



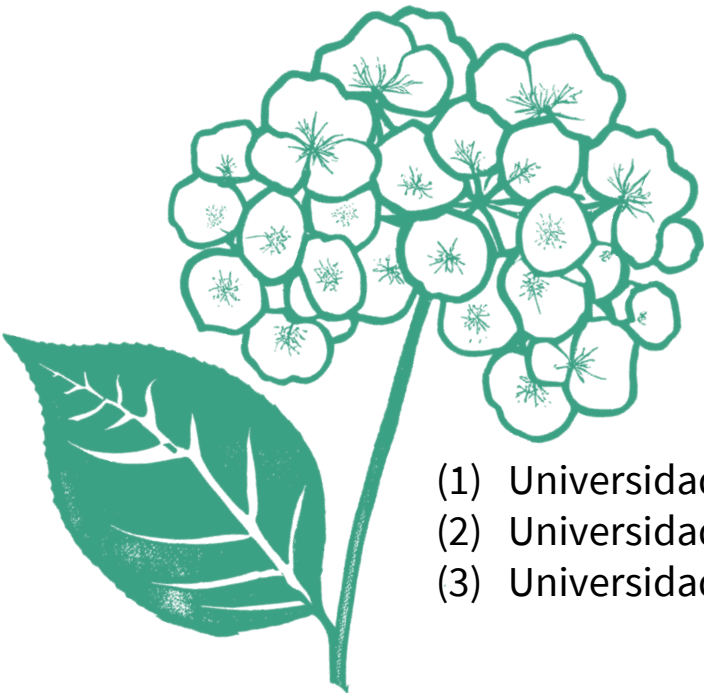
VIII ISKO BRASIL 2025



Análise da literatura sobre o uso de vocabulários no processo de avaliação de qualidade de dados Linked Data

**Ananda Fernanda de Jesus (1)
José Eduardo Santarem Segundo (2),
Mario Barité (3).**

- (1) Universidade Estadual Paulista (UNESP); af.jesus@unesp.br
- (2) Universidade de São Paulo (USP); santarem@usp.br
- (3) Universidad de la República (UDELAR) mario.barite@fic.edu.uy



Embora os dados *Linked Data* **possuam uma estrutura pré-estabelecida**, esses dados **também estão sujeitos a problemas de qualidade**

Esse cenário se intensifica pela prática de “**publicar primeiro e melhorar depois**”

O **processo** de avaliação de qualidade permite **mensurar quantitativa e qualitativamente**:

O quão adequado está um conjunto de dados frente a determinada aplicação (contextual)

O quão bem formado, livre de erros e aderente à diretrizes e melhores práticas está - (intrínseca).



AVALIAÇÃO DE QUALIDADE



VIII ISKO BRASIL 2025



Baseada em categorias, dimensões, critérios e métricas

DIMENSÕES - categorias abstratas, organizam e agrupam os diferentes aspectos que serão avaliados

CRITÉRIOS - especificam as características dos dados a serem avaliados

MÉTRICAS - indicadores que permitem mensurar a qualidade dos dados em relação a um critério.





A avaliação de qualidade pode ser baseada no uso de **ferramentas automáticas ou semiautomáticas**



Os **vocabulários** podem desempenhar um papel importante, **facilitando ou viabilizando uso dessas ferramentas**



Possuem ainda mais destaque quando se trata de dados *Linked Data*



Considerando que o uso do **RDF** implica **na aplicação de propriedades**, formalizadas e descritas em **vocabulários semânticos**.



Como tem sido discutidos e aplicados os vocabulários no processo de avaliação de qualidade de dados *Linked Data*?

Compreender como a literatura científica tem bordado o uso de vocabulários no processo de avaliação de qualidade de dados *Linked Data*.

Análise exploratória da literatura, baseada no uso de protocolo de pesquisa





Web of Science (WoS), Library, Information Science and Technology Abstracts (LISTA), Library & Information Science Abstracts (LISA)



("Linked Data" OR "Linked Open Data" OR "Linked Open Vocabularies" OR "W3C") AND ("Data Quality") AND ("Semantic Vocabularies" OR "Vocabularies" OR "Ontologies" OR "Data Quality Vocabularies" OR "Quality Vocabularies" OR "PROV Ontology" OR "Data Catalog Vocabulary")



(I) Artigos que aplicam vocabulários no processo de avaliação de qualidade de dados *Linked Data*
(I) Artigos que discutem a relação entre vocabulários e qualidade de dados *Linked Data*



(E) Apenas menciona a temática de interesse; (E) Não aborda a temática de interesse;
(E) Não foi possível obter acesso ao documento completo; (E) não está nos idiomas estabelecidos para a pesquisa



Resultados



VIII ISKO BRASIL 2025



157
Recuperados

146
Recusados

3 Duplicados

8 Aceitos





7 artigos: focados na criação/apresentação de vocabulários relacionados à qualidade de dados



1 artigo: teórico que discute a aplicação de vocabulários na qualidade em meio a outros aspectos



A literatura de qualidade de dados sempre menciona a importância dos vocabulários nesse processo mas não se aprofunda nessa questão



O uso de vocabulários no processo de avaliação de qualidade possui os seguintes aspectos:

Representação dos conjuntos de dados

Ampliação do uso de ferramentas

Divulgação dos resultados do processo de avaliação

Participação dos usuários no processo de avaliação



Refere-se ao uso de vocabulários para estruturar a descrição formal dos conjuntos de dados, descrevendo informações como:

Conteúdo	Volume	Frequência de atualização	Estrutura	Proveniência	Direitos de uso
----------	--------	---------------------------	-----------	--------------	-----------------



A divulgação dessas informações amplia as **possibilidades de uso e reuso dos dados e a confiança** dos consumidores de dados.

AMPLIA A PERCEPÇÃO DE QUALIDADE



Uso ferramentas



VIII ISKO BRASIL 2025



Vocabulário semânticos tornam viável a criação e o uso de ferramentas para auxiliar no processo de avaliação de qualidade



O uso dessas ferramentas é **indispensável em casos com grandes volumes de dados heterogêneos**



Permitem a **exportação dos resultados do processo de avaliação ou reutilização resultados de processos anteriores.**



Divulgação dos resultados



VIII ISKO BRASIL 2025



Refere-se a criação e uso de vocabulários contendo as propriedades necessárias para descrever os resultados do processo de avaliação

Podem ser usados por **publicadores, consumidores e agentes computacionais**

Permite que os resultados do processo de avaliação sejam:

Reaproveitados

Comparados

Auditados



O Data Quality Vocabulary (DQV)



VIII ISKO BRASIL 2025



Elaboração

- Em 2016, para instrumentalizar as Melhores Práticas para a Publicação de Dados na Web (MPs) do W3C
- MP06 (orienta a disponibilização de informações sobre a qualidade do conjunto de dados de maneira formal)

Estrutura

- Estruturado em RDF, com propriedades relacionadas a URIs, seguindo os princípios do *Linked Data*
- Reutiliza propriedades de outros vocabulários

Finalidade

- Provê termos que descrevem, de maneira formal, os conceitos e relações da avaliação de qualidade
- Não apresenta dimensões e métricas específicas (depende de um modelo de qualidade)



Participação dos usuários



VIII ISKO BRASIL 2025



Vocabulários podem ser utilizados para permitir o fornecimento de feedbacks de uso



Comunicação de problemas de qualidade e da sua resolução



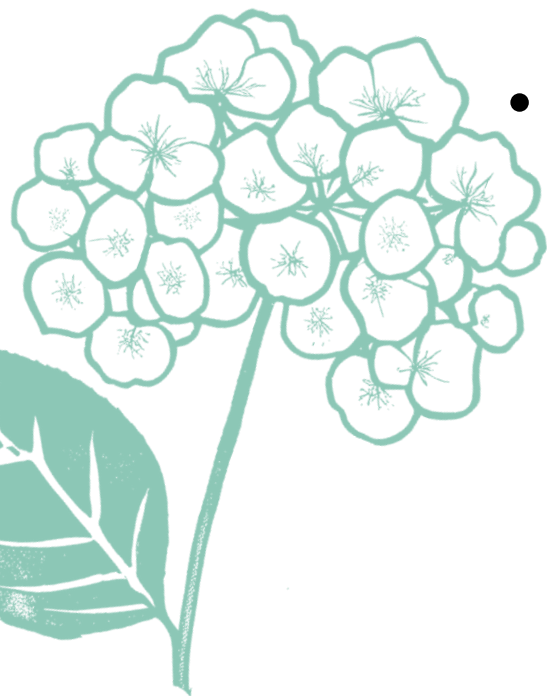
Possibilitar a criação de redes de avaliação de qualidade geridas por usuários



A avaliação dos usuários (contextual) muitas vezes terá uma perspectiva distinta da perspectiva dos publicadores (intrínseca)



- Destaca-se a necessidade de um maior aprofundamento **a respeito do DQV** outros vocabulários de qualidade, **abordando a sua aplicação no processo de avaliação, suas contribuições e limitações**
- Nota-se a lacuna de estudos que **discutam a sua ampliação e aplicação em distintos domínios.**
- Também necessitam de aprofundamento nos impactos do uso de vocabulários como o DCAT e o PROV na qualidade de dados.



- Embora os vocabulários sejam um assunto frequente na literatura sobre qualidade de dados *Linked Data*
- E o seu papel no processo de avaliação seja constantemente mencionado e ressaltados por esses estudos
- Eles ainda são abordados em um **aspecto secundário**, sendo escassos estudos que discutam de maneira aprofundada essa relação, especialmente no âmbito da Ciência da Informação.



- Parte da marginalidade dos vocabulários no âmbito da qualidade de dados *Linked Data* deve-se justamente a uma lacuna de pesquisa: **ausência de estudos que sejam conduzidos no âmbito da Ciência da Informação.**
- Nesse sentido, a relação entre vocabulários e qualidade de dados *Linked Data* **é um campo que pode ser amplamente explorado pela comunidade da Ciência da Informação**, que possui aparatos teóricos e metodológicos para fomentar essas discussões.



- Agradecemos à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento recebido para o desenvolvimento dessa pesquisa Processo FAPESP n° 2021/03349-0.

