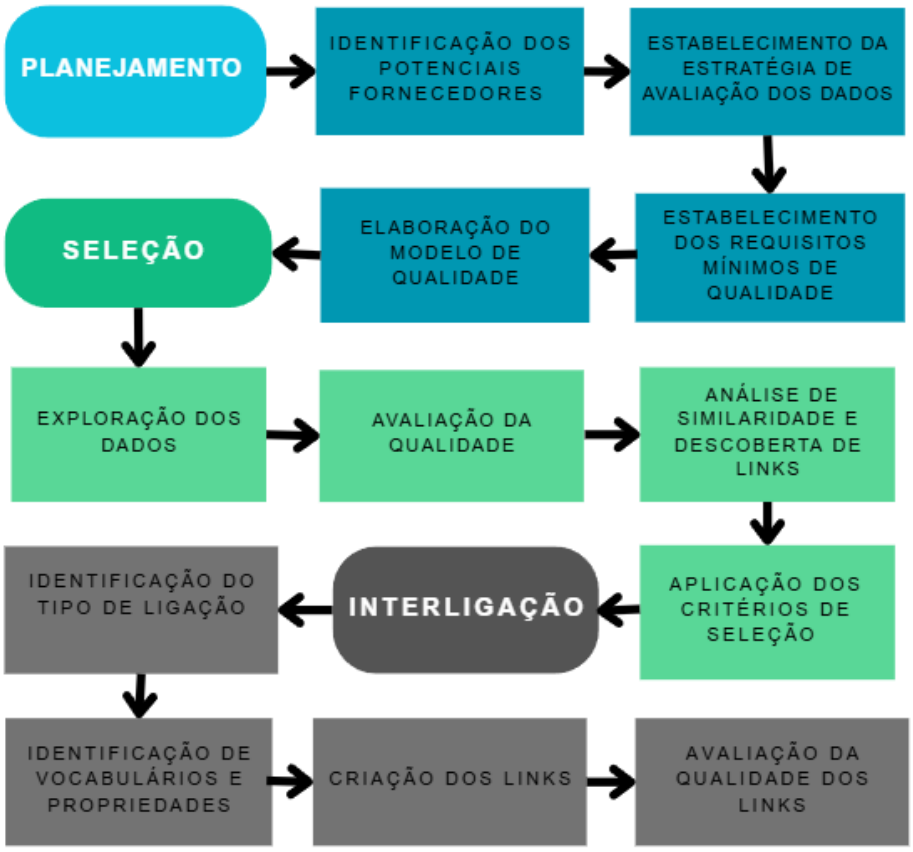
Uma vez criados os *links*, torna-se necessário avaliar a qualidade desses *links* criados. Essa avaliação pode ser feita de maneira semiautomática ou manual. Considerando que contexto do *Linked Data* geralmente é relacionado a grandes

volumes de dados, as abordagens de avaliação manual podem empregar técnicas de *crowsouricing* ou de amostragem.

Como mencionado, as duas abordagens de seleção compartilham as mesmas etapas e muitas atividades, diferenciando-se principalmente em relação aos objetivos, ao modelo de qualidade e às informações necessárias para o processo de avaliação, ressaltando-se que a maior parte das diferenças ocorre na etapa de planejamento.

A abordagem baseada em exploração parte da identificação de fontes em potencial, e pode ser uma atividade pontual ou processual, a depender dos objetivos estabelecidos. Pode ser realizada com base em sua relevância para o cenário do *Linked Data* ou para o domínio em questão. A figura 50 apresenta o fluxo de seleção de fontes baseada em análise exploratória.

Figura 50 - Fluxo para seleção de fontes baseada em análise exploratória



Fonte: Autora (2025)

A estratégia de avaliação dessa abordagem tende a ser mais abrangente e relacionada principalmente a aspectos intrínsecos e de acessibilidade dos dados, uma vez que a confiabilidade das fontes é um requisito pré-identificado. Em determinados

contextos, a frequência de atualização dos dados também é um aspecto relevante para o processo de avaliação.

Visando auxiliar no planejamento da seleção de fontes, elaborou-se um protocolo pensado para orientar a seleção de fontes baseado em análise exploratória. O quadro 26 apresenta esse protocolo.

Quadro 26 - Protocolo para seleção de fontes baseado em análise exploratória

Protocolo para seleção de fontes baseado em análise exploratória	
Descrição do projeto	Descrição detalhada do contexto e da motivação que levou a busca por fontes de dados para ligação
Frequência da busca por fontes	Descrição da frequência com que se pretende realizar a avaliação de fontes (pontual ou sistemática)
Estratégia para identificação e pré-seleção de fontes	Indicar quais serão as estratégias adotadas para encontrar as fontes e quais aspectos de confiabilidade serão considerados para a pré-seleção
Aspectos de qualidade a serem avaliados	Lista dos aspectos de qualidade que serão considerados para a avaliação
Critérios de exclusão de fontes	Conversão dos aspectos de qualidade em critérios de exclusão
Descrição da estratégia de avaliação	Descrição de como será conduzida a avaliação de qualidade dos dados, indicando ferramentas necessárias para auxiliar nesse processo
Modelo de qualidade adotado	Conversão dos aspectos de qualidade e dos critérios de exclusão em dimensões, critérios e métricas mensuráveis
Formas de compartilhamento dos resultados	Indicar se/como os resultados do processo de avaliação serão compartilhados e se será feito o uso de vocabulários para esse processo. (No caso de adoção do DQV, indicar as adaptações necessárias para que o vocabulário seja aplicável em relação ao modelo de qualidade adotado.
Estratégia para criação dos *links*	Descrever as estratégias e ferramentas utilizadas para criação dos *links*
Tipos de *links* criados e propriedades	Relacionar os tipos de *links* criados às respectivas propriedades
Estratégia de avaliação da qualidade dos *links*	Descrição de quais critérios e abordagens serão adotados para avaliar a qualidade dos *links* gerados (abordagem manual, *crowdsourcing*, uso de ferramentas automáticas e etc.)

Fonte: Autora (2025)

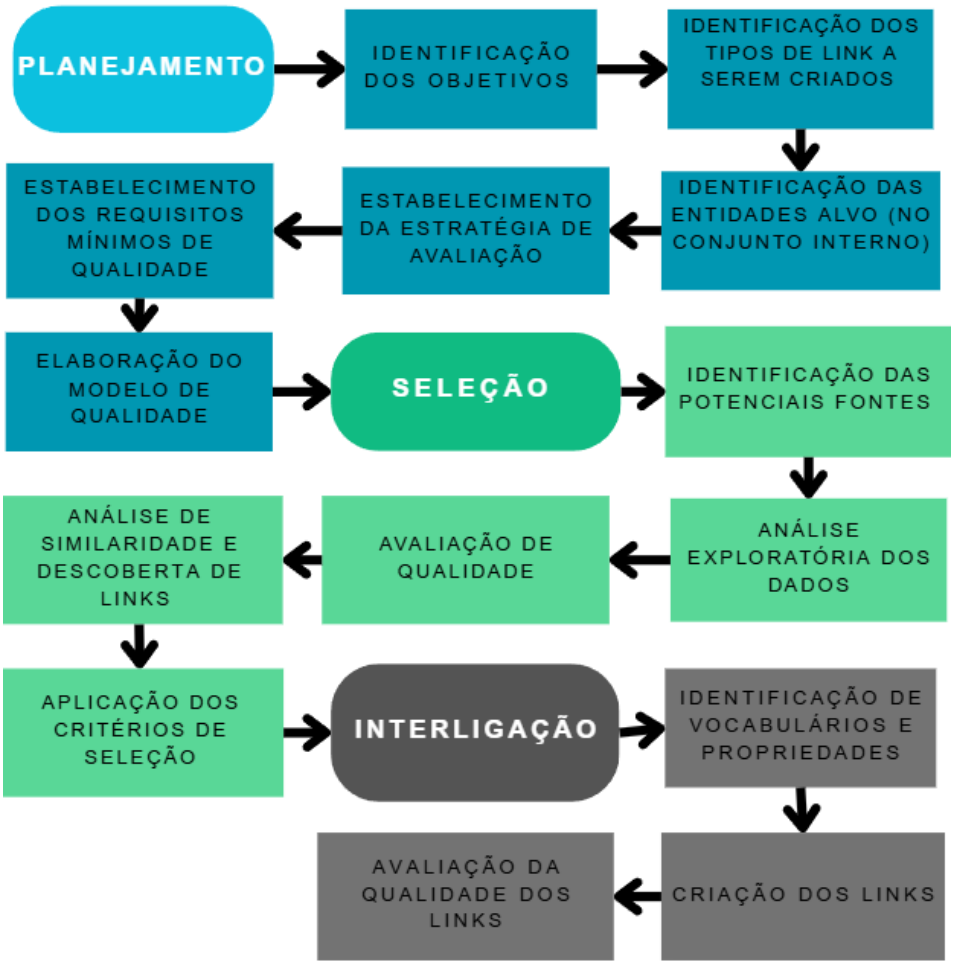
Em relação a abordagem baseada em explicitação de necessidade informacional pré-estabelecida, observa-se que, diferente da abordagem anterior, onde primeiro se identificam as fontes e depois se estabelece o processo de avaliação, essa abordagem se relaciona com a busca pela completude dos dados do consumidor e/ou com os objetivos do projeto que esse pretende realizar: consiste em processo de explicitação das necessidades informacionais.

Nessa abordagem a etapa de **planejamento** é composta por seis atividades, sendo elas: Identificação dos objetivos; Identificação dos tipos de *link* a serem criados; Identificação das entidades alvo (no conjunto interno); estabelecimento da estratégia de avaliação; estabelecimento dos requisitos mínimos de qualidade; e estabelecimento do modelo e qualidade.

Na identificação dos objetivos são descritos os motivos que levaram a busca por fontes de dados para ligação, descrevendo quais lacunas pretende-se preencher com os dados identificados.

Esses objetivos são então fragmentados em dois aspectos importantes para a seleção das fontes e para que se tenha um maior entendimento do contexto dessa seleção: quais são as entidades do conjunto dos consumidores que se pretende enriquecer, e quais são os tipos de *links* que se pretende criar. Essa abordagem é, portanto, mais contextual e tem uma etapa mais longa e completa de planejamento. A figura 51 apresenta o fluxo de seleção de fontes baseado na explicitação de necessidade informacional prévia.

Figura 51 - Fluxo de seleção de fontes baseado na explicitação de necessidade informacional prévia



Fonte: Autora (2025)

Considerando a importância do planejamento para a condução adequada dessa abordagem, elaborou-se um modelo de protocolo para auxiliar a sua condução, o quadro 27 apresenta o modelo de protocolo proposto.

Quadro 27 – Modelo de protocolo para planejamento de seleção de fontes de dados *Linked Data*

Protocolo para planejamento de seleção de fontes de *dados Linked Data*	
Contexto de uso	Fornecer as informações necessárias para que se compreenda o contexto em que os dados serão aplicados
Objetivo	Descrever os objetivos do processo de avaliação de qualidade

Identificação das entidades alvo	Apresentar uma lista que descreva o grupo de entidades do conjunto de dados interno que será alvo do processo de ligação
Identificação do tipo de ligação	Identificar a categoria e o(s) tipo(s) de *link* que se pretende criar com as fontes selecionadas
Frequência da busca por fontes	Descrição da frequência com que se pretende realizar a avaliação de fontes (pontual ou sistemática)
Estratégia para identificação de fontes	Indicar quais serão as estratégias adotadas para encontrar as fontes e quais aspectos de confiabilidade serão considerados para a pré-seleção
Critérios para avaliação da relevância das fontes	Detalhar as estratégias para a identificação de fornecedores
Critérios de exclusão de fontes	Conversão dos aspectos de qualidade em critérios de exclusão
Estratégia de exploração dos dados	Descrever como serão explorados os dados das fontes identificadas, listando as informações que precisam ser buscadas e a estratégia utilizada
Descrição da estratégia de avaliação	Descrição de como será conduzida a avaliação de qualidade dos dados, indicando ferramentas necessárias para auxiliar nesse processo
Critérios de exclusão	Indicar uma lista de critérios que leve a exclusão dos dados, associadas aos resultados obtidos por meio das métricas de avaliação de qualidade
Modelo de qualidade adotado	Conversão dos aspectos de qualidade e dos critérios de exclusão em dimensões, critérios e métricas mensuráveis
Formas de compartilhamento dos resultados	Indicar se/como os resultados do processo de avaliação serão compartilhados e se será feito o uso de vocabulários para esse processo (No caso de adoção do DQV, indicar as adaptações necessárias para que o vocabulário seja aplicável em relação ao modelo de qualidade adotado)
Estratégia para criação dos *links*	Descrever as estratégias e ferramentas utilizadas para criação dos *links*
Vocabulários e propriedades utilizados	Relacionar os tipos de *links* criados às respectivas propriedades
Estratégia de avaliação da qualidade dos *links*	Descrição de quais critérios e abordagens serão adotados para avaliar a qualidade dos *links* gerados (abordagem manual, *crowdsourcing*, uso de ferramentas automáticas e etc.)

Fonte: Autora (2025)

Embora seja complexo elaborar um modelo de qualidade para a seleção de fontes que atenda aos objetivos de ambas as abordagens, considerando que a seleção de fontes possui um aspecto fortemente contextual, partiu-se então do

objetivo de identificar aspectos gerais da qualidade que podem ser observados visando auxiliar na seleção de fontes, compondo uma *checklist* para avaliação de qualidade que pode ser adaptado as necessidades e objetivos estabelecidos.

Observa-se que aspectos relacionados a categoria acessibilidade no *Linked Data* abordam dimensões determinantes para o processo de seleção, como a disponibilidade e o desempenho dos *endpoint* SPARQL, disponibilidade de opções de *download* dos dados em RDF em diferentes formatos, disponibilidade de licença legível para humanos e máquinas, disponibilidade de metadados que permitam a recuperação e reuso dos dados e ainda os níveis de ligação com outras fontes (Zaveri *et al.*, 2012; Färber *et al.*, 2016).

As dimensões e critérios da categoria acessibilidade podem contribuir para estabelecer critérios de exclusão na medida em que se não estão disponíveis no formato necessário, se não podem ser acessados de maneira eficiente e recuperados, os dados não podem ser reutilizados. Além disso, a ausência de licença de uso explícito dificulta a reutilização dos dados.

A seleção das fontes perpassa obrigatoriamente a avaliação da pertinência e da relevância dos dados para atender a necessidade informacional pré-estabelecida. Os aspectos de relevância e pertinência dos dados são mais complexos de se avaliar, considerando que possuem um caráter qualitativo e subjetivo, com abordagem fortemente contextual. As perguntas propostas Heath e Bizer (2011, não paginado, tradução nossa), explicitam bem os aspectos de relevância e pertinência das fontes: Qual é o valor dos dados no conjunto de dados de destino? Em que medida isso agrega valor ao novo conjunto de dados?

Embora não aplicável a todos os contextos, os aspectos de temporalidade também podem influenciar no processo de seleção de dados, uma vez que para determinados contextos, a frequência de atualização dos dados pode estar diretamente ligada com a sua relevância e pertinência.

A confiabilidade das fontes é outro aspecto determinante para a seleção de fontes, e está diretamente relacionada a qualidade percebida dos dados e aos aspectos de proveniência, podendo ser medida com base na reputação do conjunto de dados ou na reputação dos fornecedores, geralmente sendo estabelecidas escalas de qualidade para a avalição dessa dimensão. Também pode ser avaliado com base na forma como são criados, como é realizada a curadoria dos dados e ainda em

relação a presença de informações e metadados necessários para avaliar os dados em relação a esse aspecto (Zaveri *et al.*, 2012; Färber *et al.*, 2016).

Com base nos modelos de Zaveri *et al.*, (2012); Färber *et al.* (2016); Cappiello *et al.*, 2016; e Melo (2017), estabeleceu-se um *checklis*t de qualidade composta apenas por critérios que podem levar a exclusão das fontes em potencial, o modelo é apresentado no quadro 28.

Quadro 28 – Modelo de *checklist* com critérios de exclusão para seleção de fontes *Linked Data*

Critérios a serem avaliados	SIM/NÃO
São disponibilizadas as informações de licença dos dados?	
A licença dos dados é adequada para os objetivos propostos?	
A quantidade de triplas é adequada para os objetivos propostos?	
A frequência de atualização dos dados é adequada para os objetivos propostos?	
É possível obter acesso aos dados no formato necessário? (*Endpoints* SPARQL funcionais ou disponibilidade para *download* nos formatos de serialização necessário)	
A fonte possui os aspectos de confiabilidade estabelecidos?	
O conjunto atende aos requisitos mínimos de relevância estabelecidos?	
Foi confirmada a validade sintática das triplas RDF?	

Fonte: Autora (2025)

Ao longo da análise dos modelos de qualidade foram identificadas métricas e ferramentas que podem auxiliar na avaliação de diferentes aspectos desse *checklist*, como a frequência de atualização das fontes, a validade sintática das triplas e validade/eficiência das formas de acesso aos dados.

Também foram listadas ferramentas que podem auxiliar nesse processo, como os validadores de RDF e as ferramentas para análise de similaridade e identificação de *links* em potencial.

Os aspectos subjetivos dessa *checklist* mostram-se mais complexos de avaliar, uma vez que sua resposta depende de uma abordagem contextual. Consideram-se aspectos subjetivos dessa lista: a confiabilidade e a relevância das fontes.

Em relação a relevância das fontes, em uma abordagem de exploração, a relevância pode ser medida pela quantidade de *links* identificados na etapa de exploração, sendo necessário o estabelecimento de uma quantidade mínima de *links* durante a etapa de planejamento.

Já em uma abordagem baseada em explicitação de necessidade informacional, uma forma de mensurar essa relevância de maneira quantitativa é medindo a relação entre o número de entidades contidas no plano de ligação e o número de *links* identificados para essa entidade, estabelecendo-se previamente um número mínimo de *links* a serem criados.

Em relação a confiabilidade das fontes, alguns indicadores de confiabilidade são:

- Os publicadores dos dados estão relacionados a organizações governamentais, associações e instituições reconhecidas no domínio?
- Os conjuntos de dados são utilizados por outros conjuntos *Linked Data* do mesmo domínio?
- O conjunto de dados se destaca por ser amplamente adotado no contexto do *Linked Data*?
- O conjunto possui bons níveis de interligação com fontes relevantes do domínio?
- Os publicadores fornecem informações de proveniência e metadados estruturados a respeito dos dados?
- Os publicadores disponibilizam informações a respeito da política de publicação e curadoria dos dados?

O estabelecimento do peso desses aspectos depende da abordagem e dos objetivos estabelecidos para a seleção.

Apresentados os fluxos da seleção de fonte de dados para ligação, os modelos de protocolo que podem auxiliar no planejamento desse processo e uma *checklist* composta apenas por aspectos que podem levar a exclusão de fontes em potencial, a próxima seção apresenta as considerações finais da presente tese.

8 CONSIDERAÇÕES FINAIS

A presente tese foi conduzida com base no objetivo de fomentar maior clareza teórica e terminológica e maior compreensão a respeito da seleção de dados *Linked Data*, por meio de sua relação com o processo de avaliação de qualidade, abordando definições, ferramentas e produtos relacionados. Mapeando os agentes, etapas e atividades da seleção de dados *Linked Data* para criação de *links* com fontes externas.

Para atender a esse objetivo, partiu-se da construção de um *corpus* teórico e documental exaustivo, auxiliando na identificação e discussão dos diferentes termos necessários para a compreensão da qualidade de dados e da estrutura dos dados *Linked Data, visando* compreender se e como essa estrutura afeta a avaliação de qualidade de dados.

O objetivo específico “Discutir a qualidade de dados enquanto objeto de estudo da Ciência da Informação” foi atingido na seção 3, onde foram apresentados os principais conceitos necessários para a compreensão da qualidade de dados e discutida a sua relação com a Ciência da Informação.

Com base na análise conjunta das categorias “5 - termos relacionados às áreas e subáreas da Ciência da Informação”; “6 - instrumentos utilizados no processo de avaliação de qualidade”; e “7 - termos relacionados aos temas e enfoques dos documentos”, é possível inferir que enquanto um objeto de estudo da Ciência da Informação, a qualidade de dados é frequentemente abordada em sua relação com a Organização e Representação da Informação e do Conhecimento.

A Ciência da Informação possui desde sua estruturação a expertise no desenvolvimento de instrumentos que visam padronizar e estruturar conjuntos de dados, tais como vocabulários, códigos de catalogação e padrões de metadados. Além disso, dados são o produto de processos importantes da área, especialmente no âmbito da Organização e Representação da Informação e do Conhecimento.

Observa-se um destaque para as abordagens de qualidade de dados no âmbito da governança de dados, onde se discute a sua inclusão como etapa importante do ciclo de vida de dados. Nesse aspecto, destacam-se os instrumentos criados pela CI para auxiliar na qualidade de dados, por meio da criação, adoção e verificação da adequação à melhores práticas, princípios e políticas voltadas para a qualidade dos dados.

Destacam-se ainda instrumentos e produtos derivados dessa área, tais como metadados, padrões de metadados, vocabulários controlados e padrões de catalogação.

Ressaltam-se as abordagens da qualidade de dados relacionada a distintos aspectos da curadoria e da preservação digital. Nessas abordagens, processos e instrumentos da qualidade de dados são inseridos como ações realizadas no âmbito da curadoria digital visando garantir o tratamento e a manutenção dos recursos informacionais digitais e ainda a preservação desses recursos a longo prazo.

A qualidade de dados e a Ciência da Informação possuem, portanto, uma relação interdisciplinar, com potencial de mútua colaboração entre as áreas. A CI pode contribuir para a melhoria da qualidade dos dados por meio da elaboração e aplicação de instrumentos, tais como diretrizes, códigos e melhores práticas. A CI pode utilizar de técnicas, processos e instrumentos da qualidade de dados para a melhoria de seus produtos.

O objetivo específico de “apresentar o estado da arte da qualidade de dados *Linked Data” foi* atingido na seção 4, com a apresentação dos resultados do estudo teórico da qualidade de dados *Linked Data.*

Nessa seção, buscou-se um aprofundamento sobre como os problemas de qualidade afetam dados *Linked Data*. Observou-se que esses problemas podem ser organizados em 3 categorias: problemas relacionados com a estrutura dos dados, problemas relacionados com as fontes dos dados e problemas relacionados com o próprio processo de avaliação. Os problemas relacionados à estrutura normalmente são ligados ao uso de URIs, a adequação ao modelo RDF, a boa aplicação de vocabulários.

A dificuldade em selecionar fonte de dados também foi destacada entre as problemáticas apontadas pela literatura, reforçada pela heterogeneidade das fontes de dados, já que os dados são provenientes de contextos diversos, com diferentes níveis de curadoria e preocupação com a acurácia, precisão e veracidade.

Além dos problemas relacionados com a estrutura e com as fontes dos dados, foram apontados problemas relacionados com o próprio processo de avaliação de qualidade desse tipo de dado, como o volume e a variedade, as discrepâncias entre as necessidades dos usuários, que dificultam a criação de ferramentas genéricas, a variedade dessas ferramentas e ainda a ausência de descrição adequada e formal dos conjuntos de dados.

Ainda como um dos desafios do processo de avaliação, destaca-se a heterogeneidade dos artefatos disponíveis para auxiliar no processo de avaliação de qualidade, como apresentado na subseção 4.4.

O próprio processo de seleção desses artefatos se torna um desafio particular, já que possuem estruturas, objetivos e funcionalidade muito distintas e nem sempre possuem interfaces amigáveis para os usuários, estando sujeitos a indisponibilidade e desatualização.

Os objetivos específicos: “Elaborar a Árvore de Domínio da qualidade de dados *Linked Data*, e “Elaborar um glossário de qualidade de dados *Linked Data*” foram atingidos na seção 5, que apresenta os resultados do estudo terminológico da qualidade de dados *Linked Data*. Foi estabelecida a Árvore de Domínio, permitindo a identificação dos principais conceitos e as relações hierárquicas existentes entre eles. A construção da Árvore de Domínio foi pautada nos resultados do estudo teórico da qualidade de dados e na análise da documentação do W3C.

Com base na Árvore de Domínio foram identificados os termos que compõe o glossário. Em seguida, foram construídas e apresentadas as definições para cada um desses termos, com base na fragmentação de definições, seguindo o método da Grade.

Observa-se que o termo qualidade de dados possui uma abordagem poliédrica, com características interdisciplinares e multidimensionais, que quando aplicada a dados *Linked Data*, possui características próprias, relacionadas com a estrutura, com o comportamento da comunidade de dados *Linked Data*, possuindo problemas de qualidade específicos e ferramentas desenvolvidas para esse domínio.

Com base na estrutura da Árvore de Domínio, destaca-se que embora se utilize da estrutura clássica da qualidade de dados, organizando muitas vezes as discussões em dimensões, critérios e métricas, o contexto do *Linked Data* influência na forma como são compreendidas as categorias e na forma como são mensurados os critérios de qualidade dos dados.

Para atingir o objetivo específico de “apresentar um fluxo para seleção de dados *Linked Data* para interligação*”* tornou-se necessária a realização do estudo das etapas, procedimentos e ferramentas que perpassam o processo de seleção. Essa discussão foi apresentada na seção 6, onde foram abordados os resultados do estudo processual da qualidade de dados *Linked Data*, com foco na seleção de fontes para criação de *links* com fontes externas.

Partiu-se da discussão dos Ciclos de Vida dos Dados, buscando situar a seleção de fontes em seu contexto. Nessa discussão, observou-se que a avaliação de qualidade é uma atividade que permeia todo o ciclo de vida dos dados, embora com objetivos diferentes.

Nota-se que a atividade de seleção de fontes pode ser inserida na etapa de coleta dos dados em CVDs gerais, enquanto no contexto específico do *Linked Data* ela está relacionada a etapa de interligação.

Considerando a importância da etapa de interligação para a seleção de fontes, discutiu-se então como ela ocorre no *Linked Data*. Em relação a esse aspecto, os *links* no *Linked Data* ocorrem quando o valor da tripla é um URI, os *links* podendo ser internos ou externos. Os *links* externos acontecem quando o URI do valor da tripla pertence a um conjunto de dados externo.

Existem diferentes relações possíveis entre as entidades interligadas, e a criação de cada *link* depende da identificação de uma propriedade apropriada, preferencialmente em um vocabulário bem estabelecido do domínio de origem, que represente o tipo de relação existente entre as entidades relacionadas ao objeto da interligação.

Em relação a análise dos processos relacionados com a avaliação de qualidade destaca-se que a avaliação pode ser realizada com os propósitos de: seleção de fontes, controle de qualidade dos dados e melhoria da qualidade dos dados. Observou-se ainda que a avaliação de qualidade perpassa todas essas etapas.

A etapa de planejamento possuí forte impacto na avaliação de qualidade, tendo como objetivo identificar as necessidades informacionais, os objetivos da avaliação e estabelecer os níveis de qualidade esperados para os dados.

O principal produto da etapa de planejamento é o modelo de qualidade, um instrumento que é construído com base nos objetivos estabelecidos e que irá guiar a condução de todo o processo. O modelo é organizado em categorias, dimensões, critérios e métricas. Para aplicar esse modelo, podem ser utilizados padrões-ouro e listas de verificação, bem como ferramentas automáticas e semiautomáticas. Na seleção de fontes, o modelo de qualidade precisa ainda estabelecer os requisitos mínimos de qualidade esperados ou critérios de inclusão e exclusão.

Um dos principais desafios em relação a avaliação de qualidade é o estabelecimento do modelo de qualidade. Esse estabelecimento envolve a seleção das categorias, dimensões e métricas a serem adotadas para compor esse modelo

em abordagens contextuais da avaliação. Foram identificadas e sistematizadas diferentes abordagens para auxiliar no estabelecimento de modelos de qualidade.

Considerando a importância do modelo de qualidade, foram discutidos os modelos de qualidade de dados *Linked Data* identificados, abordando sua estrutura e principais aspectos individualizadores.

Discutiu-se ainda como podem ser explorados os vocabulários na avaliação de qualidade, observando que eles podem facilitar a obtenção de informações relevantes, como a descrição do seu conteúdo, informações relacionadas a proveniência, temporalidade, licença e formas de acesso.

Nesse contexto, os vocabulários tornam-se fundamentais para o processo de interligação, uma vez que permitem rotular e explicitar o tipo de relação existente entre as entidades conectadas.

Os vocabulários estão diretamente relacionados a estrutura sintática e semântica dos dados, permitindo a verificação da forma como são aplicadas as propriedades, fornecendo informações sobre a acurácia e consistência. Além disso, o uso de vocabulários bem estabelecidos no domínio pode ser considerado um indicativo de boa qualidade, estando ligado a confiabilidade dos dados, tendo impacto nos aspectos de interoperabilidade.

Os vocabulários podem ainda ser utilizados para compartilhar e permitir o reuso dos resultados do processo de avaliação, como exemplificado pelo vocabulário DQV.

Compreendendo a importância e relevância dos vocabulários nesse contexto, conclui-se que esses ainda são pouco explorados e discutidos pela comunidade de dados *Linked Data*. Observa-se que os metadados necessários para descrever os dados muitas vezes são negligenciados ou sub representados.

Essa pouca abordagem pode ser observada inclusive na estrutura da qualidade de dados, que não apresenta categorias ou dimensões focadas na descrição adequada dos dados. Em alguns modelos pontuais são acrescentadas dimensões ou critérios com o propósito de avaliar esse aspecto, mas esses geralmente limitam-se apenas a verificar a presença ou ausência desses metadados, e são espalhados ao longo de diferentes categorias, sem um padrão estabelecido.

Com base no estudo teórico, terminológico e processual da qualidade de dados *Linked Data* foi possível estabelecer o fluxo da seleção de fontes para interligação, respondendo à pergunta de pesquisa: como selecionar dados *Linked Data* para criação de *links* com fontes externas?

Conclui-se que a seleção de fontes para criação de *links* pode ser dividida em três etapas: planejamento, seleção e interligação. Sendo a etapa de planejamento determinante para a boa condução das demais etapas.

A condução do planejamento depende dos objetivos que levam a busca por fontes, sendo identificadas duas abordagens principais: baseada em explicitação de necessidade informacional previamente estabelecia e baseada em análise exploratória de fontes. Foram estabelecidos fluxos para cada uma dessas abordagens e apresentados modelos de protocolos para a sua condução.

Ao final, o estabelecimento de um modelo de qualidade único para as diferentes possíveis abordagens demonstrou-se pouco funcional, considerando os objetivos e domínios que circundam os dados *Linked Data*, devido à grande heterogeneidade que permeia a comunidade. Optou-se assim pela proposta de uma *checklist* considerando os principais critérios de exclusão, baseada principalmente nas dimensões acessibilidade, confiabilidade, relevância e pertinência dos dados.

Nesse sentido, com base nas discussões realizadas e nas lacunas identificadas, propõe-se como estudos futuros:

- Aprofundamento a respeito do uso de vocabulários no processo de avaliação de qualidade de dados, identificando formas de ampliar a sua utilização nesse contexto;
- Análise da aplicabilidade das ferramentas identificadas em estudos de caso direcionados;
- Condução de estudo de caso para validação das ferramentas propostas na presente tese (fluxo, modelo de protocolo e *checklist*);
- Aprofundamento a respeito das dimensões e métricas existentes para avaliação dos metadados e dos aspectos descritivos dos conjuntos de dados;
- Análise da viabilidade de criação de uma nova categoria de qualidade para reunir os aspectos relacionados à descrição, organização e dos metadados dos conjuntos de dados (categoria descritiva).

A seleção de dados para interligação é um processo complexo e amplamente contextual, cujo sucesso depende da etapa de planejamento, do estabelecimento de um modelo de qualidade, da abordagem a ser adotada e dos objetivos a serem alcançados e que pode ser facilitado pela adoção de diferentes ferramentas.

Nesse sentido, ressalta-se o ineditismo das proposições apresentadas na presente tese, uma vez identificada inexistência de um glossário terminológico especializado na qualidade de dados *Linked Data*, de fluxos processuais claros para a seleção desses dados e lacunas nas metodologias e ferramentas disponíveis para auxiliar nesse processo.

Conclui-se, portanto, que os produtos propostos na presente tese, assim como as discussões apresentadas, preenchem essa lacuna, fomentando maior clareza teórica e terminológica. Fomentam ainda maior compreensão a respeito da seleção de dados *Linked Data,* ao sistematizar as etapas, atividades e instrumentos necessários para a realização da seleção de fontes de dados para criação de *links*.

REFERÊNCIAS

ABIÁN, David *et al*. Using contemporary constraints to ensure data consistency. In: ACM/SIGAPP SYMPOSIUM ON APPLIED COMPUTING, 34., 2019, [s.l.] **Anais [...].** [s.l.]: Acm/Sigapp, 2019. p. 2303-2310. Disponível em: https://doi.org/10.1145/3297280.3297509. Acesso em: 23 out. 2024.

ABIÁN, David. *et al*. Wikidata and DBpedia: a comparative study. **Lecture Notes In Computer Science**, [s.l.], v. 10546, n. 1, p. 142-154, 2018. Disponível em: http://dx.doi.org/10.1007/978-3-319-74497-1_14. Acesso em: 11 nov. 2024.

ABNT NBR ISO. **9000**: Sistemas de gestão da qualidade — Fundamentos e vocabulário. Rio de Janeiro: Abnt, 2015.

ACOSTA, Maribel *et al*. Crowdsourcing linked data quality assessment. **Lecture Notes In Computer Science**, [s.l.], p. 260-276, 2013. Disponível em: 10.1007/978-3-642-41338-4_17. Acesso em: 23 out. 2024.

ACOSTA, Maribel *et al*. Detecting linked data quality issues via crowdsourcing: a dbpedia study. **Semantic Web**, [s.l.], v. 9, n. 3, p. 303-335, 12 abr. 2018. Disponível em: https://doi.org/10.3233/SW-160239. Acesso em: 23 out. 2024.

AHMED, Hana Haj *et al*. Data quality assessment in the integration process of linked open data (LOD). **2017 Ieee/Acs 14Th International Conference on Computer Systems and Applications (Aiccsa)**, [s.l.], v. 628, p. 1-6, out. 2017. Disponível em: 10.1109/AICCSA.2017.178. Acesso em: 23 out. 2024.

ALBERTONI, Riccardo *et al*. Quality measures for skos. **Data Technologies And Applications**, [s.l.], v. 52, n. 3, p. 405-423, 1 ago. 2018. Disponível em: http://dx.doi.org/10.1108/dta-05-2017-0037. Acesso em: 23 out. 2024.

ALBERTONI, Riccardo; MARTINO, Monica de; QUARATI, Alfonso. Documenting context-based quality assessment of controlled vocabularies. **Ieee Transactions on Emerging Topics In Computing**, [s.l.], v. 9, n. 1, p. 144-160, 1 jan. 2021. Disponível em: 10.1109/TETC.2018.2865094. Acesso em: 11 nov. 2024.

ALBERTONI, Riccardo; MARTINO, Monica de; QUARATI, Alfonso. Linked thesauri quality assessment and documentation for big data discovery. **2017 International Conference On High Performance Computing & Simulation (Hpcs)**, [s.l.], p. 37-44, jul. 2017. Disponível em: http://dx.doi.org/10.1109/hpcs.2017.16. Acesso em: 23 out. 2024.

ALMEIDA, Cleibson Aparecido de *et al*. Melhoria na qualidade de dados com a aplicação de. **Atoz**: novas práticas em informação e conhecimento, [S.L.], v. 5, n. 2, p. 72, 2016. Disponível em: http://dx.doi.org/10.5380/atoz.v5i2.47303. Acesso em: 19 dez. 2023.

ANCIB. **Coordenações e ementas de GT**. 2024. Disponível em: https://ancib.org/sobre/. Acesso em: 27 nov. 2024.

ARAKAKI, Ana Carolina Simionato *et al*. Convergência entre organização da informação, linked data e inteligência artificial. In: ISKO BRASIL, 8., 2025, Canela. **Anais [...].** Canela: ISKO, 2025. p. 1-8.

ARAKAKI, Felipe Augusto. **Linked data: ligação de dados bibliográficos.** 2016. 144 f. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2016. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/147979/arakaki_fa_me_mar.pdf?sequence=2&isAllowed=y. Acesso em: 20 de nov. 2025

ARAKAKI, Felipe Augusto; SIMIONATO, Ana Carolina; SANTOS, Plácida Leopoldina Ventura Amorim da Costa. Catalogação e tecnologia: interseções com a web semântica. **Informação@profissões**, [s.l.], v. 6, n. 2, p. 03-19, maio 2018. Disponível em: http://dx.doi.org/10.5433/2317-4390.2017v6n2p03. Acesso em: 20 de nov. 2025

ARRUDA, Narciso *et al*. A Fuzzy Approach for data quality assessment of linked datasets. **Proceedings of The 21St International Conference on Enterprise Information Systems**, [s.l.], p. 399-406, 2019. Disponível em: http://dx.doi.org/10.5220/0007718803990406. Acesso em: 23 out. 2024.

ASSAF, Ahmad; SENART, Aline; TRONCY, Raphaël. Roomba: automatic validation, correction and generation of dataset metadata. In: THEWEBCONF, 24., 2015a, [S.L.]. **Anais [...] .** [s.l.]: Thewebconf, 2015a. p. 159-162. Disponível em: https://dl.acm.org/doi/abs/10.1145/2740908.2742827. Acesso em: 23 out. 2024.

ASSAF, Ahmad; SENART, Aline; TRONCY, Raphaël. Towards an objective assessment framework for linked data quality. **International Journal On Semantic Web and Information Systems**, [s.l.], v. 12, n. 3, p. 111-133, 1 jul. 2016b. Disponível em: http://dx.doi.org/10.4018/IJSWIS.2016070104. Acesso em: 23 out. 24

ASSAF, Ahmad; TRONCY, Raphaël; SENART, Aline. Roomba: an extensible framework to validate and build dataset profiles. In: ESWC SATELLITE EVENTS, 1., 2015b, [S.L]. **Anais [...] .** [s.l]: Eswc, 2016a. v. 1, p. 325-339. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-25639-9_46. Acesso em: 23 out. 2024.

ASSAF, Ahmad; TRONCY, Raphaël; SENART, Aline. What's up LOD cloud? **Lecture Notes in Computer Science**, [s.l.], p. 247-254, 2015c. Disponível em: http://dx.doi.org/10.1007/978-3-319-25639-9_40. Acesso em: 23 out. 2024.

AUER, Sören *et al*. Managing the life-cycle of linked data with the LOD2 stack. **Lecture Notes in Computer Science**, [s.l.], v. 2, n. 1, p. 1-16, 2012. Disponível em: http://dx.doi.org/10.1007/978-3-642-35173-0_1. Acesso em: 11 nov. 2025.

BAILEY, Martha J. *et al*. How well do automated linking methods perform? Lessons from US Historical Data. **Journal Of Economic Literature**, [s.l.], v. 58, n. 4, p. 997-1044, 1 dez. 2020. Disponível em: http://dx.doi.org/10.1257/jel.20191526. Acesso em: 11 nov. 2024.

BARATA, André Montoia. **Governança de dados em organizações brasileiras**: uma avaliação comparativa entre os benefícios previstos na literatura e os obtidos pelas organizações. 2015. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2015.

BARITÉ, Mario. **La garantía literaria como herramienta de revisión de sistemas de organización del conocimiento**: modelo y aplicación. 2011. 382 f. Tese (Doutorado) - Curso de Doctorado En, Facultad de Comunicación y Documentación, Granada, 2011.

BARITÉ, Mario; RAUCH, Mirtha. Propuesta metodológica para la elaboración de definiciones terminológicas. In: RITERM, 1., 2006, Montevideo. **Anais [...].** Montevideo: Riterm, 2006. p.1-11.

BASILI, Victor R.; CALDIERA, Gianluigi; ROMBACH, Dieter. Goal question metric paradigm. **Ncyclopedia Of Software Engineering**, [s. l], n. 2, p. 1-6, jan. 1994.

BATINI, Carlo; SCANNAPIECO, Monica. **Data quality and Information quality**: dimensions, principles and techniques. [s.l.]: Springer, 2016. 520 p.

BATINI, Carlo; SCANNAPIECO, Monica. **Data quality**: concepts, methodologies and techniques. [s.l.]: Springer, 2006. 276 p.

BAX, Marcello Peixoto. Design science: filosofia da pesquisa em ciência da informação e tecnologia. **Ciência da Informação**, [s.l.], v. 42, n. 2, p. 298-312, 6 ago. 2015. Disponível em: https://doi.org/10.18225/ci.inf.v42i2.1388. Acesso em: 09 dez. 2024.

BEEK, Wouter *et al*. Literally better: analyzing and improving the quality of literals. **Semantic Web**, [s.l.], v. 9, n. 1, p. 131-150, 30 nov. 2017. Disponível em: http://dx.doi.org/10.3233/sw-170288. Acesso em: 23 out. 2024.

BEHKAMAL, Behshid *et al*. Data accuracy: what does it mean to lod?. **2014 4Th International Conference On Computer And Knowledge Engineering (Iccke)**, [s.l.], p. 80-85, out. 2014. Disponível em: http://dx.doi.org/10.1109/iccke.2014.6993457. Acesso em: 11 nov. 2024.

BENTANCOURT, Silvia Maria Puentes; ROCHA, Rafael Port da. Metadados de qualidade e visibilidade na comunicação científica. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, [s.l.], p. 82-101, 12 dez. 2012. Disponível em: https://doi.org/10.5007/1518-2924.2012v17nesp2p82. Acesso em: 19 dez. 2023.

BERNERS-LEE, T. Linked data, 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 08 set. 2020.

BIZER, Christian; HEATH, T.; BERNERS-LEE, Tim. Linked data: the story so far. International Journal on Semantic Web and Information Systems, v. 5, n. 3, p. 1-22, 2009. Disponível em: eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf. Acesso em: 20 abr. 2023.

BLUMENTHAL, Sherman. **Management information systems**: a framework for planning and development. [s.l]: Prentice-Hall, 1969. 219 p.

BONNER, Stephen *et al*. Data quality assessment and anomaly detection via map/reduce and *Linked Data*: a case study in the medical domain. In:

BOTEGA, Leonardo Castro, et. al.QualityAware Human-Driven Information Fusion Model. In: 20th International Conference on Information Fusion, 2017, Xian. 20th International Conference on Information Fusion. 2017

BRASIL. Lei nº 13709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais (Lgpd)**. Brasília, Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 04 dez. 2025.

BRAŞOVEANU, Adrian *et al*. Framing named entity linking error types. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 11., 2018, [s.l.]. **Proceedings [...] .** [s.l]: Lrec, 2018. p. 266-271. Disponível em: https://preview.aclanthology.org/ingestion-script-update/L18-1040.pdf. Acesso em: 11 nov. 2024.

BURCH, John G.; GRUDNITSKI, Gary. **Information Systems**: theory and practice. [s.l.]: John Wiley & Sons, 1986. 674 p.

CABRÉ, María Teresa. **La terminología: representación y comunicación:** elementos para una teoría de base comunicacativa y otros artículos. Barcelona: Universitat Pompeu Fabra, 1999.

CANDELA, Gustavo *et al*. Evaluating the quality of linked open data in digital libraries. **Journal Of Information Science**, [s.l.], v. 48, n. 1, p. 21-43, 3 ago. 2020. Disponível em: http://dx.doi.org/10.1177/0165551520930951. Acesso em: 11 nov. 2024.

CAPPIELLO, Cinzia *et al*. A Quality model for linked data exploration. **Lecture Notes In Computer Science**, [s.l.], p. 397-404, 2016. Disponível em: http://dx.doi.org/10.1007/978-3-319-38791-8_25. Acesso em: 23 out. 2024.

CATALÁ, Sara Álvarez; BARITÉ, Mario (org.). **Teoría y praxis en terminología**. Montevideo: Universidad de La República., 2016. 226 p.

CATANIA, Barbara; GUERRINI, Giovanna; YAMAN, Beyza. Exploiting context and quality for *linked data* source selection. **Proceedings Of The 34Th Acm/Sigapp Symposium On Applied Computing**, [s.l.], v. 1409, p. 2251-2258, 8 abr. 2019. Disponível em: http://dx.doi.org/10.1145/3297280.3297503. Acesso em: 23 out. 2024.

CHERIX, Didier *et al*. Lessons learned — the case of CROCUS: cluster-based ontology data cleansing. **Lecture Notes In Computer Science**, [s.l.], p. 14-24, 2014. Disponível em: http://dx.doi.org/10.1007/978-3-319-11955-7_2. Acesso em: 11 nov. 2024.

COELHO JÚNIOR, Abeil. **Qualidade de dados em acervos do patrimônio cultural**: uma proposta diagnóstica semiautomática para objetos culturais sob gestão do instituto brasileiro de museus. 2021. 153 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Federal do Espírito Santo (Ppgci/Ufes), Vitória, 2021. Disponível em: https://sappg.ufes.br/tese_drupal/tese_15610_Dissertacao%20Revisada%20Final%20-%20Dirceu%20Flavio%20Macedo%20%282%29.pdf. Acesso em: 19 nov. 2024.

COELHO JÚNIOR, Abeil; LEMOS, Daniela Lucas da Silva. Qualidade de dados em acervos museais: uma avaliação semiautomática para os acervos sob gestão do ibram. In: ENCONTRO NACIONAL DE PESQUISA EM CIêNCIA DA INFORMAÇÃO, 22., 2022a, Porto Alegre. **Anais** [...]. Porto Alegre: Ancib, 2022. p. 1-11. Disponível em: https://brapci.inf.br/index.php/res/v/202046. Acesso em: 19 dez. 2023.

COELHO JÚNIOR, Abeil; LEMOS, Daniela Lucas da Silva. TRATAMENTO DA INFORMAÇÃO EM ACERVOS CULTURAIS: avaliação do uso de vocabulários controlados em coleções museológicas sob gestão do instituto brasileiro de museus. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 5., 2022b, Vitória. **Anais [...].** Vitória: Ufes, 2023. p. 235-241. Disponível em: https://widat2022.ufes.br/wp-content/uploads/2023/04/widat-2022-anais.pdf. Acesso em: 19 dez. 2023.

CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. **Dicionário de Biblioteconomia e Arquivologia**. Brasília: Briquet de Lemos, 2008. xvi, 451 p.

DAMA INTERNATIONAL. **DAMA-DMBOK**: data management body of knowledge. [s.l.]: Technics Publications, 2015.

DARARI, Fariz; PRASOJO, Radityo Eko; NUTT, Werner. CORNER: a completeness reasoner for sparql queries over rdf data sources. Lecture Notes In **Computer Science**, [s.l.], v. 8798, p. 310-314, 2014. Disponível em: http://dx.doi.org/10.1007/978-3-319-11955-7_40. Acesso em: 11 nov. 2024.

DAVIS, Charles H.; RUSH, James. **Guide to Information Science.** [s.l.]: Greenwood, 1979. 305 p.

DEBATTISTA, Jeremy *et al*. Evaluating the quality of the LOD cloud: an empirical investigation. **Semantic Web**, [s.l.], v. 9, n. 6, p. 859-901, 12 set. 2018. Disponível em: http://dx.doi.org/10.3233/sw-180306. Acesso em: 23 out. 2024.

DEBATTISTA, Jeremy *et al*. Quality Assessment of linked datasets using Probabilistic Approximation. **Eswc**, [s.l.], v. 1, n. 1, p. 1-15, 2015. Disponível em: http://dx.doi.org/10.48550/ARXIV.1503.05157. Acesso em: 23 out. 2024.

DEBATTISTA, Jeremy; AUER, Soren; LANGE, Christoph. Luzzu -- A Framework for linked data quality assessment. **2016 Ieee Tenth International Conference on Semantic Computing (ICSC)**, [s.l.], v. 35, p. 124-131, fev. 2016. Disponível em: 10.1109/ICSC.2016.48. Acesso em: 23 out. 24.

DIAS, Cláudia Augusto. Hipertexto: resumo histórico e efeitos sociais. Ci. Inf. Brasília, v. 28, n. 3, p. 269-277, dez 1999. Disponível em: https://www.scielo.br/j/ci/a/WB4h7bH3yM3YM89Z4JhjdVs/. Acesso em: 15 jul. 2020

DIAS, Guilherme Ataíde *et al*. Garbage in, garbage out (GIGO): enfrentando esta máxima nos conjuntos de dados associados ao programa dinheiro direto na escola (pdde). In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 5., 2023, Brasilía. **Anais [...] .** Brasilia: Ibict, 2023. p. 1-11. Disponível em: https://labcotec.ibict.br/widat/index.php/widat2023/article/view/32/60. Acesso em: 05 dez. 2024.

DIMOU, Anastasia *et al*. Assessing and refining mappings to RDF to Improve dataset quality. In: THE SEMANTIC WEB - ISWC, 1., 2015, [S.L.]. **Anais [...].** [s.l.]: Iswc, 2015. v. 9367, p. 133-149. Disponível em: https://doi.org/10.1007/978-3-319-25010-6_8. Acesso em: 11 nov. 2024.

DORN, P.H. **Business information in the eighties**. In: PAPPENHEIM, Business information systems. Maidenhead: Pergamon Infotech,1981. p. 245-260.

DUBLIN CORE. **Dublin Core**. 2025. Disponível em: https://www.dublincore.org/resources/glossary/dublin_core/. Acesso em: 11 nov. 2025.

ELBATTAH, Mahmoud; RYAN, Conor. Learning triple sequence patterns in knowledge graphs to predict inconsistencies. **Proceedings Of The 11Th International Joint Conference on Knowledge Discovery, Knowledge Engineering And Knowledge Management**, [s.l.], p. 435-441, jan. 2019.

ESPÍNDOLA, Priscilla Lüdtke *et al*. Governança de dados aplicada à ciência da informação. **Rdbci Revista Digital de Biblioteconomia e Ciência da Informação**, [s.l.], v. 16, n. 3, p. 274-298, 16 ago. 2018. Disponível em: https://doi.org/10.20396/rdbci.v16i3.8651080. Acesso em: 11 nov. 2024.

ESTEVES, Diego *et al*. Toward veracity assessment in RDF knowledge bases. **Journal Of Data and Information Quality**, [s.l.], v. 9, n. 3, p. 1-26, 30 set. 2017. Disponível em: http://dx.doi.org/10.1145/3177873. Acesso em: 23 out. 2024.

FAGUNDES, Priscila Basto; MACEDO, Douglas Dyllon Jeronimo de;

FAPESP. **Gestão de dados**. 2024. Disponível em: https://fapesp.br/gestaodedados. Acesso em: 22 nov. 2024.

FÄRBER, Michael *et al*. *Linked data* quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. **Semantic Web**, [s.l.], v. 9, n. 1, p. 77-129, 30 nov. 2017. Disponível em: https://doi.org/10.3233/SW-170275. Acesso em: 23 out. 2024.

FEENEY, Kevin Chekov *et al*. Improving curated web-data quality with structured harvesting and assessment. **International Journal on Semantic Web and Information Systems**, [s.l.], v. 10, n. 2, p. 35-62, 1 abr. 2014. Disponível em: http://dx.doi.org/10.4018/ijswis.2014040103. Acesso em: 23 out. 2024.

FEENEY, Kevin *et al*. The dacura data curation system. **IFIP Advances in Information and Communication Technology**, [s.l.], v. 482, n. 1, p. 15-20, 2016. Springer International Publishing. Disponível em: http://dx.doi.org/10.1007/978-3-319-46224-0_2. Acesso em: 11 nov. 2024.

FEENEY, Kevin; GLEASON, Gavin Mendel; BRENNAN, Rob. *Linked Data* schemata: fixing unsound foundations. **Semantic Web**, [s.l.], v. 9, n. 1, p. 53-75, 30 nov. 2017. Disponível em: 10.3233/SW-170271. Acesso em: 23 out. 2024.

FELIZARDO, Katia Romero *et al*. **Revisão Sistemática da Literatura em Engenharia de Software**: teoria e prática. [s.l]: Gen Ltc, 2017. 144 p.

FLEISCHHACKER, Daniel *et al*. Detecting errors in numerical linked data using cross-checked outlier detection. **Lecture Notes In Computer Science**, [s.l.], p. 357-372, 2014. Disponível em: http://dx.doi.org/10.1007/978-3-319-11964-9_23. Acesso em: 23 out. 2024.

FONT, Ludovic; ZOUAQ, Amal; GAGNON, Michel. Assessing the quality of domain concepts descriptions in dbpedia. In: INTERNATIONAL CONFERENCE ON SIGNAL-IMAGE TECHNOLOGY & INTERNET-BASED SYSTEMS (SITIS), 11., 2015, [s.l]. **Anais [...] .** [S.L.]: Sitis, 2015. v. 1, p. 254-261. Disponível em: https://ieeexplore.ieee.org/abstract/document/7400574. Acesso em: 23 out. 2024.

FOX, Christopher; LEVITIN, Anany; REDMAN, Thomas. The notion of data and its quality dimensions. **Information Processing & Management**, [s.l.], v. 30, n. 1, p. 9-19, jan. 1994. Disponível em: http://dx.doi.org/10.1016/0306-4573(94)90020-5. Acesso em: 26 nov. 2024.

FREUND, Gislaine Parra. A produção científica sobre qualidade de dados em big data: um estudo na base de dados web of science**. Rdbci: Revista Digital de Biblioteconomia e Ciência da Informação**, [s.l.], v. 16, n. 1, p. 194, 9 nov. 2017. Disponível em: https://doi.org/10.20396/rdbci.v16i1.8650412. Acesso em: 19 dez. 2023.

FRY, James; SIBLEY, Edgar. Evolution of data-base management systems. **Computing Surveys**, [s.l], v. 8, n. 1, p. 8-42, mar. 1976.

FÜRBER, Christian; HEPP, Martin. Using SPARQL and SPIN for data quality management on the semantic web. **Lecture Notes in Business Information Processing**, [s.l.], p. 35-46, 2010. Disponível em: http://dx.doi.org/10.1007/978-3-642-12814-1_4. Acesso em: 23 out. 2024.

GAMBLE, Matthew *et al*. MIM: a minimum information model vocabulary and framework for scientific *Linked Data*. **2012 Ieee 8Th International Conference On E-Science**, [s.l.], p. 1-8, out. 2012. Disponível em: http://dx.doi.org/10.1109/escience.2012.6404489. Acesso em: 23 out. 2024.

GUALDANI, Fabrício Amadeu *et al*. Critérios de qualidade de dados em saúde: uma análise quantitativa. **Informação & Informação**, Londrina, v. 27, n. 2, p. 466–490, 2022. Disponível em: https:// ojs.uel.br/revistas/uel/index.php/informacao/article/view/43782. Acesso em: 20 dez. 2023.

GÜRDÜR, Didem; EL-KHOURY, Jad; NYBERG, Mattias. Methodology for linked enterprise data quality assessment through information visualizations. **Journal Of Industrial Information Integration**, [s.l.], v. 15, p. 191-200, set. 2019. Disponível em: https://doi.org/10.1016/j.jii.2018.11.002. Acesso em: 23 out. 2024.

HADHIATMA, A. Improving data quality in the linked open data: a survey. **Journal Of Physics**: Conference Series, [s.l.], v. 978, p. 012026, mar. 2018. Disponível em: 10.1088/1742-6596/978/1/012026. Acesso em: 23 out. 2024.

HALLER, Armin *et al*. What are links in linked open data? A characterization and evaluation of links between knowledge graphs on the web. **Journal of Data and Information Quality**, [s.l.], v. 12, n. 2, p. 1-34, 6 maio 2020. Disponível em: https://doi.org/10.1145/3369875. Acesso em: 23 out. 2023.

HANLON, Rolando *et al*. Towards an effective user interface for data exploration, data quality assessment and data integration. **2021 Ieee 15Th International Conference On Semantic Computing (Icsc)**, [s.l.], v. 7, p. 431-436, jan. 2021. Disponível em: http://dx.doi.org/10.1109/icsc50631.2021.00077. Acesso em: 23 out. 2024.

HARPER, Gillian. Linkage of maternity hospital episode statistics data to birth registration and notification records for births in england 2005–2014: quality assurance of linkage of routine data for singleton and multiple births. **Bmj Open**, [s.l.], v. 8, n. 3, p. 1-13, mar. 2018. Disponível em: http://dx.doi.org/10.1136/bmjopen-2017-017898. Acesso em: 18 nov. 2024.

HASSAN, Umair Ul *et al*. ACRyLIQ: leveraging dbpedia for adaptive crowdsourcing in *Linked Data* quality assessment. **Lecture Notes In Computer Science**, [s.l.], p. 681-696, 2016. Disponível em: http://dx.doi.org/10.1007/978-3-319-49004-5_44. Acesso em: 23 out. 2024.

HEALY, Mark. *et al*. The accuracy of chemotherapy ascertainment among colorectal cancer patients in the surveillance, epidemiology, and end results registry program. **Bmc Cancer**, [s.l.], v. 18, n. 1, p. 1-12, 27 abr. 2018. Disponível em: http://dx.doi.org/10.1186/s12885-018-4405-7. Acesso em: 11 nov. 2024.

HEATH, Tom Talis; BIZER, Christian Freie. **Linked data: evolving the web into a global data space**. Berlim: Morgan & Claypool, 2011. 136 p. Disponível em: http://linkeddatabook.com/editions/1.0/#htoc8. Acesso em: 10 de set. 2020.

HELING, Lars. Quality-driven query processing over federated RDF data sources. **Lecture Notes in Computer Science**, [s.l.], v. 1, n. 1, p. 209-219, 2019. Disponível em: http://dx.doi.org/10.1007/978-3-030-32327-1_40. Acesso em: 23 out. 2024.

HEVNER *et al*. Design science in information systems research. **Mis Quarterly**, [s.l.], v. 28, n. 1, p. 75, 2004. Disponível em: http://dx.doi.org/10.2307/25148625. Acesso em: 09 dez. 2024.

HITZLER, Pascal; KRÖTZSCH, Markus; RUDOLPH, Sebastian. **Foundations of semantic web technologies**. Boca Raton: CRC Press, 2010.

HOMBURG, Timo. Connecting semantic situation descriptions with data quality evaluations: towards a framework of automatic thematic map evaluation. **Information**, [s.l.], v. 11, n. 11, p. 532, 15 nov. 2020. Disponível em: http://dx.doi.org/10.3390/info11110532. Acesso em: 11 nov. 2024.

HUANG, Li *et al*. An RDF data set quality assessment mechanism for decentralized systems. **Data Intelligence**, [s.l.], v. 2, n. 4, p. 529-553, out. 2020. Disponível em: http://dx.doi.org/10.1162/dint_a_00059. Acesso em: 11 nov. 2024.

IBÁÑEZ, Luis-Daniel *et al*. An assessment of adoption and quality of linked data in european open government data. **Lecture Notes In Computer Science**, [s.l.], p. 436-453, 2019. Disponível em: http://dx.doi.org/10.1007/978-3-030-30796-7_27. Acesso em: 23 out. 2024.

IBÁÑEZ, Luis-Daniel *et al*. Col-Graph: towards writable and scalable linked open data. **Lecture Notes In Computer Science**, [s.l.], p. 325-340, 2014. Disponível em: http://dx.doi.org/10.1007/978-3-319-11964-9_21. Acesso em: 23 out. 2024.

IETF. *Hypertext Transfer Protocol -- HTTP/1.1*, 1999. Disponivel em: https://www.ietf.org/rfc/rfc2616.txt

IHMC. **Ferramentas cmaptools**. Disponível em: https://cmap.ihmc.us/cmaptools/. Acesso em: 18 out. 2024.

INTERNATIONAL CONFERENCE ON BIG DATA (BIG DATA), 1., 2015, [S.L]. **Anais [...] .** [s.l]: Ieee, 2015. p. 737-746. Disponível em: https://ieeexplore.ieee.org/document/7363818. Acesso em: 23 out. 2024.

ISAAC, Antoine; BAKER, Thomas. *Linked Data* practice at different levels of semantic precision: the perspective of libraries, archives and museums. **Bulletin of the Association for Information Science and Technology**, [s.l.], v. 41, n. 4, p. 34-39, abr. 2015. Disponível em: http://dx.doi.org/10.1002/bult.2015.1720410411. Acesso em: 11 nov. 2024.

ISO. **ISO 9001:2008**: Sistemas de gestão da qualidade. [s.l], 2008.

ISO. **Quality management principles**. Geneva: 2015. 20 p.

ISO/IEC. **ISO/IEC 2382**: Iformation technology — Vocabulary. Vernier, 2015.

ISO/IEC. **ISO/IEC 25012**. 2022. Disponível em: https://iso25000.com/index.php/en/iso-25000-standards/iso-25012. Acesso em: 17 abr. 2023.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados abertos conectados**. São Paulo: Novatec, 2015. 175 p. Disponível em: http://www.icmc.usp.br/e/b0477. Acesso em: 19 nov. 2020.

ISSA, Subhi *et al*. Knowledge graph completeness: a systematic literature review. **Ieee Access**, [s.l.], v. 9, p. 31322-31339, 2021. Disponível em: https://ieeexplore.ieee.org/document/9344615. Acesso em: 23 out. 2024.

JESUS, Ananda Fernanda de *et al*. O uso do método design science research na Ciência da Informação: uma revisão sistemática da literatura. **Atoz**: novas práticas em informação e conhecimento, [S.L.], v. 12, p. 1, 26 jul. 2023. Disponível em: http://dx.doi.org/10.5380/atoz.v12i0.87478. Acesso em: 09 dez. 2024.

JESUS, Ananda Fernanda de. **Recomendações teórico-metodológicas para a publicação de dados bibliográficos abertos e conectados**. 2021. 165 f. Tese (Doutorado) - Curso de Programa de Pós-graduação em Ciência da Informação, Universidade Federal de São Carlos, São Carlos, 2021.

JESUS, Ananda Fernanda de; SANTAREM SEGUNDO, José Eduardo. A questão da qualidade em dados publicados como linked data: um mapeamento sistemático da literatura. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 22., 2022, Porto Alegre. **Anais** [...]. Porto Alegre: Ancib, 2022. p. 1-16. Disponível em: https://enancib.ancib.org/index.php/enancib/xxiienancib/paper/viewFile/811/729. Acesso em: 19 dez. 2022.

JESUS, Ananda Fernanda de; SANTAREM SEGUNDO, José Eduardo. Qualidade de dados linked data: análise da temática sob a perspectiva da Ciência da Informação. **Informação@Profissões**, [s.l.], v. 11, n. 2, p. 153-169, 20 set. 2023. Disponível em: 10.5433/2317-4390.2022v11n2p153. Acesso em: 19 dez. 2023.

JESUS, Ananda Fernanda de; SEGUNDO, José Eduardo Santarem. A descrição formal da qualidade de dados publicados na web: análise do data quality vocabulary (dqv). **Em Questão**, [s.l.], v. 29, n. -129415, p. 1-34, 2023. Disponível em: http://dx.doi.org/10.1590/1808-5245.29.129415. Acesso em: 19 nov. 2024.

JULIANI, Jordan Paulesky et al. Governança de dados aplicada no processo de catalogação. **Revista Brasileira de Biblioteconomia e Documentação**, [s. l.], v. 15, n. 2, p. 81–105, 2019. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/1153. Acesso em: 19 dez. 2023.

JURAN, Joseph M. *et al* (ed.). **JURAN'S QUALITY HANDBOOK**. 5. ed. New York: McGraw-Hill, 1998. 1699 p.

KAHLAWI, Adham. An Ontology Driven ESCO LOD Quality Enhancement. **International Journal of Advanced Computer Science and Applications**, [s.l.], v. 11, n. 3, p. 60-70, 2020. Disponível em: http://dx.doi.org/10.14569/ijacsa.2020.0110308. Acesso em: 23 out. 2024.

KAMDAR, Maulik R.; MUSEN, Mark A. An empirical meta-analysis of the life sciences linked open data on the web. **Scientific Data**, [s.l.], v. 8, n. 1, p. 1-21, 21 jan. 2021. Disponível em: http://dx.doi.org/10.1038/s41597-021-00797-y. Acesso em: 23 out. 2024.

KETTOUCH, Mohamed Salah. **A newapproach for interlinking and integrating semi-structured and linked data**. 2017. 208 f. Tese (Doutorado) –

KETTOUCH, Mohamed Salah; LUCA, Cristina. LinkD: element-based data interlinking of rdf datasets in linked data. **Computing**, [s.l.], v. 104, n. 12, p. 2685-

2709, 16 jul. 2022. Disponível em: http://dx.doi.org/10.1007/s00607-022-01107-z. Acesso em: 11 nov. 2025.

KITCHENHAM, Barbara *et al*. **Procedures for Performing Systematic Reviews**. Eversleigh: Empirical Software Engineering National Ict Australia Ltd., 2004. 33 p.

KNUTH, Magnus. Linked data cleansing and change management. **Lecture Notes in Computer Science**, [s.l.], p. 201-208, 2015. Disponível em: http://dx.doi.org/10.1007/978-3-319-17966-7_29. Acesso em: 23 out. 2024.

KONTOKOSTAS, Dimitris *et al*. Databugger. **Proceedings Of The 23Rd International Conference On World Wide Web**, [s.l.], p. 115-118, 7 abr. 2014a. Disponível em: http://dx.doi.org/10.1145/2567948.2577017. Acesso em: 23 out. 2024.

KONTOKOSTAS, Dimitris *et al*. Test-driven evaluation of *linked data* quality. **Proceedings of the 23Rd International Conference On World Wide Web**, [s.l.], p. 747-758, 7 abr. 2014b. Disponível em: http://dx.doi.org/10.1145/2566486.2568002. Acesso em: 23 out. 2023.

KONTOKOSTAS, Dimitris *et al*. TripleCheckMate: a tool for crowdsourcing the quality assessment of linked data. **Communications In Computer and Information Science**, [s.l.], p. 265-272, 2013. Disponível em: 10.1007/978-3-642-41360-5_22. Acesso em: 23 out. 2024.

KROMPAß, Denis; NICKEL, Maximilian; TRESP, Volker. Querying factorized probabilistic triple databases. **Lecture Notes In Computer Science**, [s.l.], p. 114-129, 2014. Disponível em: http://dx.doi.org/10.1007/978-3-319-11915-1_8. Acesso em: 11 nov. 2024.

LANGER, André *et al*. SemQuire: assessing the data quality of linked open data sources based on dqv. **Lecture Notes in Computer Science**, [s.l.], p. 163-175, 2018. Springer International Publishing. Disponível em: Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-03056-8_14. Acesso em: 23 out. 2024.

LANGER, André; GAEDKE, Martin. DaQAR - An ontology for the uniform exchange of comparable linked data quality assessment requirements. **Lecture Notes In Computer Science**, [s.l.], p. 234-242, 2018. Disponível em: 10.1007/978-3-319-91662-0_18. Acesso em: 23 out. 2024.

LEMOS, Daniela Lucas da Silva; COELHO JUNIOR, Abeil. Qualidade de dados em acervos do patrimônio cultural: uma avaliação diagnóstica semiautomática nos objetos culturais sob gestão do instituto brasileiro de museus. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, [s.l.], v. 28, p. 1-22, 7 fev. 2023. Disponível em: https://doi.org/10.5007/1518-2924.2023.e90510. Acesso em: 19 dez. 2023.

LIU, Wenqiang *et al*. A new truth discovery method for resolving object conflicts over linked data with scale-free property. **Knowledge And Information Systems**, [s.l.], v. 59, n. 2, p. 465-495, 3 mai. 2018. Disponível em: 10.1007/s10115-018-1192-z. Acesso em: 23 out. 2024.

LIU, Wenqiang *et al*. Exploiting Source-Object Networks to Resolve Object

LIU, Wenqiang *et al*. TruthDiscover. **Proceedings of the 26Th International Conference On World Wide Web Companion - WWW '17 Companion**, [s.l.], p. 243-246, 2017. Disponível em: https://doi.org/10.1145/3041021.3054722. Acesso em: 23 out. 2024.

LÓSCIO, Bernadette Farias; OLIVEIRA, Marcelo Iury S.; BITTENCOURT, Ig Ibert. Publicação e consumo de dados na web: conceitos e desafios. In: SBBD, 30., 2015, Petrópolis. **Anais [...].** Petrópolis: Sbbd, 2015. p. 39-68.

LOV. **In LOV at a glance...** [s.d.]. Disponível em: https://lov.linkeddata.es/dataset/lov/about. Acesso em: 20 nov. 2025.

MACEDO, Dirceu Flavio. **Dados abertos governamentais:** modelo de governança voltado a qualidade de dados para publicação em rede. 2021. 151 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Federal do Espírito Santo, Vitória, 2021.

MANGEL, Simon *et al*. Data reliability and trustworthiness through digital transmission contracts. **Lecture Notes in Computer Science**, [s.l.], p. 265-283, 2021. Disponível em: http://dx.doi.org/10.1007/978-3-030-77385-4_16. Acesso em: 11 nov. 2024.

MARTINS, Dalton Lopes et al. Requisitos de qualidade para dados de agregação em museus. **Tendências da Pesquisa Brasileira em Ciência da Informação**, [s.l], v. 14, n. 1, p. 1-25, 2021b. Disponível em: https://brapci.inf.br/index.php/res/v/197846. Acesso em: 19 dez. 2023.

MARTINS, Dalton Lopes *et al*. Requisitos de qualidade para dados de agregação em museus: o caso ibram. In: ENANCIB, 21., 2021a, Rio de Janeiro. **Anais [...].** Rio de Janeiro: Ancib, 2021. p. 1-15. Disponível em: https://enancib.ancib.org/index.php/enancib/xxienancib/paper/view/575. Acesso em: 19 nov. 2024.

MCKENNA, Lucy; DEBRUYNE, Christophe; O'SULLIVAN, Declan. Modelling the provenance of linked data interlinks for the library domain. **Companion Proceedings of the 2019 World Wide Web Conference**, [s.l.], p. 954-958, 13 maio 2019. Disponível em: http://dx.doi.org/10.1145/3308560.3316518. Acesso em: 11 nov. 2025.

MELO, Jessica Oliveira de Souza Ferreira. **Metodologia de avaliação de qualidade de dados no contexto do *linked data*.** 2017. 111 f. Tese (Doutorado) - Curso de Programa de Pós-graduação em Ciência da Informação, Faculdade de Filosofia e Ciências de Marília, Universidade Estadual Paulista, Marilia, 2017. Disponível em: https://repositorio.unesp.br/server/api/core/bitstreams/a4efef73-6c80-4060-b48b-27dddf311409/content. Acesso em: 19 dez. 2023.

MELO, Jessica Oliveira de Souza Ferreira; BOTEGA, Leonardo Castro; SANTARÉM SEGUNDO, José Eduardo. Metodologia de avaliação de qualidade para dados conectados. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais** [...]. Marilia: Ancib, 2017a. p. 1-19.

Disponível em: http://enancib.marilia.unesp.br/index.php/XVIII_ENANCIB/ENANCIB/paper/viewFile/257/1266. Acesso em: 19 dez. 2023.

MELO, Jessica Oliveira de Souza Ferreira; BOTEGA, Leonardo Castro; SANTAREM SEGUNDO, José Eduardo. Metodologia de avaliação de qualidade para dados conectados. **Informação & Tecnologia**, [s.l.], v. 4, n. 2, p. 80-101, 20 out. 2017b. Disponível em: http://dx.doi.org/10.22478/ufpb.2358-3908.2017v4n2.40539. Acesso em: 11 nov. 2024.

MIHINDUKULASOORIYA, Nandana *et al*. A Linked data profiling service for quality assessment. **Lecture Notes In Computer Science**, [s.l.], p. 335-340, 2017. Disponível em: http://dx.doi.org/10.1007/978-3-319-70407-4_42. Acesso em: 23 out. 2024.

MIHINDUKULASOORIYA, Nandana *et al*. An analysis of the quality issues of the properties available in the spanish dbpedia. **Lecture Notes In Computer Science**, [s.l.], p. 198-209, jan. 2015. Disponível em: http://dx.doi.org/10.1007/978-3-319-24598-0_18. Acesso em: 11 nov. 2024.

MIHINDUKULASOORIYA, Nandana; GARCÍA-CASTRO, Raúl; GÓMEZ-PÉREZ, Asunción. LD Sniffer: a quality assessment tool for measuring the accessibility of *Linked Data*. **Lecture Notes In Computer Science**, [s.l.], p. 149-152, 2017. Disponível em: http://dx.doi.org/10.1007/978-3-319-58694-6_20. Acesso em: 23 out. 2024.

MIHINDUKULASOORIYA, Nandana; RICO, Mariano. Type prediction of rdf knowledge graphs using binary classifiers with structural data. **Lecture Notes In Computer Science**, [s.l.], v. 11153, n. 1, p. 279-287, 2018. Disponível em: http://dx.doi.org/10.1007/978-3-030-03056-8_27. Acesso em: 23 out. 2024.

MOREIRA, Fábio Mosso et al. A qualidade na recuperação de dados governamentais: um estudo sobre dados de políticas públicas na internet. **Perspectivas em Ciência da Informação**, [s.l.], v. 25, n. 2, p. 103-132, jun. 2020. Disponível em: http://dx.doi.org/10.1590/1981-5344/3994. Acesso em: 19 dez. 2023.

MOSS, Laura; CORSAR, David; PIPER, Ian. A *Linked Data* approach to assessing medical data. **Ieee International Symposium on Computer-Based Medical Systems (CBMS)**, [s.l.], v. 61, n. 25, p. 1-4, jun. 2012. Disponível em: http://dx.doi.org/10.1109/cbms.2012.6266391. Acesso em: 23 out. 2024.

MOURA JUNIOR, Pedro Jácome de; ARAGÃO, Maicon Henrique Ferreira. Metas, ações e indicadores como subsídios para análise da qualidade de dados: uma inversão necessária entre consequentes e antecedentes. **Anais do Workshop de Informação, Dados e Tecnologia - WIDaT**, [s. l.], v. 2, p. 4–10, 2018. DOI: 10.22477/ii.widat.132. Disponível em: https://labcotec.ibict.br/widat/index.php/widat2023/article/view/132. Acesso em: 19 nov. 2024.

MUÑOZ, Emir. On learnability of constraints from RDF data. **Lecture Notes in Computer Science**, [s.l.], v. 9678, n. 1, p. 834-844, 2016. Disponível em: http://dx.doi.org/10.1007/978-3-319-34129-3_52. Acesso em: 23 out. 2024.

NAHARI, Mohammad Khodizadeh *et al*. A framework for *linked data* fusion and quality assessment. **2017 3Th International Conference On Web Research (Icwr)**, [s.l.], v. 782, p. 67-72, abr. 2017. Disponível em: http://dx.doi.org/10.1109/icwr.2017.7959307. Acesso em: 23 out. 2024.

NOOGHABI, Mahdi Zahedi; DASTGERDI, Akram Fathian. Proposed metrics for data accessibility in the context of linked open data. **Program**, [s.l.], v. 50, n. 2, p. 184-194, 4 abr. 2016. Disponível em: http://dx.doi.org/10.1108/prog-01-2015-0007. Acesso em: 11 nov. 2024.

OCLC. **Linked Data Overview**. Disponível em: https://www.oclc.org/research/areas/data-science/linkeddata/linked-data-overview.html. Acesso em: 03 dez. 2024.

ODONI, Fabian *et al*. On the importance of drill-down analysis for assessing gold standards and named entity linking performance. **Procedia Computer Science**, [s.l.], v. 137, p. 33-42, 2018. Disponível em: http://dx.doi.org/10.1016/j.procs.2018.09.004. Acesso em: 11 nov. 2024.

PAULHEIM, Heiko; BIZER, Christian. Improving the quality of linked data using statistical distributions. **International Journal on Semantic Web and Information Systems**, [s.l.], v. 10, n. 2, p. 63-86, 1 abr. 2014. Disponível em: http://dx.doi.org/10.4018/ijswis.2014040104. Acesso em: 23 out. 2024.

PEFFERS, Ken *et al*. A design science research methodology for information systems research. **Journal Of Management Information Systems**, [s.l.], v. 24, n. 3, p. 45-77, dez. 2007. Disponível em: http://dx.doi.org/10.2753/mis0742-1222240302. Acesso em: 09 dez. 2024.

PICCOLO, Daiane Marcela et al. Qualidade de dados em gestão de dados de pesquisa. **Em Questão**, [s.l.], p. 159-184, 7 dez. 2021. Disponível em: https://doi.org/10.19132/1808-5245281.159-184. Acesso em: 19 dez. 2021.

PICCOLO, Daiane Marcela. Qualidade de dados dos sistemas de informação do Datasus: análise crítica da literatura. **Ciência da Informação em Revista**, [s.l.], v. 5, n. 3, p. 13-19, 31 dez. 2018. Disponível em: https://doi.org/10.28998/cirev.2018v5n3b. Acesso em: 19 dez. 2023.

POSSEMATO, Tiziana. How RDA is essential in the reconciliation and conversion processes for quality linked data. **Jlis**, [s.l.], v. 9, n. 1, p. 48-60, 2018. Disponível em: http://dx.doi.org/10.4403/jlis.it-12447. Acesso em: 11 nov. 2024.

PRIBERAM (Brasil). **Dicionário Priberam da Língua Portuguesa**. [s.l.]: Priberam, 2025. 1 p. Disponível em: https://dicionario.priberam.org/. Acesso em: 11 nov. 2025.

RADULOVIC, Filip *et al*. A comprehensive quality model for linked data. **Semantic Web**, [s.l.], v. 9, n. 1, p. 3-24, 30 nov. 2017. Disponível em: 10.3233/SW-170267. Acesso em: 23 out. 2023.

RAHOMAN, Md-Mizanur; ICHISE, Ryutaro. Automatic erroneous data detection over type-annotated linked data.**Transactions on Information and Systems**, [s.l.], v. 99, n. 4, p. 969-978, 2016. Disponível em: http://dx.doi.org/10.1587/transinf.2015dap0022. Acesso em: 23 out. 2024.

RALSTON, Anthony; REILLY, Edwin D. (ed.). **Encyclopedia of Computer Science and Engineering**. [s.l.]: Van Nostrand Reinhold Company, 1983. 1664 p.

RASHID, Mohammad *et al*. A quality assessment approach for evolving knowledge bases. **Semantic Web**, [s.l.], v. 10, n. 2, p. 349-383, 21 jan. 2019. Disponível em: http://dx.doi.org/10.3233/sw-180324. Acesso em: 11 nov. 2024.

REDMAN, Thomas. Data quality management past, present, and future: towards a management system for data. In: SADIQ, Shazia (ed.). **Handbook of Data Quality**: research and practice. New York: Springer, 2013. p. 1-93.

RICO, Mariano *et al*. Predicting incorrect mappings: a data-driven approach applied to dbpedia. In: ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 33., 2018, [s.l.]. **Anais [...].** [S.L.]: Sac, 2018. v. 33, p. 323-330. Disponível em: https://doi.org/10.1145/3167132.3167164. Acesso em: 11 nov. 2024.

RINSER, Daniel; LANGE, Dustin; NAUMANN, Felix. Cross-lingual entity matching and infobox alignment in Wikipedia. **Information Systems**, [s.l.], v. 38, n. 6, p. 887-907, set. 2013. Disponível em: http://dx.doi.org/10.1016/j.is.2012.10.003. Acesso em: 11 nov. 2024.

RUAN, Tong *et al*. On evaluating web-scale extracted knowledge bases in a comparative way. **International Journal on Semantic Web and Information Systems**, [s.l.], v. 14, n. 1, p. 98-120, jan. 2018. Disponível em: http://dx.doi.org/10.4018/ijswis.2018010104. Acesso em: 23 out. 2024.

RUCKHAUS, Edna; BALDIZÁN, Oriana; VIDAL, María-Esther. Analyzing linked data quality with liquate. **Lecture Notes in Computer Science**, [S.L.], v. 1, n. 8186, p. 629-638, 2013. Disponível em: 10.1007/978-3-642-41033-8_80. Acesso em: 23 out. 2024.

RUCKHAUS, Edna; BALDIZÁN, Oriana; VIDAL, María-Esther. Analyzing linked data quality with liquate. **Lecture Notes In Computer Science**, [s.l.], p. 629-638, 2013. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-41033-8_80. Acesso

RULA, Anisa. DC Proposal: towards *linked data* assessment and linking temporal facts. **Lecture Notes In Computer Science**, [s.l.], p. 341-348, 2011. Disponível em: 10.1007/978-3-642-25093-4_27. Acesso em: 23 out. 2023.

SADIQ, Shazia (ed.). **Handbook of data quality**: research and practice. New York: New York, 2013. 440 p

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, [s.l.], v. 21, n. 2, p. 116, 20 dez. 2016. Disponível em: http://dx.doi.org/10.5433/1981-8920.2016v21n2p116. Acesso em: 11 nov. 2023.

SANTAREM SEGUNDO, Jose Eduardo. Web semântica, dados ligados e dados abertos: uma visão dos desafios do brasil frente às iniciativas internacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, [s. l.], v. 8, n. 2, 2015. Disponível em: https://revistas.ancib.org/tpbci/article/view/359. Acesso em: 20 nov. 2025.

SANTOS, Helton Douglas dos *et al*. Investigations into data published and consumed on the Web: a systematic mapping study. **Journal of the Brazilian Computer Society**, [s.l.], v. 24, n. 1, p. 1-22, 7 nov. 2018. Disponível em: http://dx.doi.org/10.1186/s13173-018-0077-z. Acesso em: 17 nov. 2025.

SANTOS, Marcelo Nair dos. **Fundamentos estruturais do registro bibliográfico:** revisitando a compreensão de Seymour Lubetzky sobre a entrada principal representativa da obra e sua manifestação. 2019. 263 f. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2019. Disponível em: https://repositorio.ufmg.br/handle/1843/33494. Acesso em: 08 set. 2020.

SANTOS, Nirian Martins Silveira dos; STREIT, Rosalvo Ermes. O processo decisório de governança de dados. **Brazilian Journal of Information Studies:** [s. l], v. 2, n. 12, p. 64-73, jan. 2018.

SANTOS, Plácida L. V. Amorim da Costa; SANT'ANA, Ricardo César Gonçalves. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação. **Ciência da Informação**, [s.l.], v. 42, n. 2, p. 199-209, 27 jan. 2015. Disponível em: http://dx.doi.org/10.18225/ci.inf.v42i2.1382. Acesso em: 19 jun. 2025.

SAPNA, R; RANI, Monika; MISHRA, Shakti. An Investigative study on the quality aspects of linked open data. **Proceedings of the 2018 International Conference on Cloud Computing and Internet of Things**, [s.l.], p. 33-39, 29 out. 2018. Disponível em: https://doi.org/10.1145/3291064.3291074. Acesso em: 23 out. 2024.

SCITEPRESS - Science and technology publications. Disponível em: 10.5220/0008494604350441. Acesso em: 11 nov. 2025.

SIKOS, Leslie. Mastering structured data on the semantic web: from HTML5 microdata to linked open data. EUA: Apress, 2015.

SILK. **Silk the linked data integration framework**. 2024. Disponível em: http://silkframework.org/. Acesso em: 11 nov. 2025.

SILVA, Cleiton Rodrigo Queiroz. **Critérios para priorização de estudos primários identificados por snowballing com conjunto inicial gerado por string de busca**. 2017. 156 f. Dissertação (Mestrado) - Curso de O Programa de Pósgraduação em Ciência da Computação, Universidade Federal de São Carlos, São Carlos, 2017.

SILVA, Jordana *et al*. Desenvolvimento de ontologia ciente de qualidade de informações para a melhoria de consciência situacional no domínio de gerenciamento de emergências. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 1., 2017, Florianópolis. **Anais** [...]. Florianópolis: Ufsc, 2017. p. 53-58. Disponível em:

https://repositorio.ufsc.br/xmlui/bitstream/handle/123456789/180265/Anais.do.WIDAT2017.pdf?sequence=1&isAllowed=y. Acesso em: 19 dez. 2023.

SILVA, Kátia Regina da *et al*. Glocal clinical registries: pacemaker registry design and implementation for global and local integration: methodology and case study. **Plos One**, [S.L.], v. 8, n. 7, p. 1-12, 25 jul. 2013. Disponível em: http://dx.doi.org/10.1371/journal.pone.0071090. Acesso em: 11 nov. 2024.

SINDICE. **Sindice:** data web services. Disponível em: https://www.sindice.com/index.html. Acesso em: 11 nov. 2025.

TALLERÅS, Kim. Quality of linked bibliographic data: the models, vocabularies, and links of data sets published by four national libraries. **Journal Of Library Metadata**, [S.L.], v. 17, n. 2, p. 126-155, 3 abr. 2017. Disponível em: http://dx.doi.org/10.1080/19386389.2017.1355166. Acesso em: 03 dez. 2025.

TOMOYOSE, Kazumi. **O data catalog vocabulary (DCAT) para a publicação de dados de pesquisa nos princípios linked data**. 2021. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de São Carlos, São Carlos, 2021. Disponível em: https://repositorio.ufscar.br/handle/20.500.14289/14116.

TOURINO, Emanuelle. **Arquitetura de dados no contexto da ciência da informação**. 2023. 334 f. Tese (Doutorado) - Curso de Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília, 2023. Disponível em: https://repositorio.unesp.br/entities/publication/b2192b88-8362-488f-9b85-2c173eb66e48. Acesso em: 20 nov. 2025.

TRIQUES, Maria Lígia; ARAKAKI, Ana Carolina Simionato; CASTRO, Fabiano Ferreira de. Aspectos da representação da informação na curadoria digital. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, [s.l.], v. 25, p. 01-21, 8 maio 2020. Disponível em: https://doi.org/10.5007/1518-2924.2020.e69898. Acesso em: 14 abr. 2025.

TSICHRITZIS, Dionysios C.; LOCHOVSKY, Frederick H. **Data Models**. [s.l.]: Prentice Hall; 1982. 395 p.

TURI, Leandro Furlam; COMARELA, Giovanni. IMPACTO DA ADEQUAÇÃO À LEI GERAL DE PROTEÇÃO DE DADOS PESSOAIS NA METRIFICAÇÃO DA QUALIDADE DE DADOS. In: Workshop de informação, dados e tecnologia, 5., 2022, Vitória. **Anais** [...]. Vitória: Ufes, 2022. p. 121-126. Disponível em: https://widat2022.ufes.br/wp-content/uploads/2023/04/widat-2022-anais.pdf. Acesso em: 19 dez. 2023.

VAN HOEVEN, Loan R. *et al*. Validation of multisource electronic health record data: an application to blood transfusion data. **Bmc Medical Informatics and Decision Making**, [s.l.], v. 17, n. 1, p. 1-10, 14 jul. 2017. Disponível em: 10.1186/s12911-017-0504-7. Acesso em: 11 nov. 2024.

VARMDAL, Torunn *et al*. Data from national health registers as endpoints for the Tromsø Study: correctness and completeness of stroke diagnoses. **Scandinavian Journal of Public Health**, [s.l.], v. 51, n. 7, p. 1042-1049, 14 jun. 2021. Disponível em: http://dx.doi.org/10.1177/14034948211021191. Acesso em: 11 nov. 2024.

VILLAZÓN-TERRAZAS, Boris; VILCHES-BLÁZQUEZ, Luis. M.; CORCHO, Oscar; GÓMEZ-PÉREZ, Asunción. Methodological guidelines for publishing government linked data. **Linking Government Data**, [S.L.], p. 27-49, 2011. Disponível em: http://dx.doi.org/10.1007/978-1-4614-1767-5_2. Acesso em: 11 nov. 2025.

W3C. **Best practices for publishing *linked data***. 2014. Disponível em: https://www.w3.org/TR/ld-bp/. Acesso em: 26 jan. 2021.

W3C. **Data on the web best practices: data quality vocabulary**. 2016a. Disponível em: https://www.w3.org/TR/vocab-dqv/. Acesso em: 17 abr. 2023.

W3C. **Links in html documents**. 2018. Disponível em: https://www.w3.org/TR/html401/cover.html#minitoc. Acesso em: 08 set. 2020.

W3C. **Primer RDF.** 2004. Disponível em: https://www.w3.org/TR/rdfprimer/#intro. Acesso em: 08 set. 2025.

W3C. **URIs, URLs e URNs:** clarifications and recommendations 1.0. 2011. Disponível em: https://www.w3.org/TR/uri-clarification/. Acesso em: 08 set. 2020.

W3C. **Data catalog vocabulary (DCAT) - Version 2**. 2020. Disponível em: https://www.w3.org/TR/vocab-dcat/. Acesso em: 16 jan. 2023.

W3C. **Data catalog vocabulary (DCAT) - Version 3**. 2024. Disponível em: https://www.w3.org/TR/vocab-dcat-3/. Acesso em: 11 nov. 2025.

W3C. **Data quality vocabulary (DQV)**. 2015a. Disponível em: https://www.w3.org/2013/dwbp/wiki/Data_Quality_Vocabulary_(DQV). Acesso em: 17 abr. 2023.

W3C. **Describing linked datasets with the VoID vocabulary**. 2011. Disponível em: https://www.w3.org/TR/void/. Acesso em: 11 nov. 2025.

W3C. **Hypertext transfer protocol**. 1999. Disponível em: https://www.w3.org/Protocols/HTTP/1.1/draft-ietf-http-v11-spec-rev-05.txt. Acesso em: 16 dez. 2024.

W3C. **Linked data glossary**. 2013. Disponível em: https://www.w3.org/TR/ld-glossary/#dereferenceable-uris. Acesso em: 04 dez. 2024.

W3C. **Linked data**. 2023. Disponível em: https://www.w3.org/wiki/LinkedData. Acesso em: 20 nov. 2025.

W3C. **List of DQV implementations list of DQV implementations**. 2016b. Disponível em: https://www.w3.org/2013/dwbp/wiki/List_of_DQV_implementations. Acesso em: 17 abr. 2023.

W3C. **Naming and addressing: URIs, URLs, ...** 2001. Disponível em: https://www.w3.org/Addressing/. Acesso em: 20 nov. 2025.

W3C. **RDF 1.2 Concepts and abstract data model**. 2025. Disponível em: https://www.w3.org/TR/rdf12-concepts/#dfn-literal. Acesso em: 20 nov. 2025.

W3C. **SKOS simple knowledge organization system reference**. 2009. Disponível em: https://www.w3.org/TR/skos-reference/. Acesso em: 11 nov. 2025.

W3C. **Vocabularies**. 2015b. Disponível em: https://www.w3.org/standards/semanticweb/ontology#:~:text=There%20is%20no%20clear%20division,in%20a%20very%20loose%20sense. Acesso em: 18 abr. 2023.

WANG, Richard, STRONG, Diane. M. Beyond accuracy: what data quality means to data consumers. **J. Manage. Inf. Syst**. v. 12, n. 4, p 5–33, jan. 1996.

ZAVERI, Amrapali *et. al*. Quality assessment methodologies for linked open data. **SWJ**, v.1, p. 1-5, 2012. Disponível em: https://www.semantic-web-journal.net/system/files/swj414.pdf. Acesso em: 17 abr. 2023

ZHANG, Shuxin; BENIS, Nirupama; CORNET, Ronald. Assessing resolvability, parsability, and consistency of RDF resources: a use case in rare diseases. **Journal Of Biomedical Semantics**, [s.l.], v. 14, n. 1, p. 1-19, 5 dez. 2023. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s13326-023-00299-3.

GLOSSÁRIO

Termos relacionados à Qualidade de dados

Qualidade

Qualidade pode ser definida como a medida de adequação de determinada entidade à requisitos de qualidade. A qualidade é objeto dos processos de avaliação, melhoria e controle. Os requisitos podem ser pré-estabelecidos ou implícitos, de caráter contextual ou intrínseco. Requisitos pré-estabelecidos são formalizados em um documento que irá orientar os diferentes processos, sendo exemplos as melhores práticas, políticas, normas e regulamentos. Esses documentos possuem diferentes níveis de abrangência, como nacional, regional e local. Os requisitos implícitos precisam ser identificados antes da condução dos processos mencionados. Em uma abordagem contextual, a qualidade está relacionada a capacidade da entidade de atender às necessidades dos usuários, sendo considerada um investimento, que terá um custo mais elevado. Em uma abordagem intrínseca, a qualidade está relacionada a ausência de erros e deficiências que possam levar a falhas ou retrabalho, resultando na economia de recursos. O termo qualidade não pode ser empregado como sinônimo de excelência ou para comparação entres diferentes entidades, para denotar essa acepção, pode ser empregado um termo composto, como: alta qualidade, baixa qualidade, boa qualidade, má qualidade, melhor qualidade ou pior qualidade.

Dado

Um dado é a unidade de conteúdo de granularidade mais fina, composto por dois aspectos: o abstrato e o registrado. Em seu aspecto abstrato, o dado se refere a entidades do mundo real (conceituais ou físicas) e a relação existente entre elas, podendo um dado possuir mais de uma representação. Para permitir a sua interpretação, o dado depende de um modelo, abstração do domínio no qual se insere, podendo esse modelo estar implícito ou explícito. O modelo estabelece as entidades (e) do domínio, os atributos (a) dessa entidade e os valores (v) possíveis/esperados desse atributo para a entidade em questão. Nesse sentido, cada dado equivale a uma tripla <e,a,v>. Quando registrado, seja em meio analógico ou digital, o dado se torna passível de armazenamento, recuperação, processamento (por humanos ou máquinas), interpretação e reinterpretação e avaliação de qualidade. Por suas características, o dado é um insumo para sistemas, análises e aplicações, permitindo a compreensão de fenômenos, a predição de cenários e a identificação de padrões.

Qualidade de Dados

O termo qualidade de dados pode ser considerado poliédrico, de caráter interdisciplinar e multidimensional, contando com as acepções: **1. Enquanto um campo, ou domínio**, a qualidade de dados possui uma comunidade de discurso e uma produção cientifica, técnica e tecnológica

bem estabelecida, de escopo interdisciplinar, de interesse de diversas áreas do conhecimento no qual são discutidas, investigadas e propostas formas de mensurar e melhorar os níveis de qualidade dos conjuntos de dados, assim como formular artefatos e metodologias que permitam auxiliar na condução de processos de avaliação e melhoria de qualidade de dados. **2. Enquanto um problema/desafio** a qualidade de dados é entendido como uma barreira a ser superada por pessoas, empresas e organizações para o uso efetivo e eficiente de dados em diversos contextos. **3. Enquanto uma medida ou uma ponderação** pode ser abordada como sinônimo dos conceitos “dados de qualidade”, “qualidade dos dados” ou “qualidade aplicada a dados”, onde a qualidade dos dados é um aspecto que pode ser mensurado e melhorado. A medida de qualidade dos dados não pode ser mensurada de maneira única ou absoluta, e pode estar relacionada à: adequação a normas, princípios, padrões e melhores práticas; a ausência de erros e anomalias; e a adequação ao domínio ou uso pretendido dos dados **4. Enquanto processo**, a qualidade de dados é abordada pela literatura como um sinônimo dos processos de avaliação e controle de qualidade de dados. O processo de controle de qualidade busca atingir e manter as medidas de qualidade necessárias ou estabelecidas. O processo de avaliação de qualidade é realizado por meio do estabelecimento de critérios, dimensões e métricas que permitem mensurar os níveis de qualidade dos conjuntos de dados. O caráter interdisciplinar e multidimensional do termo afeta todas as acepções mencionadas, uma vez que tanto o domínio da qualidade de dados, como os problemas, a sua medida, manutenção e avaliação são objeto de interesse de diversas áreas do conhecimento. Podem ser considerar as principais dimensões de qualidade de dados a contextual, intrínseca, representacional e a acessibilidade.

Termos relacionados ao *Linked Data*

Linked Data

O *Linked Data* é um conjunto de boas práticas, de caráter não prescritivo ou normativo, para a publicação e conexão de dados estruturados na W*eb*. Tem como objetivo fortalecer e facilitar a busca, recuperação, acesso, reuso, intercambio e interoperabilidade de dados tanto para usuários humanos como para agentes computacionais. Busca prover a serendipidade (descoberta acidental) e o enriquecimento semântico dos dados. As boas práticas se utilizam de padrões da *Web*, como o protocolo HTTP e o uso de identificadores únicos, como URIs e IRIs. O modelo RDF é a base para a estrutura dos dados *Linked Data*, tendo seu uso combinado ao de formatos de serialização. O RDF prevê o uso de vocabulários e ontologias, que permitem o enriquecimento semântico dos dados. Essas conexões ocorrem por meio de *Links* que conectam dados de uma mesma fonte ou de fontes diversas. O *Linked Data* foi proposto em 2006 por Tim Berners-Lee, sendo posteriormente expandido, explicado e completado por uma série de documentos e boas práticas do W3C e pela própria comunidade de usuários. Sua adoção também é afetada por princípios e normas do domínio ao qual se relacionam os dados. As práticas do *Linked*

Data podem ser adotadas em conjunto com práticas de dados abertos, sendo, nesse caos, nomeado pela comunidade como *Linked Open Data*. Existe uma pluralidade em relação a tradução do termo para o português, que pode ser traduzido como dados conectados, dados vinculados, dados ligados ou dados interligados.

Universal Resource Identifiers (URIs)

Os Universal Resource Identifiers (URIs) são identificadores globais padronizados que representam, por meio do uso de uma estrutura e caracteres preestabelecidos, os recursos na Web. Recursos podem ser coisas diversas, como pessoas e objetos do mundo real, conceitos e abstrações e ainda recursos e documentos da Web. Existem diferentes tipos e nomenclaturas, sendo exemplos os Uniform Resource Locator (URL), que definem a localização do recurso a partir de um protocolo, o URN (Unified Resource Name) que definem o nome dos recursos, e ainda os International Resource Identifier (IRIs) que estendem as possibilidades de caracteres permitidos para a representação dos recursos. Os URIs podem ser URIs HTTPs, que permitem que esses sejam desreferenciados/ resolvidos, utilizando o procolo de compartilhamento da Web para permitir acesso a informações sobre as entidades, no caso de representações de objetos e pessoas do mundo real, ou ainda aos próprios recursos no caso de recursos Web. Os URIs são essenciais para a estruturação dos dados em RDF pois as entidades e propriedades devem, necessariamente, ser URIs. Os objetos das declarações em RDF podem ser URIs ou Literais. Quando são utilizados URIs como objetos, criam-se Links entre dados de uma mesma fonte ou de fontes diversas.

Literal

No contexto do *Linked Data* um Literal é um conjunto de valores em linguagem natural e que não é representado por um URI em uma declaração RDF. Os literais são números, datas, textos, *strings*, dentre outros. Referem-se a entidades do mundo real, como pessoas, objetos, conceitos e abstrações. Literais podem ser associados a uma *tag* que indique em que idioma estão. Devem ser acompanhados de uma indicação de tipo de dados em algum dos esquemas aceitos pelo RDF, representados por um URI, essa identificação facilita a avaliação da qualidade dos literais. Literais podem ser utilizados apenas na posição de valor em uma declaração RDF, não podendo ser utilizado nas posições de recurso e propriedade.

Resource Description Framework (RDF)

O RDF é um *framework* (estrutura), ou modelo padrão de dados, que permite identificar e descrever as características dos recursos na *Web*, bem como explicitar e nomear as relações existentes entre recursos. Essa estrutura facilita o compartilhamento, exposição e combinação de dados de domínios distintos. Baseia-se na estrutura de declarações em formato de tripla <sujeito> <predicado> <objeto>. O sujeito é representado por um URI do recurso cuja caraterística ou relação está sendo descrita. O predicado deve ser o URI da propriedade de um vocabulário que nomeei a característica ou a relação descrita. O objeto pode ser um URI ou um Literal

que represente o valor da propriedade para a característica descrita ou um outro recurso com o qual o primeiro possui uma relação. Cada característica ou relação resulta em uma nova declaração. Um conjunto de triplas pode ser representado no formato de grafo. Um conjunto de grafos pode ser reunido em um catálogo de dados RDF

Vocabulário

Vocabulários são coleções de termos criados para um contexto específico. Os termos podem ser classes, propriedades e tipos de dados, que permitem identificar, descrever e representar características possíveis dos recursos de um domínio bem como as potenciais relações entre recursos. Os vocabulários são utilizados como predicados nas triplas RDF, permitindo o enriquecimento dos dados. Podem ser criados por qualquer pessoa para atenderem a necessidade de um domínio específico, entretanto, preferencialmente deve-se reutilizar um vocabulário já estabelecido. Podem ser feitos mapeamentos entre termos correlatos de distintos vocabulários. No contexto do *Linked Data* não existe uma diferenciação rígida entre os termos “vocabulário” e “ontologia”, sendo comumente utilizado pelo W3C e pela comunidade de maneira intercambiável. Entretanto, é possível considerar o termo vocabulário como um termo abrangente, do qual são tipos ontologias e vocabulários controlados/sistemas de organização do conhecimento. Nessa acepção, as ontologias são vocabulários caracterizados pela complexidade formal, pela presença de axiomas e restrições e pelo caráter estruturante, fornecendo a base para a descrição de outros vocabulários. Os vocabulários controlados/sistemas de organização do conhecimento, nesse contexto, fornecem a base para caracterizar e descrever as características dos recursos e as relações de um determinado domínio. Os vocabulários são elaborados em RDF, com base na estrutura de vocabulários do tipo ontologia, e utilizando URIs para representar os termos.

SPARQL

O SPARQL pode ser entendido como um conjunto de especificações e protocolos do W3C, composto por uma linguagem de consulta e por orientações para a criação de *Endpoints* SPARQL. Tem o objetivo de permitir a busca e manipulação de conjuntos de dados e catálogo de dados RDF. A linguagem de consulta permite diferentes tipos de buscas e é baseada no uso de triplas, como no RDF, entretanto, cada uma das partes da tripla <sujeito> <predicado> <objeto> pode ser substituída por uma variável.

Termos relacionado à avaliação de qualidade de dados *Linked Data*

Qualidade de Dados *Linked Data*

O termo qualidade de dados *Linked Data* refere-se às implicações da adoção do *Linked Data* na qualidade dos dados, e pode contar com as seguintes acepções: **1. Enquanto um campo ou domínio** a Qualidade de Dados *Linked Data* possui uma comunidade própria, heterogênea e

interdisciplinar, focada principalmente na criação de ferramentas, metodologias e modelos para possibilitar a realização dos processos de avaliação e controle de qualidade de dados. A comunidade também se concentra em avaliar e comparar os níveis de qualidade das fontes de dados disponíveis. **2. Enquanto um problema a ser superado**, a Qualidade de Dados *Linked Data* possui desafios relacionados principalmente com as fontes de dados, com a estrutura dos dados e com o próprio processo de avaliação. As fontes de dados são heterogêneas e possuem diferentes níveis de curadoria dos dados. Os problemas de estrutura em dados *Linked Data* são relacionados principalmente com aplicação incorreta do RDF, a criação de URIs inconsistente e com problemas relacionados a seleção e aplicação de vocabulários e propriedades. Em relação ao processo de avaliação de qualidade, os dados *Linked Data* enfrentam problemas relacionados ao volume e variedade dos dados, às ferramentas que estão indisponíveis ou não são amigáveis para os usuários, e as variações nos objetivos de qualidade da comunidade, que dificultam a criação de ferramentas únicas e generalistas. **3. Enquanto uma medida** a Qualidade de Dados *Linked Data* pode ser medida com base na adequação dos dados aos princípios e melhores práticas disponibilizados pelo W3C e pela comunidade ao qual os dados se relacionam, pode ser medida em relação a adequação ao uso ou ainda em relação a sua adequação sintática. **4. Enquanto um processo a avaliação da Qualidade de dados *Linked Data*,** em sua maioria, adota a estrutura proposta por Wang e Strong (1996) ou a estrutura proposta pela norma ISO de qualidade de dados vigente. São adotadas as categorias e dimensões clássicas, sendo adaptadas as suas definições para o contexto. São criados critérios e dimensões próprios para o contexto do *Linked Data*, levando em consideração principalmente sua estrutura pautada em triplas seguindo o modelo RDF.

Avaliação de qualidade

A avaliação de qualidade de dados pode ser definida como um processo ou como uma atividade realizada no âmbito de outros processos, como gestão e controle de qualidade de dados, que busca realizar um diagnóstico da qualidade dos dados em relação a diferentes aspectos preestabelecidos. A avaliação pode ser realizada de maneira pontual ou cíclica, sendo necessária em diversos momentos do ciclo de vida dos dados. É organizada de maneira hierárquica, composta por categorias, dimensões, critérios e métricas. É considerada multifacetada, pois pode ser avaliada com base em diferentes perspectivas, sendo elas: intrínseca, contextual, representacional e acessibilidade. A avaliação de qualidade possuí quatro etapas principais: estabelecimento do modelo de qualidade, mensuração, avaliação dos resultados e realização de atividades necessárias. O modelo de qualidade estabelece quais dimensões, critérios e métricas serão avaliados. Para mensurar a qualidade podem ser necessárias ferramentas automáticas, semiautomáticas e/ou manuais. A avaliação dos resultados pode ser feita por meio de uma comparação entre os resultados obtidos e os resultados esperados, que devem ser pré-estabelecidos. Pode ainda ter como base a comparação dos resultados de diferentes conjuntos de dados. Ao final da avaliação, os resultados podem

ser utilizados para selecionar fontes de dados ou para identificar e corrigir problemas de qualidade. Os principais agentes da avaliação de qualidade são os consumidores e os publicadores. A avaliação de qualidade depende do domínio de criação e uso dos dados, ou da previsão de uso esperada pelos publicadores. No contexto do *Linked Data*, precisam ser levados em consideração tanto os princípios e melhores práticas do W3C, quanto os relacionados ao domínio do conteúdo dos dados. Essa característica diminui a relevância de modelos de qualidade muito abrangentes ou genéricos.

Ferramenta de avaliação de qualidade de dados

As ferramentas de avaliação de qualidade de dados são instrumentos criados para auxiliar ou realizar completamente a avaliação de qualidade de dados. Existe uma grande pluralidade em relação as ferramentas, com diferentes tipos, que possuem nomenclaturas próprias. Elas variam em relação ao domínio para o qual foram criadas, existindo ferramentas para domínios específicos e ferramentas de aspecto geral. Existem diversas ferramentas criadas especificamente para a avaliação de dados *Linked Data*. As ferramentas variam quanto as dimensões e métricas avaliadas, existindo ferramentas focadas em dimensões específicas, permitindo ou não a personalização dessas dimensões e métricas e do peso que cada uma delas tem em relação ao resultado da avaliação. Elas podem desempenhar diferentes atividades, como avaliação, a avaliação e correção dos problemas de qualidade identificados ou ainda concentrar-se na exportação de resultados do processo de avaliação de qualidade, como é caso dos vocabulários de qualidade de dados. Variam quanto a forma como desempenham essas atividades, existindo ferramentas com abordagens automáticas, semiautomáticas e manuais. Elas podem ser criadas ainda para um público específico, existindo ferramentas focadas nas necessidades dos consumidores e dos publicadores bem como ferramentas voltadas para ambos, sendo mais comum a criação de ferramentas focadas em publicadores de dados.

Modelo de qualidade de dados

Um modelo de qualidade é uma estrutura que fornece descrição detalhada sobre como será conduzida a avaliação de qualidade. É composto pelas dimensões, critérios e métricas que serão utilizados e por detalhes relevantes, como os pesos das métricas, definições e orientações. O fornecimento detalhado dessas informações é a base da avaliação de qualidade, sendo fundamental para que os resultados possam ser compreendidos, auditados, atualizados, compartilhados e comparados. Os modelos de qualidade podem variar em relação ao nível de especificidade, existindo modelos genéricos e modelos criados para aplicação apenas em um cenário específico de uso de dados. Mesmo os modelos genéricos precisam ser ajustados para aplicação, geralmente são mantidas as dimensões, sendo feitos ajustes nos critérios e especialmente nas métricas e nos pesos que recebem essas métricas em relação ao resultado geral da avaliação de qualidade.

Categoria de Qualidade

Uma categoria de qualidade é uma classe que permite a organização e o agrupamento de distintos aspectos de qualidade com base em sua semelhança. É considerado o nível mais abrangente na hierarquia da avaliação de qualidade. As categorias são compostas por dimensões, que agrupam critérios, que podem ser mensurados com base em métricas de qualidade. Como categorias e dimensões possuem um caráter abstrato, para se avaliar os níveis de qualidade dos dados em relação a determinada categoria torna-se necessário a aplicação de um conjunto de métricas, indicadores que permitem mensurar quantitativa e qualitativamente a qualidade dos dados. A literatura tradicionalmente organiza o processo de avaliação de qualidade e quatro categorias, sendo elas: Intrínseca, Contextual, Acessibilidade e Representacional.

Categoria Contextual

A categoria contextual permite a organização e o agrupamento de dimensões cuja avaliação depende das características, políticas e boas práticas do domínio de criação/uso dos dados; da tarefa que será realizada com esses dados; e/ou das necessidades e objetivos do usuário. É a perspectiva mais adotada pela literatura, sinônimo da expressão "*fitness for use*". Nessa perspectiva um conjunto de dados pode possuir boa qualidade para um contexto e não ser adequado para outro. Por sua dependência de fatores externos, é um desafio para os produtores atingirem altos níveis de qualidade contextual. No contexto do *Linked Data*, os domínios de origem e os contextos de aplicação em potencial dos dados são diversos e o paradigma de compartilhamento que circunda a publicação desses dados torna ainda mais complexo prever em que cenário esses dados serão empregados no futuro, ampliando esse desafio. As principais dimensões contextuais de qualidade de dados são: relevância, confiabilidade, compreensibilidade, completude e temporalidade.

Categoria Intrínseca

A categoria intrínseca permite a organização e o agrupamento de dimensões relacionadas às características inerentes dos dados, que podem ser mensuradas de maneira independente da tarefa a ser realizada ou das necessidades e objetivos do usuário, embora sua avaliação seja, em determinada medida, influenciada pelo domínio de criação dos dados. Geralmente a qualidade intrínseca é definida do ponto de vista do produtor. Nessa perspectiva, para que os dados sejam considerados de boa qualidade, eles precisam ser logicamente consistentes, sintaticamente corretos, coerentes, compactos e corretos, estarem livres de anomalias e representarem adequadamente a realidade a qual estão relacionados. No contexto do *Linked Data*, a qualidade intrínseca está relacionada com a criação de bons URIs, a aplicação correta da estrutura do RDF e com a seleção e aplicação de vocabulários e propriedades. Nesse contexto as principais dimensões intrínsecas são: acurácia, confiabilidade, validade sintática, validade semântica, precisão semântica, consistência, concisão, completude.

Categoria Representacional

A categoria representacional agrupa e organiza dimensões relacionadas com a qualidade da estrutura responsável pela materialização dos dados, que possibilita o seu registro, a sua codificação e armazenamento, partindo do entendimento de dados como “unidades de registro abstratas que necessitam de um suporte”. Nessa perspectiva, para serem considerados de boa qualidade, precisam ter clareza estrutural, estarem livres de redundâncias, coerentes em relação a sua sintaxe e semântica, e serem fáceis de interpretar para usuários humanos e para agentes computacionais. No contexto do *Linked Data* a qualidade de dados representacional está relacionada com os formatos de serialização dos dados, com a presença de comentários e anotações que facilitem o seu entendimento, com a análise estrutural dos URIs, que não devem ser demasiadamente longos ou de difícil compreensão, e com a análise dos literais, que devem seguir os padrões de escrita e forma estabelecidos pelo domínio. Está relacionado ainda com a interoperabilidade dos dados, fortemente impactada pela aplicação dos vocabulários. As principais dimensões de qualidade representacional são: concisão representacional, interoperabilidade, interpretabilidade, versatilidade, facilidade de compreensão.

Categoria Acessibilidade

A categoria acessibilidade permite a organização e o agrupamento de dimensões que buscam avaliar em que medida os dados podem ser recuperados, estão disponíveis e acessíveis, permitindo seu uso e reuso. Também reúne dimensões que buscam avaliar a performance e a eficiência dos meios de acesso e a autenticidade dos dados. Em uma perspectiva de acessibilidade são considerados de boa qualidade dados que podem ser facilmente recuperados, obtidos e que possuem uma licença de uso clara e disponível tanto para usuários humanos como para agentes computacionais. Nesse contexto do *Linked Data*, a qualidade de acessibilidade está relacionada com a disponibilidade e o funcionamento dos *Endpoints* SPARQL, com a disponibilidade de *download* do conjunto de dados em diferentes formatos de serialização, com a presença de licença legível por humanos e máquinas. Nesse contexto, também busca verificar em que medida os dados estão conectados com fontes externas relevantes. As principais dimensões de acessibilidade são: disponibilidade, licenciamento, interligação, segurança, performance e acessibilidade.

Dimensão

Uma dimensão de qualidade é uma categoria que reúne características semelhantes dos dados, relevantes para a avaliação da qualidade. Consiste em um conjunto de atributos abstratos que representam um aspecto único no âmbito geral da qualidade de dados, fornecendo assim vocabulário para estabelecer os requisitos de qualidade esperados para um conjunto de dados. Por seu caráter abstrato, podem ser materializadas em um ou mais critérios mensuráveis. Para avaliar a qualidade dos dados em relação a uma dimensão é necessário o estabelecimento de uma ou mais métricas, indicadores que permitem a avaliação quantitativa e qualitativa da qualidade em relação a determinada dimensão. A escolha e a definição das dimensões e dos critérios e métricas que a compõe depende do domínio

no qual estão inseridos os dados e dos objetivos do processo de avaliação de qualidade. Existe um conjunto de dimensões estabelecido na comunidade de avaliação de qualidade, que podem ser adaptadas para atender a necessidades específicas. Também podem ser criadas dimensões para atender aos propósitos do processo de avaliação, entretanto, a adoção de dimensões bem estabelecidas facilita o uso de ferramentas automáticas e a interoperabilidade dos resultados do processo de avaliação.

Critério

Um critério de qualidade representa uma característica mensurável dos dados. Os critérios de qualidade são agrupados, com base em sua semelhança, em categorias e dimensões de qualidade. Podem possuir um caráter subjetivo ou objetivo, o que irá impactar na forma como serão avaliados. Para que a qualidade dos dados em relação a um critério possa ser avaliada são necessárias métricas de qualidade, podendo um critério ser avaliado com base em mais de uma métrica.

Métrica

Métricas de qualidade são procedimentos que permite calcular em que medida determinada critério de qualidade é atendido e os níveis de qualidade dos dados em relação a uma dimensão. Uma métrica é considerada o aspecto mais granular na hierarquia da qualidade de dados, permitindo mensurar e comparar os níveis de qualidade de um ou mais conjuntos de dados em relação a um aspecto específico. As métricas podem possuir um caráter quantitativo ou qualitativo, subjetivo ou objetivo. Podem ser estruturadas como escalas ou como fórmulas. Variam em relação a complexidade do processo de avaliação e podem exigir o uso de instrumentos e ferramentas específicos como validadores de estrutura e conteúdo, listas de verificação e padrões-ouro de qualidade.

ÍNDICE ALFABÉTICO DE TERMOS DO GLOSSÁRIO