

Enriquecendo dados abertos conectados do QualisBrasil com Melhores Práticas W3C para publicação e consumo de dados na Web

Lucélia de Souza

Universidade Estadual do Centro Oeste – UNICENTRO, E-mail: lucelia@unicentro.br

Bruno de Carvalho Câmara

Universidade Estadual do Centro Oeste – UNICENTRO, E-mail: b.camara@live.com

Sandro Rautenberg

Universidade Estadual do Centro Oeste – UNICENTRO, E-mail: srautenberg@unicentro.br

Josiane M. H. Dall’Agnol

Universidade Estadual do Centro Oeste – UNICENTRO, E-mail: jhauagge@unicentro.br

Gisane A. Michelon

Universidade Estadual do Centro Oeste – UNICENTRO, E-mail: gisane@unicentro.br

RESUMO

A iniciativa de Dados Abertos Conectados incentiva o compartilhamento de conjuntos de dados conforme os princípios básicos (diretrizes e melhores práticas) estabelecidos pelo World Wide Web Consortium (W3C). Um desafio difundido pelo W3C é a promoção da interoperabilidade entre conjuntos de dados compartilhados, mediante a adoção de licenças, vocabulários e padrões. Um exemplo de compartilhamento de Dados Abertos Conectados é o grafo QualisBrasil, disponibilizando os recursos de dados do histórico do índice Qualis em pesquisas da Ciência da Informação. Todavia, a publicação do QualisBrasil é realizada mediante o uso do vocabulário SCOVO, que apesar de ser reconhecido por sua simplicidade, não é mais atualizado. Em face disso, esta pesquisa visa representar o grafo QualisBrasil com vocabulários recentes, conforme preconizado pelas Melhores Práticas do W3C. Especificamente, o grafo QualisBrasil é enriquecido com o vocabulário *Data Cube*, recomendado pelo W3C para publicação de dados multidimensionais. Como resultado, argumenta-se que ao adicionar a semântica do *Data Cube* ao QualisBrasil, contribui-se para a confiança, o reuso, a compreensão e a descoberta dos dados do referido grafo, ampliando as opções de interligação, interoperabilidade e processabilidade automatizada de seus recursos na web.

Palavras-chave: Web Semântica. Representação da Informação. Vocabulário Data Cube. Metadados. Armazenamento de dados.

1 INTRODUÇÃO

Como uma plataforma atual de compartilhamento de dados, a Web Semântica (BERNES-LEE *et al.*, 2001) tem-se configurado como um campo interdisciplinar de pesquisa e atuação, com aportes importantes da Ciência da Informação. Nesse contexto, vários pesquisadores têm estudado

e aplicado a base conceitual da Web Semântica (FIGUEIREDO; ALMEIDA, 2017; SANTARÉM SEGUNDO, 2015; MOSTAFA; SANTARÉM SEGUNDO; SABBAG, 2016; GABRIEL Jr, 2016; CATARINO, 2014; RAMALHO, 2015). Em seus trabalhos, os referidos autores discutem principalmente os desafios a serem sobrepostos no Brasil e a representação de dados com o uso de vocabulários e ontologias, visando a exploração informacional em vários domínios. Em suma, os trabalhos que exploram a Web Semântica se baseiam em diretrizes e melhores práticas reconhecidas internacionalmente. Uma das organizações promotoras dessas recomendações é o *World Wide Web Consortium (W3C)*. Pontualmente, essa organização incentiva o uso de 35 Melhores Práticas (MPs 1-35) relacionadas à publicação e uso de dados na web, as quais repercutem em oito benefícios (W3C, 2018):

- **Reuso (*Reuse*)** - aumenta as chances de reutilização de dados publicados por consumidores, independentemente do domínio de aplicação.
- **Compreensão (*Comprehension*)** - com a publicação de conjuntos de dados e seus respectivos metadados, os seres humanos têm melhor entendimento da natureza, da estrutura e da semântica dos dados compartilhados.
- **Interligação (*Linkability*)** - por publicar os dados em consonância com os protocolos da Internet, possibilita-se a criação de relações (*hiperlinks*) entre recursos de dados (conjuntos de dados e itens de dados).
- **Descoberta (*Discoverability*)** - mediante os *hiperlinks*, autonomamente, os computadores descobrem os conjuntos de dados e navegam entre os recursos de dados, aferindo o conhecimento relacional entre os itens compartilhados.
- **Confiança (*Trust*)** - com a publicação de metadados de proveniência, os publicadores expressam como os conjuntos de dados são preservados ao longo do tempo.
- **Acesso (*Access*)** - utilizando os protocolos de acesso da Internet, promove-se a encontrabilidade e a exploração dos dados pelos agentes (humanos e de software).
- **Interoperabilidade (*Interoperability*)** - permite que os dados sejam automaticamente convertidos em formatos abertos diversos (CSV¹, JSON², RDF³, XML⁴, entre outros).

¹ CSV é um acrônimo de *Comma-Separated Values*

² JSON é um acrônimo de *JavaScript Object Notation*

³ RDF é um acrônimo de *Resource Description Framework*

- **Processabilidade (*Processability*)** - adicionalmente à interoperabilidade, a processabilidade permite que as aplicações computacionais processem os dados contidos na Web de Dados de forma automática.

Cabe ressaltar que tais benefícios dão suporte à iniciativa de Dados Abertos Conectados - *Linked Open Data* - LOD (LOD, 2018). Essa iniciativa incentiva as atividades de publicação, uso e reuso de conjuntos de dados. Essas atividades podem ser guiadas pelas MPs estabelecidas pelo W3C (W3C, 2018). Para cada uma das 35 MPs propostas, associa-se: uma explicação do porquê da MP; resultados pretendidos com sua adoção; uma ou mais possíveis abordagens de implementação; exemplos de utilização por humanos e máquinas; como testá-las; a evidência com requisitos relevantes; assim como quais são os benefícios obtidos com sua adoção.

No contexto da Ciência da Informação, como exemplo de publicação de LOD, cita-se o grafo QualisBrasil disponível no *endpoint* <<http://lod.unicentro.br/sparql/>>. Este é um grafo na web de dados que disponibiliza o histórico dos dados do índice Qualis, no qual encontram-se índices de qualidade de periódicos científicos, permitindo sua utilização e interligação em pesquisas bibliométricas e cientométricas (RAUTENBERG *et al.*, 2014).

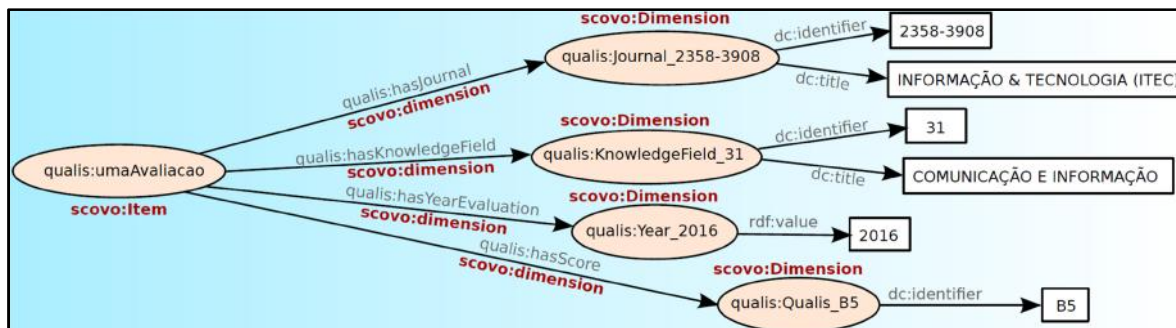
O grafo QualisBrasil é baseado na tabela de classificação do índice Qualis, atualmente disponibilizando 865.281 avaliações para o período 2005 a 2016 (RAUTENBERG; BURDA, 2016; RAUTENBERG; HILD; SOUZA, 2018). Em sua semântica são utilizados o Padrão de Metadados Dublin Core⁵ e o Vocabulário SCOVO⁶ (RAUTENBERG *et al.*, 2017). A Figura 1 apresenta uma avaliação, destacando o uso do Vocabulário SCOVO e seus recursos semânticos como `scovo:Item`, `scovo:dimension` e `scovo:Dimension`.

⁴ XML é um acrônimo de *eXtensible Markup Language*

⁵ Disponível em: <<http://dublincore.org/>>

⁶ SCOVO é um acrônimo de *Statistical Core Vocabulary*. Disponível em: <vocab.deri.ie/scovo>

Figura 1 - Avaliação de um periódico, destacando os recursos do Vocabulário SCOVO



Fonte: Elaborada pelos autores (2018).

Apesar do grafo QualisBrasil ter sido desenvolvido conforme os princípios LOD, o uso do Vocabulário SCOVO, embora reconhecido por sua simplicidade, atualmente não é mais atualizado. Nesse sentido, podem haver desafios na geração e no consumo de suas informações semânticas segundo as MPs propostas pelo Consórcio W3C. Assim, um exemplo de MP a ser revista na publicação do grafo QualisBrasil é a MP15, a qual aponta ao uso de um vocabulário atual para representação dos recursos.

Dessa forma, esta pesquisa tem como objetivo enriquecer os recursos do grafo QualisBrasil por meio de sua atualização com o vocabulário *Data Cube*, uma recomendação do W3C para publicação de dados multidimensionais.

Nesse sentido, esse trabalho contribui para a adoção de um subconjunto das 35 MPs para publicação e consumo de recursos do índice Qualis. Com isso, tem-se a expectativa de proporcionar interoperabilidade ao grafo QualisBrasil em conformidade aos vocabulários recomendados pelo W3C.

2 REVISÃO DE LITERATURA

Esta seção é reservada à apresentação do vocabulário denominado *The RDF Data Cube Vocabulary*⁷ e do procedimento metodológico adotado no enriquecimento do grafo QualisBrasil usando o referido vocabulário. Adicionalmente, é apresentado o processo de validação da modelagem desenvolvida.

⁷ *The RDF Data Cube Vocabulary*. Disponível em: <<https://www.w3.org/TR/vocab-data-cube/>>

2.1 THE RDF DATA CUBE VOCABULARY

O *Data Cube* (prefixo qb:) é um vocabulário difundido pelo W3C para publicar dados multidimensionais na web (DATA CUBE VOCABULARY, 2018). Os dados a serem publicados com este vocabulário têm sua semântica descrita conforme a Figura 2.

Pontualmente, um conjunto de dados é representado pelo recurso qb:Dataset, que é uma coleção de dados estatísticos correspondentes a uma estrutura definida. Os dados em um conjunto de dados podem ser descritos como pertencentes a um dos seguintes tipos:

- **Observações** – são os dados atuais, os valores medidos.
- **Estrutura organizacional** – para localizar uma observação em um modelo multidimensional é preciso especificar os valores de cada dimensão. De acordo com esses valores, as observações são recuperadas. Conjuntos de dados também podem ter estruturas organizacionais adicionais na forma de *Slices*, agrupando subconjuntos de observações dentro do conjunto de dados.
- **Metadados estruturais** – tais metadados são fornecidos como atributos e podem ser anexados às observações individuais.
- **Metadados de referência** – descrevem o conjunto de dados como um todo, incluindo sua categorização, o publicador e um *endpoint* para acesso aos dados.

Conforme o objetivo do trabalho, a modelagem do grafo QualisBrasil com o *Data Cube* é apresentada na Subseção 3.2.

2.2 METODOLOGIA CICLO DE VIDA DE DADOS CONECTADOS


Para o enriquecimento do grafo QualisBrasil com o vocabulário *Data Cube*, a metodologia adotada nesta pesquisa é denominada *Linked Data Lifecycle* (AUER, 2014), ou Ciclo de Vida de Dados Conectados. O referido ciclo de vida compreende oito estágios (Extração, Armazenamento/Consulta, Revisão Manual/Autoria, Interligação/Fusão, Classificação/Enriquecimento, Análise de Qualidade, Evolução/Reparação e Busca/Navegação/Exploração), os quais são executados de acordo com os requisitos exigidos em um projeto de publicação de dados em consonância aos princípios LOD.

2.3 CONSULTAS DE VALIDAÇÃO SPARQL ASK W3C

Como forma de certificação da representação de recursos multidimensionais com o *Data Cube*, o W3C (DATA CUBE VOCABULARY, 2018) propõe um *checklist* de validação. Tal *checklist* contém 21 consultas SPARQL ASK (IC-1 até IC-21)⁸, as quais devem retornar *false* a cada teste de validação. Se a consulta SPARQL ASK for aplicada e retornar *true* (verdadeiro), significa que o grafo contém uma ou mais instâncias *Data Cube* que violam a correspondente restrição de integridade. Um exemplo da execução de um teste de validação SPARQL ASK é apresentado na Figura 3, onde:

⁸ Disponível em: <<https://www.w3.org/TR/vocab-data-cube/>>

Figura 3 - Consulta para verificar se todo *Slice* tem uma estrutura única (IC-9)



```
Default Graph IRI http://lod.unicentro.br/QualisBrasilQB/
Query (A)
PREFIX qb: <http://purl.org/linked-data/cube#>
ASK {
  {
    # Slice has a key
    ?slice a qb:Slice .
    FILTER NOT EXISTS { ?slice qb:sliceStructure ?key }
  } UNION {
    # Slice has just one key
    ?slice a qb:Slice ;
    qb:sliceStructure ?key1, ?key2 .
    FILTER (?key1 != ?key2)
  }
}
Execute Save Load Clear
false (B)
```

Fonte: Elaborada pelos autores (2018).

-) **Figura 3(A)** - codifica uma consulta de integridade (IC-9), a qual verifica se todo qb:Slice tem uma estrutura única, atestando que cada qb:Slice tem exclusivamente um qb:sliceStructure associado.
-) **Figura 3(B)** – retorna o resultado da consulta. Neste caso, foi obtido *false* (falso), significando que o modelo não viola a referida restrição de integridade.

3 PROCEDIMENTOS METODOLÓGICOS

Nesta seção é apresentada a modelagem do grafo QualisBrasil em consonância ao vocabulário *Data Cube* e a validação do modelo proposto mediante as 21 consultas SPARQL ASK (IC-1 até IC-21) como prova de conceito.

3.1 ESTÁGIOS DA METODOLOGIA ADOTADA

Com o propósito de enriquecimento do grafo QualisBrasil com o vocabulário *Data Cube*, nesta pesquisa, são executados os seguintes estágios da Metodologia Ciclo de Vida de Dados Conectados:

- **Extração** – os dados primários em formato CSV foram modelados com o vocabulário *Data Cube*, associados a dados modelados com o vocabulário SCOVO, usando a ferramenta Sparqlify⁹.
- **Armazenamento/Consulta** – os recursos RDF são armazenados no servidor OpenLink Virtuoso¹⁰.
- **Classificação/Enriquecimento** – os recursos RDF modelados com SCOVO são enriquecidos semanticamente com o vocabulário *Data Cube*.
- **Busca/Navegação/Exploração** – como prova de conceito, consultas SPARQL são desenvolvidas para recuperar recursos de dados de um grafo de acordo com o vocabulário *Data Cube*.

3.2 DESENVOLVENDO O GRAFO QUALISBRASILQB

Inicialmente, para a modelagem realizou-se um comparativo entre os vocabulários SCOVO (prefixo scovo:) e *Data Cube* (prefixo qb:), visando encontrar estruturas semânticas correlacionadas. Verificou-se que: **i)** um recurso do tipo scovo:Item corresponde a um recurso do tipo qb:Observation; e **ii)** um predicado scovo:dimension tem uma equivalência com o predicado qb:observation ou uma propriedade instanciada de qb:MeasureProperty. Diante disso, ao criar um qb:Observation no *Data Cube* compatível com um scovo:Item, alguns recursos scovo:dimension de um scovo:Item em particular são referenciados com predicados qb:observation a uma instância de qb:Observation. Baseando-se na representação do grafo QualisBrasil na Figura 1, essas atualizações são representadas no núcleo da Figura 4.

⁹ A Sparqlify é uma ferramenta disponibilizada na plataforma Linux e possui como característica principal a conversão de um arquivo CSV para um arquivo com recursos RDF, conforme um arquivo de configuração declarado segundo a linguagem *Sparqlification Mapping Language* – SML (SPARQLIFY, 2018).

¹⁰ OpenLink Virtuoso é um sistema universal para acesso, integração e gerenciamento de dados baseados no modelo RDF na web (VIRTUOSO, 2018).

-) **dc:publisher** – especifica o publicador do conjunto de dados;
-) **dc:title** – nomeia o conjunto de dados;
-) **rdfs:comment** – usado similarmente ao dc:description; e
-) **rdfs:label** – usado similarmente ao dc:title.

Quadro 1 - Resumo da atualização do grafo QualisBrasil com o *Data Cube*

	Grafo QualisBrasilQB	Grafo QualisBrasil
Classes	qb:DataSet	-
	qb:MeasureProperty [QualisQB:refScore]	scovo:dimension
	qb: Observation	scovo:Item
	qb:Slice	scovo:Dimension
Propriedades	dc:description	-
	dc:issued	-
	dc:license	-
	dc:Publisher	-
	dc:title	-
	qb:dataSet	-
	qb:observation	scovo:dimension
	rdfs:comment	-
rdfs:label	-	

Fonte: Elaborado pelos autores (2018).

O Quadro 1 resume as atualizações realizadas, perfazendo: **i)** a correspondência conceitual entre os recursos existentes no grafo QualisBrasil para com o grafo QualisBrasilQB; e **ii)** o enriquecimento semântico do grafo QualisBrasilQB, ao especificar um recurso qb:DataSet e suas propriedades.

3.3 VALIDANDO O GRAFO QUALISBRASILQB COM CONSULTAS SPARQL ASK W3C

Como prova de conceito foram aplicadas as 21 consultas SPARQL ASK de validação propostas pelo W3C, para certificar a aderência da representação de um grafo RDF conforme os preceitos do modelo do *Data Cube*.

Todas as 21 consultas SPARQL ASK executadas retornaram *false* (falso) devido a sua adequação ao *Data Cube*. Com isso, denota-se que não há violação do uso deste vocabulário para com o grafo QualisBrasilQB. O Quadro 2 sumariza os testes realizados. A cada consulta de

Enriquecendo dados abertos conectados do QualisBrasil com Melhores Práticas W3C para publicação e consumo de dados na Web

validação SPARQL ASK (IC1-IC21) são apresentados: o código; o título; a descrição; a adequação ao grafo QualisBrasilQB; e o resultado.

Quadro 2 - Restrições de Integridade W3C aplicadas no grafo QualisBrasilQB (IC1-IC21)

CONSULTA	TÍTULO	DESCRIÇÃO	ADEQUAÇÃO	RESULTADO
IC-1	Conjunto de dados único	Todo qb:Observation tem exatamente um qb:DataSet associado.	Os dados do grafo QualisBrasilQB são observações relacionadas ao respectivo conjunto de dados.	<i>False</i>
IC-2	DSD único	Todo qb:DataSet tem exatamente um qb:DataStructureDefinition associado.	O conjunto de dados tem exatamente uma definição de estrutura de dados.	<i>False</i>
IC-3	DSD inclui medidas	Todo qb:DataStructureDefinition deve incluir no mínimo uma medida declarada.	A definição de estrutura dos dados inclui a medida <i>Score</i> .	<i>False</i>
IC-4	Dimensões tem intervalo	Toda dimensão declarada em um qb:DimensionProperty deve ter um rdfs:range declarado.	Todas as dimensões possuem um respectivo rdfs:range declarado baseado no modelo SCOVO.	<i>False</i>
IC-5	Dimensões Concept tem <i>code lists</i>	Toda dimensão com intervalo skos:Concept deve ter um qb:codeList.	No grafo QualisBrasilQB não foram declaradas dimensões com o intervalo skos:Concept.	<i>False</i>
IC-6	Somente atributos podem ser opcionais	Os únicos componentes de um qb:DataStructureDefinition que podem ser assinalados como opcionais usando qb:componentRequired são atributos.	No grafo QualisBrasilQB não foram declarados atributos.	<i>False</i>
IC-7	<i>Slice Keys</i> devem ser declarados	Todo qb:SliceKey deve ser associado com um qb:DataStructureDefinition.	Toda chave <i>Slice</i> é associada à definição da estrutura dos dados.	<i>False</i>
IC-8	<i>Slice Keys</i> consistentes com DSD	Todo qb:componentProperty em um qb:SliceKey deve também ser declarado como um qb:component da qb:DataStructureDefinition associada.	Todos os componentes de propriedades em uma chave <i>Slice</i> foram declarados como componentes da definição da estrutura de dados.	<i>False</i>
IC-9	Única estrutura de <i>Slice</i>	Cada qb:Slice deve ter exatamente um qb:sliceStructure associado.	Todo <i>Slice</i> tem sua respectiva chave <i>Slice</i> .	<i>False</i>
IC-10	Dimensões <i>Slice</i> completas	Todo qb:Slice deve ter um valor para toda dimensão declarada em seu qb:sliceStructure.	Todo <i>Slice</i> tem sua respectiva chave <i>Slice</i> com seu respectivo componente de propriedade.	<i>False</i>

CONSULTA	TÍTULO	DESCRIÇÃO	ADEQUAÇÃO	RESULTADO
IC-11	Todas as dimensões requeridas	Toda qb:Observation tem um valor para cada dimensão declarada em seu qb:DataSetDefinition associado.	Toda dimensão está declarada na Observação com seu respectivo valor.	<i>False</i>
IC-12	Nenhuma observação duplicada	Nenhuma de duas qb:Observation no mesmo qb:DataSet deve ter o mesmo valor para todas as dimensões.	Não há duplicação de Observações.	<i>False</i>
IC-13	Atributos requeridos	Toda qb:Observation tem um valor para cada atributo declarado que é marcado como requerido.	Não existem componentes requeridos.	<i>False</i>
IC-14	Todas as medidas presentes	Em um qb:DataSet que não usa uma dimensão <i>Measure</i> então cada qb:Observation individual deve ter um valor para cada medida declarada.	Nesse caso existe uma medida declarada que é o <i>Score</i> .	<i>False</i>
IC-15	Dimensão de medida consistente	Em um qb:DataSet que usa uma dimensão <i>Measure</i> então cada qb:Observation deve ter um valor para a medida correspondente a seu dado qb:measureType.	Os valores da medida <i>Score</i> são atribuídos à Observação.	<i>False</i>
IC-16	Simples medidas em uma <i>Observation</i> de dimensão <i>Measure</i>	Em um qb:DataSet que usa uma dimensão <i>Measure</i> então cada qb:Observation deve somente ter um valor para uma medida (pela IC-15 isto será a medida correspondente a seu qb:measureType).	Os valores da medida <i>Score</i> possuem somente um valor atribuído em sua Observação.	<i>False</i>
IC-17	Todas as medidas presentes na dimensão medida do cubo	Em um qb:DataSet que usa uma dimensão <i>Measure</i> então se existe uma <i>Observation</i> para alguma combinação de dimensões não-medidas deve ter outras <i>Observations</i> com os mesmos valores de dimensões não-medidas para cada uma das medidas declaradas.	Não há Observações que possuam valores de medidas que não estejam na medida <i>Score</i> .	<i>False</i>
IC-18	<i>Links</i> de conjuntos de dados consistentes	Se um qb:DataSet D tem um qb:slice S, e S tem um qb:Observation O, então o qb:DataSet correspondente a O deve ser D.	Todos os <i>links</i> são consistentes.	<i>False</i>
IC-19	Códigos a partir de <i>code list</i>	Se uma <i>dimension property</i> tem uma qb:codeList, então o valor de <i>dimension property</i> em toda qb:Observation deve estar na <i>code list</i> .	Não são usadas <i>code lists</i> .	<i>False</i>
IC-20	Códigos a partir da	Se uma <i>dimension property</i> tem um qb:HierarchicalCodeList com	Não são usadas <i>code lists</i> .	<i>False</i>

CONSULTA	TÍTULO	DESCRIÇÃO	ADEQUAÇÃO	RESULTADO
	hierarquia	um qb:parentChildProperty <i>non-blank</i> , então o valor daquela <i>dimension property</i> em toda qb:Observation deve ser alcançável a partir da raiz da hierarquia usando zero ou mais saltos ao longo dos <i>links</i> qb:parentChildProperty.		
IC-21	Códigos a partir da hierarquia (inverso)	Se uma <i>dimension property</i> tem uma qb:HierarchicalCodeList com qb:parentChildProperty inversa, então o valor daquela <i>dimension property</i> em toda qb:Observation deve ser alcançável a partir da raiz da hierarquia usando zero ou mais saltos ao longo dos <i>links</i> qb:parentChildProperty inversos.	Não são usadas <i>code lists</i> .	<i>False</i>

Fonte: Elaborado pelos autores (2018).

Em um endereço de Internet¹¹ encontra-se disponível o relatório da execução das 21 consultas SPARQL ASK (IC-1 até IC-21), aferindo a consonância do grafo QualisBrasilQB em relação ao modelo *Data Cube*.

4 RESULTADOS

Adicionalmente à modelagem do grafo QualisBrasil com o vocabulário *Data Cube*, nesta pesquisa foram adotadas algumas MPs (MP2, MP3, MP4, MP7, MP9, MP10, MP12, MP15 e MP34). Estas contribuem: **i)** no enriquecimento semântico do grafo QualisBrasil, resultando no grafo QualisBrasilQB; e **ii)** na agregação dos benefícios de Reuso, Compreensão, Interligação, Descoberta, Confiança, Interoperabilidade e Processabilidade de dados. Na sequência é descrito o subconjunto de MPs adotadas e sua aplicação ao grafo QualisBrasilQB com os respectivos benefícios obtidos:

- **MP2 - Forneça metadados descritivos** – esta MP diz respeito ao fornecimento de metadados que descrevem as características dos conjuntos de dados. Para adoção desta MP são definidos os metadados do qb:DataSet contendo: rótulo (rdfs:label), título (dc:title), comentário (rdfs:comment), descrição (dc:description), publicador (dc:publisher), data da publicação (dc:issued) e licença de uso (dc:license). Como

¹¹ Disponível em: <<https://goo.gl/oq6Tc9>>.

benefícios obtidos com a adoção da MP2, destacam-se: **i)** o Reuso do grafo QualisBrasilQB por diversos grupos de consumidores de dados; **ii)** a Compreensão pelos agentes humanos da estrutura dos dados e metadados, seu significado e sua natureza; e **iii)** a Descoberta dos recursos do grafo QualisBrasilQB autonomamente pelos agentes de software.

- **MP3 - Forneça metadados estruturais** – para sua adoção, devem ser fornecidos metadados que descrevem o esquema e a estrutura interna de um conjunto de dados em um formato legível por máquina e sem erros de sintaxe. Para adoção da MP3, é feita a descrição dos dados do grafo QualisBrasil por meio de classes, relações e propriedades, considerando a semântica do vocabulário padronizado, atualizado e multidimensional *Data Cube*. Como benefícios a partir da MP3, além do Reuso e da Compreensão, corrobora o benefício da Processabilidade de recursos semânticos do grafo QualisBrasilQB por parte dos agentes de software.
- **MP4 - Forneça informações sobre a licença dos dados** – para adotar esta MP, deve-se fornecer um *link* ou uma cópia do acordo de licença que controla o uso dos dados. Para sua adoção, no qb:DataSet é descrita a licença de uso dos dados. No caso do grafo QualisBrasilQB, a licença adotada é a *Creative Commons*¹². Adotando a MP4, como benefícios alcançados têm-se o Reuso e Confiança por parte dos consumidores de dados do grafo QualisBrasilQB.
- **MP7 - Forneça um indicador de versão** – esta MP sugere a declaração de um número de versão ou data para cada conjunto de dados. Para adotar esta MP, é fornecida a data de publicação do conjunto de dados por meio do metadado dc:issued em qb:DataSet. Em consonância à MP4, a MP7 também repercute nos benefícios de Reuso e Confiança.
- **MP9 - Use URIs persistentes como identificadores de conjuntos de dados** – esta MP propõe a identificação de cada conjunto de dados por um URI¹³ persistente, cuidadosamente escolhido. Para adotar a MP9, URIs do grafo QualisBrasilQB disponíveis no *endpoint* <<http://lod.unicentro.br/QualisBrasilQB/>> são mantidos persistentes juntamente com as URIs do grafo QualisBrasil. Como benefícios com a

¹² Disponível em: <<https://br.creativecommons.org>>

¹³ Acrônimo de *Uniform Resource Identifier*

adoção da MP9, além do Reuso; a Interligação entre recursos de dados do grafo QualisBrasilQB; a Descoberta dos recursos semânticos por agentes de software; e a Interoperabilidade possibilitando que o grafo QualisBrasilQB seja utilizado por diversos sistemas.

- **MP10 - Use URIs persistentes como identificadores dentro de conjuntos de dados** – esta MP propõe o uso de URIs persistentes de outros conjuntos de dados sobre o mesmo recurso para que sejam procurados e mais dados sejam descobertos. Para adotar a MP10, foi feita a integração entre o grafo QualisBrasil, desenvolvido com o vocabulário SCOVO e o grafo modelado com o vocabulário *Data Cube*, sendo mantidas suas URIs persistentes, resultando no grafo QualisBrasilQB, a partir do *endpoint* <<http://lod.unicentro.br/QualisBrasilQB/>>. Para geração das URIs persistentes foram seguidas as regras conforme (RAUTENBERG *et al.*, 2017). No caso da MP10, obtém-se os mesmos benefícios da MP9, sendo, o Reuso; a Interligação; a Descoberta; e a Interoperabilidade de seus recursos semânticos.
- **MP12 - Use formatos de dados padronizados legíveis por máquina** – é proposto por meio desta MP que os dados sejam disponíveis em um formato de dados padronizado, legível por máquina, sendo adequado ao seu uso pretendido ou potencial. A linguagem padrão adotada para a geração de dados semânticos a partir do grafo QualisBrasilQB usando o vocabulário *Data Cube* é a RDF, a qual especifica os dados na forma de triplas RDF, considerando um Sujeito, um Predicado e um Objeto (SPO), facilitando a legibilidade por máquinas. Como benefícios obtidos com a MP12, destacam-se: o Reuso e a Processabilidade automatizada do grafo QualisBrasilQB.
- **MP15 - Reuse vocabulários** – esta MP propõe o uso de termos compartilhados a partir de vocabulários, preferencialmente os padronizados, para codificar dados e metadados. Vocabulários padronizados podem ser consultados no repositório *Linked Open Vocabularies* (LOV)¹⁴. O vocabulário padronizado escolhido para uso é o *Data Cube* (prefixo qb:), recomendado pelo W3C para modelagem de dados estatísticos multidimensionais. Possui 15 classes e 21 propriedades, sendo usado atualmente em 25

¹⁴ Disponível em: <<https://lov.linkeddata.es/dataset/lov/>>

Enriquecendo dados abertos conectados do QualisBrasil com Melhores Práticas W3C para publicação e consumo de dados na Web

conjuntos de dados¹⁵. Com a adoção da MP15, obtém-se como benefícios: o Reuso, a Processabilidade, a Interoperabilidade, a Compreensão, e a Confiança dos consumidores de dados no grafo QualisBrasilQB.

- **MP34 - Siga termos da licença** – para sua utilização, deve-se encontrar e seguir os requerimentos de licença a partir do publicador original do conjunto de dados. Para adotar a MP34, o grafo QualisBrasilQB possui os mesmos termos da licença *Creative Commons* usada pelo grafo QualisBrasil. No caso da adoção da MP34, os benefícios nesse caso são os mesmos das MP4 e MP7, sendo o Reuso do conjunto de dados e a Confiança nos dados pelos consumidores de dados no grafo QualisBrasilQB.

O Quadro 3 sumariza o mapeamento entre os benefícios obtidos com a adoção do subconjunto de MPs adotadas no grafo QualisBrasilQB.

Quadro 3 - Resumo do mapeamento dos benefícios W3C alcançados perante a adoção das MPs

MELHORES PRÁTICAS	BENEFÍCIOS							
	Reuso	Compreensão	Interligação	Descoberta	Confiança	Acesso	Interoperabilidade	Processabilidade
MP 02	✓	✓		✓				
MP 03	✓	✓						✓
MP 04	✓				✓			
MP 07	✓				✓			
MP 09	✓		✓	✓			✓	
MP 10	✓		✓	✓			✓	
MP 12	✓							✓
MP 15	✓	✓			✓		✓	✓
MP 34	✓				✓			

Fonte: Elaborado pelos autores (2018).

Dos oito benefícios que podem ser alcançados a partir da adoção das MPs, o benefício de Acesso é o único que ainda não foi obtido. Todavia, este pode ser alcançado com a adoção de outras MPs que estão fora do escopo deste trabalho.

6 CONSIDERAÇÕES FINAIS

No contexto da Web Semântica, apesar do grafo QualisBrasil ter sido desenvolvido conforme princípios de Dados Abertos Conectados, podem haver desafios na geração e no

¹⁵ Disponível em: <<https://www.w3.org/TR/vocab-data-cube/>>

consumo de seus recursos segundo as Melhores Práticas W3C. Um exemplo de MP que pode ser atualizada no grafo QualisBrasil é a MP15. A referida MP sugere o uso de vocabulários padronizados e recomendados internacionalmente. Nesse trabalho, o grafo QualisBrasil é atualizado com o vocabulário multidimensional *Data Cube*. Isso permite a melhoria na geração de dados cientométricos no grafo QualisBrasil, aumentando as possibilidades de reuso de seus recursos na web de dados.

Nesse sentido, nessa pesquisa é apresentado o enriquecimento semântico do grafo QualisBrasil com a adoção de um subconjunto de Melhores Práticas internacionais, descrevendo os benefícios alcançados com a adoção de cada uma das MPs.

Especificamente, quanto à adoção das MP2 e MP3, respectivamente, são disponibilizados metadados descritivos e estruturais, descrevendo as características do conjunto de dados, aumentando o Reuso, a Compreensão, a Descoberta e a Processabilidade dos dados.

Quanto à MP4, no qb:DataSet é fornecida a licença de uso dos dados. Em relação à MP34, a versão proposta do grafo QualisBrasilQB segue os mesmos termos da licença adotada originalmente no grafo QualisBrasil. Tais MPs resultam nos benefícios de Reuso e Confiança dos dados.

Para a adoção da MP7, é fornecida a descrição da data de publicação por meio do metadado dc:issued no qb:DataSet. Isso permite sua legibilidade por agentes (humanos e máquinas), resultando nos benefícios de Reuso e Confiança do conjunto de dados.

Ao adotar as MP9 e MP10, objetiva-se manter as URIs persistentes do grafo QualisBrasil que foram modeladas anteriormente com o vocabulário SCOVO com as URIs da versão proposta com o Vocabulário *Data Cube*. Isso resulta nos benefícios de Reuso, Interligação, Descoberta e Interoperabilidade do grafo QualisBrasilQB.

Quanto à adoção da MP12, os dados RDF do grafo QualisBrasil são armazenados no grafo QualisBrasilQB, no servidor OpenLink Virtuoso, a partir do *endpoint* <<http://lod.unicentro.br/QualisBrasilQB/>>, possibilitando a elaboração de consultas SPARQL para geração de novos conhecimentos. Dessa forma obtém-se os benefícios de Reuso e Processabilidade.

Com a adoção da MP15, o objetivo é enriquecer semanticamente o grafo QualisBrasil com o vocabulário multidimensional *Data Cube* por este ser um modelo de referência W3C. Com a

adoção da MP15, obtém-se os benefícios de Reuso, Processabilidade, Compreensão, Confiança e Interoperabilidade dos recursos de dados.

Assim, foi realizada a integração do grafo QualisBrasil modelado com o vocabulário SCOVO com a representação por meio do vocabulário multidimensional amplamente difundido pelo W3C denominado *Data Cube*. Como resultado, disponibiliza-se na web o grafo RDF QualisBrasilQB, a partir do qual o consumidor de dados pode escolher em qual vocabulário pode realizar as consultas SPARQL.

Dessa forma, contribui-se para a geração de consultas mais ricas semanticamente e atualizadas, aumentando a confiança, o reuso, a compreensão e a descoberta dos dados do grafo QualisBrasilQB, ampliando-se as opções de interligações, interoperabilidade e processabilidade automatizada de seus recursos na web de dados.

Ressalta-se que o uso do vocabulário *Data Cube* para representação de dados multidimensionais é uma recomendação atual do W3C. Entretanto, a semântica deste vocabulário é complexa e abstrata, demandando um maior esforço para sua compreensão e consequente utilização em comparação com a representação com o vocabulário SCOVO.

Como trabalhos futuros, sugere-se a geração de um modelo *Data Cube* que não mantenha o vocabulário SCOVO para identificar qual a implicação quanto à realização das consultas RDF. Também sugere-se a modelagem do grafo QualisBrasil com o vocabulário *Model for Tabular Data and Metadata on the Web* (MODEL, 2018) com o objetivo de comparar as diferentes modelagens quanto ao reuso facilitado dos recursos.

Enriching QualisBrasil Open Data with W3C Best Practices to Publish and Consuming Data on the Web

ABSTRACT

The Linked Open Data initiative encourages to share datasets according to established principles (guidelines and best practices) by the World Wide Web Consortium (W3C). According W3C, for publishing data on the web, we need to deal with interoperability issues among datasets by adopting licenses, vocabularies, and standards. An example of publishing Linked Open Data is the QualisBrasil graph, which provides historical resources of the Qualis index for Information Science researches. However, the publication of QualisBrasil is carried out using the SCOVO Vocabulary, which, despite being recognized for its simplicity, is no longer up-to-date. Therefore, this research aims to represent the QualisBrasil with recent vocabularies as recommended by W3C best practices. The QualisBrasil is enriched with the Data Cube vocabulary, a W3C recommendation for publishing multidimensional data.

As a result, we argue that by adding the Data Cube semantics to QualisBrasil, this effort promotes the trust, reuse, comprehension and the discovery of the graph, expanding the options of linkability, interoperability and automatized processing of its resources in the data web.

Keywords: Semantic Web. Information Representation. Data Cube Vocabulary. Metadata. Data Storage.

REFERÊNCIAS

AUER, S. Introduction to lod2. In AUER, S.; BRYL, V.; TRAMP, C (ed). **Linked Open Data – Creating Knowledge Out of Interlinked Data**. Springer-Verlag, 2014. 215p.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American Magazine**, v. 284, n. 5, p. 34-43, 2001.

CATARINO, M. E. Simple Knowledge Organization System: construindo sistemas de organização do conhecimento no contexto da Web Semântica. **Informação & Tecnologia (ITEC)**, v. 1, n. 1, p. 17-28, 2014.

DATA CUBE VOCABULARY. Disponível em: <https://www.w3.org/TR/vocab-data-cube/>. Acesso em: 13 fev. 2018.

FIGUEIREDO, F. de C.; ALMEIDA, F. G. Ontologias em ciência da informação: um estudo bibliométrico no Brasil. **Ciência da Informação**, v.46, n.1, p.23-33, 2017.

GABRIEL Jr., R. F. Utilização da Web Semântica e RDF em Estudos Métricos da Informação: aplicação na base Brapci. In: XVII Encontro Nacional de Pesquisa em Ciência da Informação, 2016. Salvador. **Anais...** Salvador-BA: UFBA, 2016.

LOD. **Linked Open Data Initiative**. Disponível em: <http://linkeddata.org/>. Acesso em: 14 jun. 2018.

MODEL. **Model for Tabular Data and Metadata on the Web: W3C Recommendation 17 December 2015**. Disponível em: <https://www.w3.org/TR/tabular-data-model/>. Acesso em: 21 set. 2018.

MOSTAFA, S. P.; SANTARÉM SEGUNDO, J. E.; SABBAG, D. M. A. Descrição bibliográfica na era da Web Semântica: por uma nova noção de documento. **Informação & Sociedade: Estudos**, v.26, n.2, p. 25-35, 2016.

RAMALHO, R. A. S. Análise do Modelo de Dados SKOS: sistema de Organização do Conhecimento Simples para a Web. **Informação & Tecnologia (ITEC)**, v. 2, n. 1, p. 66-79, 2015.

RAUTENBERG, S.; HILD, T. A.; SOUZA, L. de. Curadoria Digital e Dados Abertos Conectados: o endpoint lod.unicentro.br como fonte informacional da Web de Dados para estudos

Enriquecendo dados abertos conectados do QualisBrasil com Melhores Práticas W3C para publicação e consumo de dados na Web

bibliométricos e cientométricos. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. **Anais...** Rio de Janeiro: UFRJ, 2018.

RAUTENBERG, S.; MARX, E.; NGOMO, A.; LEHMANN, J. QualisBrasil: Disponibilizando dados via linked open data para estudos cientométricos. In: CONGRESSO LINKED OPEN DATA, 1., 2014, Florianópolis. **Anais...** Florianópolis-SC: PPEGC, UFSC, 2014.

RAUTENBERG, S.; SOUZA, L. de; HAUAGGE, J.; HILD, T.; MICHELON, G.; BURDA, A. Representando Índices Cientométricos como Dados Abertos Conectados. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017. Marília. **Anais...** Marília-SP: PPGCI, Unesp, 2017.

RAUTENBERG, S.; BURDA, A. Linked Open Data para Cientometrica: Compartilhando e Mantendo o índice Qualis na Web de Dados. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais...** São Paulo, 2016.

SANTARÉM SEGUNDO, J. E. 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**, v. 8, n. 2, 2015.

SPARQLIFY. **Sparqlify - Agile Knowledge Engineering and Semantic Web (AKSW)**. Disponível em: <http://aksw.org/Projects/Sparqlify.html>. Acesso em: 21 set. 2018.

VIRTUOSO. **OpenLink Virtuoso Home Page**. Disponível em: <https://virtuoso.openlinksw.com/>. Acesso em: 21 set. 2018.

W3C. **Data on the Web Best Practices. W3C Candidate Recommendation 31/01/2017**. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 14 jul. 2018.