Uso de Dicionário Semântico de Dados na anotação de modelos de dados dimensionais para geração de indicadores de desempenho

Marcello Peixoto Bax
Pós-Doutorado pela Rensselaer Polytechnic Institute (RPI) - Estados Unidos. Doutor em Informática, Anal. Sistemas e Tratamento de Sinal pela Université Montpellier 2 - Sciences et Techniques (UM2) - França. Professor da Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG - Brasil.
http://lattes.cnpq.br/1864473087690223
E-mail: bax.ufmg@gmail.com

Evaldo de Oliveira da Silva
Doutorando Gestão & Organização do Conhecimento pela Universidade Federal de Minas Gerais (UFMG) - Brasil. Mestre em Ciência da Computação pela Universidade de Federal de Viçosa (UFV) - Viçosa, MG – Brasil.
http://lattes.cnpq.br/7337125039379689
E-mail: evaldo.oliveira@gmail.com

Submetido em: 30/04/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

RESUMO

Key Performance Indicators (KPIs) são usados por organizações para avaliar o desempenho de suas atividades, apoiando a decisão. Com esses indicadores, elas reveem seus processos, buscando a sua melhoria contínua. Em modelagem de dados, os modelos dimensionais estruturam os dados agrupando-os em "fatos" e "dimensões". Os fatos são representados por campos numéricos que permitem gerar KPIs. É importante, contudo, seguir técnicas e boas práticas de anotação de dados com metadados que minimizem interpretações divergentes. O modo de anotar dados é ilustrado com a técnica "Dicionário Semântico de Dados" (SDD), que os associa a conceitos e tem potencial para apoiar a geração de KPIs, enriquecendo-os e formalizando-os com ontologias. Seguindo essa técnica, é apresentado um breve experimento que anota um modelo de dados para cálculos de KPIs usando SDDs. Como resultado, o potencial dos SDDs no contexto da geração de KPIs em organizações é examinado. Conclui-se que, além da integração semântica dos dados, outra contribuição é a estruturação formal (em lógica) dos indicadores em grafos de conhecimento fundamentados por ontologias. Finalmente, o experimento contribui para a curadoria dos dados, já que o SDD segue as boas práticas e os princípios FAIR.

Palavras-chave: Modelos Dimensionais. Indicadores de desempenho. KPI. Dicionário de Dados. Ontologia. Anotação semântica. FAIR.

Annotation of data for generation of performance indicators in organizations

ABSTRACT

Key performance indicators (KPIs) are used by organizations to assess their performance, supporting the decision. With these indicators, they review their processes seeking continuous improvement. Dimensional models structure data by grouping it into "facts" and "dimensions." The facts represent numeric fields that leverage the generation of KPIs. However, it is essential to follow data annotation techniques and recommended practices with metadata seeking to minimize divergent interpretations. In the context of the generation of KPIs, described how an annotation occurs according to the method "Semantic Data Dictionary" (SDD), which associates data with concepts to generate these indicators, enriching and formalizing them using ontologies. A "use case" (experiment) of data annotation of a dimensional model for KPI calculations is presented, based on SDDs. As a result, the experiment examines the potential of applying SDDs in the context of generating organizational performance indicators (KPIs). Besides the conceptual integration of the data, it is possible to consider another contribution, which is the formal structuring (in logic) of the KPIs in graphs of knowledge grounded on ontologies. Finally, this work contributes to data curation since the SDD follows acceptable modeling practices (FAIR principles).

Keywords*: Dimensional Data. Performance Indicators. KPI. Data Dictionary. Ontology. Semantic Annotation.*

Anotación de datos para generar indicadores de desempeño en organizaciones

RESUMEN

Las organizaciones utilizan los principales indicadores de desempeño (KPI) para evaluar su desempeño, respaldando la decisión. Con estos indicadores, revisan sus procesos buscando una mejora continua. Los modelos dimensionales estructuran los datos agrupándolos en "hechos" y "dimensiones". Los hechos representan campos numéricos que aprovechan la generación de KPI. Sin embargo, es esencial seguir las técnicas de anotación de datos y las mejores prácticas con metadatos para minimizar las interpretaciones divergentes. En el contexto de la generación de KPIs, describió cómo se produce una anotación según el método "Diccionario de datos semánticos" (SDD), que asocia datos con conceptos para generar estos indicadores, enriqueciéndolos y formalizándolos a través de ontologías. Se presenta un "caso de uso" (experimento) de anotación de datos de un modelo dimensional para los cálculos de KPI, basados en SDD. Como resultado, el experimento examina el potencial de aplicar los SDD en el contexto de la generación de indicadores de desempeño organizacional (KPI). Además de la integración conceptual de datos, es posible considerar otro aporte, que es la estructuración formal (en lógica) de KPIs en grafos de conocimiento basados en ontologías. Finalmente, este trabajo contribuye a la conservación de datos, ya que el SDD sigue prácticas de modelado aceptables (principios FAIR).

Palabras clave*: Modelos dimensionales. Indicadores de desempeño. KPI. Diccionario de datos. Ontología. Anotación semántica.*

1 – INTRODUÇÃO

Um indicador chave de desempenho (KPI ou *Key Performance Indicator*) é um valor que pode ser medido e que demonstra a eficácia da organização em alcançar resultados (PARMENTER, 2015). KPIs permitem avaliar o atingimento de metas e rever processos para melhoria contínua das atividades, criando a base analítica para a tomada de decisões que priorizam as ações avaliadas (empiricamente) como mais relevantes. Eles medem, a título de ilustração, receitas, lucros, preços e custos, atividades, qualidade ou satisfação. Gestores e executivos interpretam KPIs para decidirem com base empírica, científica. Exemplo comum de mensuração é o percentual de aderência da realização de atividades ao que foi planejado anteriormente. Os KPIs podem ser vistos também no meio acadêmico. De acordo com Kolar, Harrison e Gliksohn (2018), eles podem avaliar o grau do alcance de objetivos de instituições de ensino ou programas de pesquisa. Além disso, são insumos para gerenciar e monitorar o atingimento de objetivos e auxiliar no planejamento estratégico. Para Kimball e Ross (2013), a criação de KPIs deve ser disciplinada por boas práticas de nomeação de dados. Em casos em que o conjunto de dados (*datasets*) utilizado para gerar os cálculos não seja de compreensão trivial, nomes podem ser atribuídos segundo diferentes interpretações. Com isso, os KPIs acabam resultando de combinações de dados incompatíveis, comprometendo os valores e prejudicando a tomada de decisão. No âmbito da geração de KPIs, é crucial garantir a qualidade dos dados, e a Curadoria Digital propõe técnicas de descrição de dados com metadados que favorecem a qualidade, a preservação e facilitam a descoberta de novas informações e novos conhecimentos pelo reuso de dados (MEDEIROS, 2018). No entanto, somente a definição dos metadados não basta para extrair dados dos sistemas de informação existentes e compartilhar *datasets*. Dados usados para geração de KPIs vêm de sistemas e modelos de dados distintos e requerem informações adicionais para que seus significados sejam explicitados.

Wise *et al.* (2019) afirmam que a pesquisa e o desenvolvimento na indústria biofarmacêutica também estão se tornando cada vez mais orientados por dados. Os autores destacam os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*), propostos por Wilkinson *et al.* (2016), para aumentar a eficiência e a eficácia da gestão de dados científicos, não somente na produção biofarmacêutica, mas em outras áreas da indústria. Projetos de *data warehouse* têm sido construídos para alavancar a produção (WISE *et al.*, 2018; VAUDANO, 2013). Porém, Wise *et al.* (2019) destacam a necessidade de utilizar procedimentos para anotar e gerenciar os dados nesses projetos também, visando a atender aos princípios FAIR.

Rashid *et al.* (2020) avaliaram a aderência dos SDDs aos princípios FAIR, comparando-os a outras abordagens de anotação de dados. Na comparação, o SDD recebeu a maior nota (detalhes na Seção 3). Geralmente acompanhando *datasets*, os Dicionários de Dados tradicionais (não semânticos) facilitam o gerenciamento de dados. Para formalizar a descrição dos dados com os metadados dos dicionários, ontologias podem ser aplicadas. Elas enriquecem semanticamente e formalizam logicamente o significado dos dados, evitando interpretações discrepantes.

Este artigo apresenta um "caso de uso" de anotação de dados de um modelo dimensional para cálculos de KPIs, baseado na abordagem de Dicionários Semânticos de Dados (*Semantic Data Dictionary*, SDD) proposta por Rashid *et al.* (2017). Os SDDs contribuem para a curadoria dos dados e estão alinhados aos princípios FAIR. Trata-se do relato de um experimento que procurou examinar o potencial de uso de SDDs no contexto da geração de indicadores de desempenho organizacional (KPIs). A abordagem emprega ontologias para oferecer uma descrição detalhada de *datasets* a ponto de ser automatizada e tratada por computador. Ela permite o enfoque na semântica dos dados, incluindo informações que possam prontamente ser processadas.

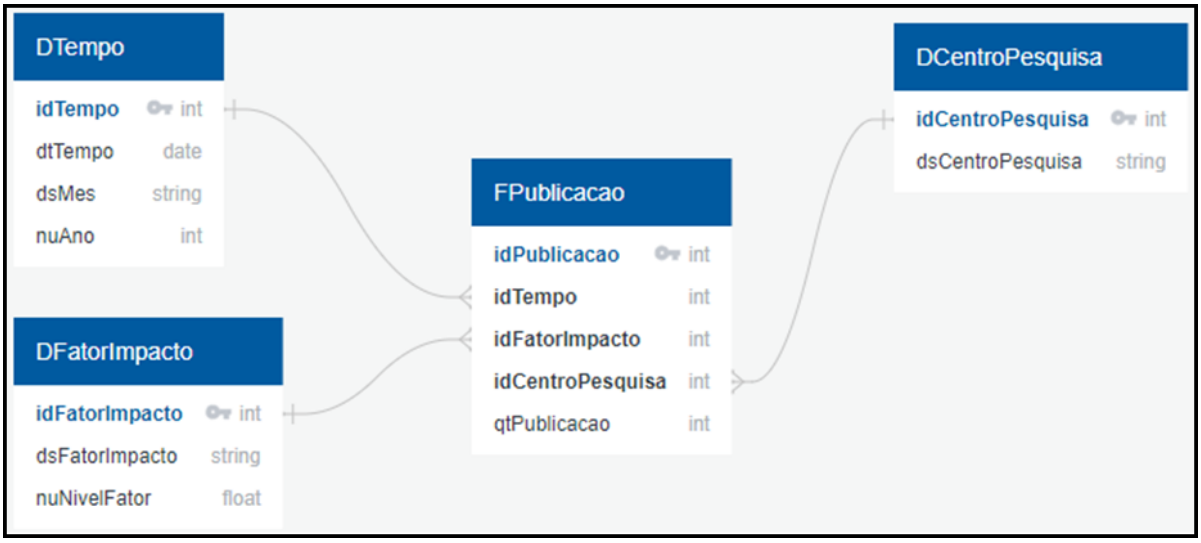
Ao considerar essas características dos SDDs, argumenta-se que é alcançada uma representação legível e padronizada por máquina para o registro de metadados com base em *datasets*, e em metodologias que usam linguagens de mapeamento. Isso é alcançado simplificando os requisitos de conhecimento de programação, separando a parte de anotação da abordagem do componente de software. Uma vantagem de aplicar SDDs em modelos dimensionais é alinhar e harmonizar interpretações de conceitos e diferentes escalas e unidades de medida que descrevem os dados. Pode-se assim integrar semanticamente dados de diferentes fontes e unidades de negócios ou até mesmo de organizações diferentes. Rashid *et al.* (2020) comparam a abordagem SDD com trabalhos correlatos por meio de métricas que possibilitam compreender as diferentes iniciativas de integração de dados. De acordo com os autores constatou-se desempenho superior da abordagem de SDD em relação, sobretudo, aos dicionários de dados tradicionais.

O texto está organizado como segue: a Seção 2 traz o conceito de modelagem dimensional de dados como KPIs e sua anotação por meio de ontologias. A Seção 3 descreve o processo de anotação usando o SDD. A Seção 4 aplica o SDD, relatando a anotação necessária para criar KPIs voltados ao monitoramento de quantitativos de publicações científicas em instituições de pesquisa. A Seção 5 analisa e relaciona os trabalhos correlatos de destaque na literatura. A Seção 6 traz as conclusões e considerações finais, além de sugerir trabalhos futuros.

2 – MODELANDO KPIS COM ONTOLOGIAS

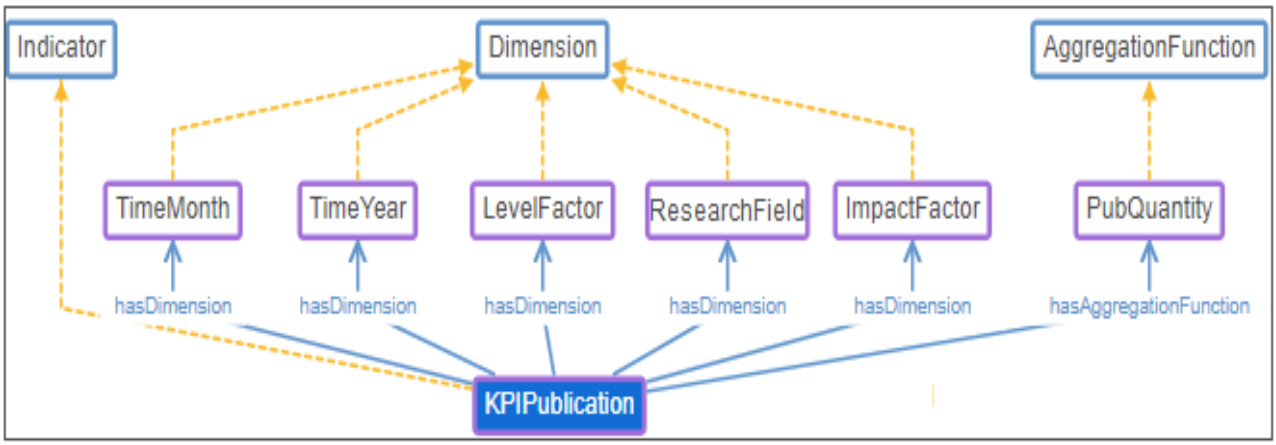
Um modelo de dados dimensional agrupa dimensões ou facetas ao redor de dados numéricos, chamados de fatos. A análise dos fatos usa as dimensões (ou facetas) para combinar filtros que atendem às necessidades do usuário e apoiam a tomada de decisão (KIMBAL; ROSS, 2013). O modelo dimensional da figura 1, organizado como uma estrela (fatos no centro), permite calcular os montantes de publicações por meio das seguintes dimensões: fator de impacto, centro de pesquisa, mês e ano. A tabela fato (FPublicacao) possui chaves (estrangeiras) oriundas das outras tabelas e um atributo quantidade de publicações (qtPublicacao), com o qual se extrai os KPIs a partir de funções de agregação e fórmulas.

Figura 1 – Exemplo de modelo dimensional de dados



Fonte: Elaborada pelos autores.

Figura 2 – Ontologia KPIOnto



Fonte: Adaptado pelos autores com base em Diamantini, Potena e Storti (2016).

Anotar dados semanticamente exige compreender o domínio representado por um modelo conceitual. Deve-se selecionar os dados e anotar os termos existentes que designam conceitos do domínio, explicitando e formalizando a sua semântica com o uso de ontologias. Esta anotação permite gerar fragmentos de conhecimento do domínio que podem ser representados por grafos de conhecimento (Hogan *et al.,* 2020). Um grafo de conhecimento representa objetos de interesse e conexões entre eles. Restrições na sua estrutura são impostas por meio de ontologias. Além disso, grafos de conhecimento permitem que pessoas e diferentes aplicações reutilizem as definições nele modeladas e gerem inferência de novos fatos, enriquecendo o conhecimento e o seu compartilhamento (Pan *et al.*, 2017).

Eles agregam entidades (e seus atributos), relacionando-as por relacionamentos expressivos e formais. Nessa perspectiva, eles são vistos como ontologias, nas quais as entidades formam o vocabulário que descreve os entes/indivíduos (instâncias) mais salientes e recorrentes do domínio.

O padrão RDF[1] (W3C, 2014) pode ser usado para representar grafos de conhecimento a partir de declarações no formato de triplas. Tais declarações (RDF) são geradas como fragmentos do conhecimento, como resultado da anotação por SDD. A conceitualização do domínio por meio da formalização semântica de ontologias é uma das premissas dos SDDs e este artigo utiliza a ontologia KPIOnto[2] de Diamantini, Potena e Storti (2016).

Ela nos permitirá anotar e alinhar conceitualmente a compreensão de diferentes profissionais sobre os KPIs empregados. A KPIOnto constitui-se de classes como: *Indicator, Dimension, AggregationFunction* e *Formula*; sendo "*Indicator*" a classe principal. Ela especifica um indicador pelas propriedades *hasDimension, hasFormula* e *hasAggrFunction* (para uso de funções de agregação). Como forma de representação das relações existentes na KPIOnto, a figura 2 apresenta instâncias que se relacionam com os conceitos sobre modelos dimensionais discutidos anteriormente. As instâncias foram criadas com o WebProtegé (MUSEN, 2015). Dessa forma, como mostra a figura 2, a instância *KPIPublication* (em azul) é um tipo de *Indicator*.

[1] *Resource Description Framework.*
[2] KPIOn to, https://github.com/KDMG/kpionto

A relação "*is-a*" é representada pela linha pontilhada amarela. A instância *KPIPublication* relaciona-se com os demais conceitos da KPIOnto, tais como *AggregationFunction* e *Dimension* pelas propriedades *hasAggrFunction* e *hasDimension*, respectivamente, representadas por linhas azuis sólidas. A instância *PubQuantity* é um de tipo de *AggregationFunction* pois representa o somatório da quantidade de publicações. As instâncias de *researchField, LevelFactor, ImpactFactor, TimeYear* e *TimeMonth* são um tipo de *Dimension*.

3 – ANOTANDO DADOS COM DICIONÁRIOS SEMÂNTICOS DE DADOS

Rashid *et al.* (2017) propõem o Dicionário Semântico de Dados (SDD) para anotar *datasets*. O SDD é uma abordagem de anotação de dados que emprega um conjunto de padrões de metadados fundamentados em ontologias que descrevem objetos (representados por dados) em classes e relacionamentos. A fim de enriquecer os dados presentes em um *dataset*, a anotação por SDD associa-os a conceitos/classes das ontologias. Recomenda-se o uso da ontologia *Semanticscience Integrated Ontology* (SIO) (2020), que fornece propriedades para descrever os relacionamentos entre objetos e atributos como modelo de representação do conhecimento e que permite também facilitar a descoberta do conhecimento. A abordagem deve ser orientada por especialistas de domínio, engenheiros de conhecimento (ontologistas) ou cientistas da informação que compreendem o domínio, tanto dos conceitos relacionados na ontologia, quanto dos *datasets* anotados. A anotação é um processo manual e utiliza um conjunto de documentos (templates de metadados) explicados mais abaixo no texto (*InfoSheet*; *Dictionary Mapping*; *CodeBook*; *Code Mapping*; *TimeLine*; *Properties Table*). A Seção 3 detalha esses templates.

A ferramenta *sdd2rdf*[3] (SEMANTIC DATA DICTIONARY, 2019) é um *script/software* que interpreta o SDD e converte os dados do *dataset* descrito por ele em um grafo de conhecimento expresso no padrão RDF.

[3] https://github.com/tetherless-world/SemanticDataDictionary

Para exemplificar o acesso aos dados anotados no grafo, o *sdd2rdf* cria alguns exemplos de consultas SPARQL[4]. O grafo de conhecimento (no formato RDF) gerado pelo script utiliza a ontologia e possibilita a interoperabilidade dos dados. É a formalização do vocabulário da anotação que abre caminho para interoperabilidade dos dados, que podem ser integrados de fontes diversas. Após escolher quais dados do *dataset* anotar, segue-se para a criação dos artefatos abaixo, em um processo cujas etapas são descritas a seguir:

- *Ontologia de domínio.* A ontologia formaliza os conceitos do problema de pesquisa. Deve-se buscar reutilizar ontologias consolidadas no domínio do problema;
- *Dictionary Mapping (DM).* Anota a semântica das colunas do *dataset*. Cada linha do DM mapeia uma coluna do *dataset*, formalizando-a conceitualmente, explicitando suas relações com os outros dados do mesmo *dataset*, bem como a sua proveniência[5] [6];
- *CodeBook*. Um *codebook* estrutura os dados categóricos[7] de um *dataset* mapeando-os para conceitos correspondentes na ontologia. Dessa forma, o cientista da informação preocupa-se com o tratamento dos dados, criando categorias e estabelecendo um código para cada uma. O *codebook* possui os seguintes campos para anotação: Coluna (entidade a ser anotada), Código, Descrição e Classe da Ontologia;

[4] *SPARQL Protocol and RDF Query Language* - Linguagem de consulta elaborada pelo W3C para acesso a dados em formato RDF.

[5] Proveniência de dados é a descrição das origens de um dado e o processo pelo qual ele chegou a uma base de dados (BUNEMAN; KHANNA; WANG-CHIEW, 2001).

[6] Como forma de anotar a proveniência do dado, o DM mapeia as entidades pré-existentes que são relevantes na anotação dos dados por meio do campo "*wasDerivedFrom*". Já o campo "*wasGeneratedBy*" descreve a atividade de geração associada à anotação de dados no DM (RASHID *et al*, 2020).

[7] Dados categóricos são dados agrupados. Podem derivar de observações feitas de dados qualitativos ou de observações de dados quantitativos agrupados em determinados intervalos (AGRESTI, 2003).

- *Infosheet*. Organiza os metadados de descrição do SDD. É importante principalmente para o seu compartilhamento em redes, conforme o princípio de "encontrabilidade" FAIR;
- *Grafo de Conhecimento (RDF).* Resulta da interpretação da dupla "SDD (templates de metadados) + Dados" pelo *script sdd2rdf*, gerando o grafo RDF. Caso seja necessário persistir os dados, o usuário pode armazenar o grafo em *triplestore*[8] para consulta posterior dos dados.

Inicialmente, os dados mapeados para as ontologias pelo SDD são as colunas do próprio *dataset*. Os objetos caracterizados nos *datasets* podem estar implicitamente representados. Os objetos implícitos serão explicitados no SDD e formalizados no grafo final gerado. A explicitação dos objetos implícitos favorece a integração semântica dos dados nos níveis conceituais mais abstratos do projeto, permitindo alinhar e homogeneizar, harmonizar, interpretações de conceitos que descrevem os dados que se deseja integrar.

Rashid *et al*. (2020) elaboraram métricas a fim de avaliar a aderência dos SDDs aos princípios FAIR. O SDD foi avaliado juntamente com outras abordagens de dicionário de dados, tais como: dicionários de dados tradicionais, abordagens envolvendo linguagens de mapeamento e ferramentas gerais de integração de dados. Para medir se o SDD atendeu a cada métrica, os autores forneceram um valor entre 0, 0,5 ou 1, dependendo do quanto o SDD responde a um parâmetro de avaliação. Em comparação às outras abordagens de dicionários de dados, o SDD recebeu nota 1.

Para pontuar a métrica localizável, foi avaliado o uso de identificadores persistentes exclusivos, como URLs, bem como a inclusão de metadados pesquisáveis, para que o conhecimento seja descoberto na web. O SDD permite a representação do conhecimento, podendo ser persistente e detectável. Para a métrica acessível, os autores consideraram que a representação de conhecimento gerada pelo SDD permite que os dados disponíveis estejam acessíveis abertamente, podendo ser divulgados publicamente. Em relação à métrica interoperável, os usos de vocabulários estruturados e de ontologias, como melhores práticas compatíveis com RDF, foram analisados. O SDD permite a representação do conhecimento a partir de vocabulários formais ou ontologias que são compatíveis com RDF. Sendo assim, para testar se o SDD permite reutilização dos dados, foi analisada a reutilização irrestrita deles. Também foi discutido se os metadados são detalhados o suficiente para um novo usuário entender. Constatou-se que o SDD permite o reuso irrestrito dos dados disponíveis publicamente.

4 – ANOTANDO DADOS PARA GERAR KPIS

Descreve-se, nesta seção, o exemplo de anotação de dados para geração de KPIs dada a necessidade de acompanhar índices quantitativos de publicação em centros de pesquisa. O modelo relacional da figura 1 foi utilizado como esquema de dados para recuperação dos *datasets*. Foram inseridos dados fictícios nas tabelas *DTempoMes, DTempoAno, DFatorImpacto, FPublicação, DCentroPesquisa*, onde o prefixo "D" é a designação de tabela de dimensão e o prefixo "F" é relativo à tabela fato para análise e geração de indicadores. O PostgreSQL foi usado como sistema gerenciador de banco de dados para persistir os *datasets*.

Para gerar o *dataset* a ser anotado, foi necessário relacionar as tabelas de dados (conf. figura 1). Esse estabelecimento de relação permitiu a criação de uma visão de dados (ou *view*) definida abaixo, gerando, como resultado, o *dataset* representado na figura 3 e que possui as colunas e dados correspondentes com a *view*.

[8] O armazenamento em *triplestore* (ou RDF) é um banco de dados com o propósito de armazenar e recuperar triplas por meio de consultas semânticas (DBPEDIA, 2020).

Figura 3 − *Dataset* a ser anotado

Id_Kpi	TimeMonth	TimeYear	ReasearchField	ImpactFactor	LevelFactor	PubWauantity
1	JANEIRO	2000	1	1	4	63
2	JANEIRO	2000	3	2	7	6

Fonte: Elaborada pelos autores.

Além da recuperação do *dataset*, é importante ressaltar que os elementos (*Dictionary Mapping*, *CodeBook* e *Infosheet*) utilizados em um SDD são definidos por meio de arquivos em formato CSV[9]. Abaixo segue a descrição da execução do processo de anotação:

- *Ontologia de Domínio*. As ontologias utilizadas foram a KPIOnto e a SIO. A KPIOnto possui conceitos consensuados que descrevem KPIs; já a SIO é a ontologia padrão utilizada nos SDDs;
- *Dictionary Mapping (DM)*. O DM (tabelas 1 e 2) mapeia, para ontologias (SIO e KPIOnto), as seguintes propriedades dos KPIs: dimensões *TimeMonth* (tempo na dimensão mês), *TimeYear* (tempo na dimensão ano), *ResearchField* (descrição do centro de pesquisa), *ImpactFactor* (descrição do fator de impacto) e *LevelFactor* (nível do fator de impacto); e função de agregação de *PubQuantity* (quantidade de publicação). Especificamente em relação à tabela 2, é possível identificar conceitos implícitos sobre o domínio. Dessa forma, os dados mapeados são de um *reseachInstitute*.Os dados implícitos são anotados no artefato DM do SDD para que sejam também enriquecidos semanticamente. Com isso, novos dados anotados aparecem, servindo de pontes para representar mais amplamente o conhecimento. É uma preparação para considerar novos dados na análise explicitando relacionamentos que até o momento estavam implícitos;
- *Codebook*. A tabela 3 traz o *Codebook*, que descreve os dados categoriais do *dataset,* mapeando-os para ontologias, nesse caso, a KPIOnto. São mapeadas as dimensões *DTempoMes, DTempoAno, DCentroPesquisa, DFatorImpacto, DNivelImpacto* e a função de agregação *QtPublicacao*;
- *Infosheet*. A tabela 4 possui os metadados (e seus vocabulários, *dct*[10], *owl*[11], *schema*[12]) que descrevem o SDD a fim de melhorar a sua "encontrabilidade" na rede (um princípio FAIR):
  - dct:creator: responsável pelo preenchimento;
  - dct:contributor: contribuidores na criação do *Infosheet* e execução do processo;dct:created: data de criação;
  - dct:description: propósito do SDD;
  - owl:imports: endereço das ontologias utilizadas no SDD;
  - schema:keywords: palavras-chave;
  - dct:publisher: responsável por publicar;
  - dct:title: título do SDD;
- *Grafo de Conhecimento*. O grafo RDF representando o conhecimento sobre os KPIs é persistido no Virtuoso[13], onde é possível manipular os dados empregando a linguagem SPARQL (ERLING; MIKHAILOV, 2009). Segue abaixo o trecho do grafo RDF (em sintaxe TTL[14]) que representa a anotação e integração semântica dos dados nas primeiras 4 linhas da tabela de dados da figura 3. Cada linha do *dataset* tem seus dados anotados pelos metadados do DM usando ontologias (tabelas 1 e 2).

[9] *Comma-Separated-Values*

[10] http://purl.org/dc/terms/
[11] https://www.w3.org/2002/07/owl
[12] http://schema.org/
[13] https://virtuoso.openlinksw.com/
[14] https://www.w3.org/TR/turtle/

Tabela 1 − Especificação do DM para objetos explícitos

Column	Attribute	sio:AttributeOf	rdfs:Label
Id_Kpi	sio:Identifier	??kpiPublication	Identificador do KPI
ResearchField	kpiOnto:hasDimension	??kpiPublication	Descrição do Centro de Pesquisa
ImpactFactor	kpiOnto:hasDimension	??kpiPublication	Descrição do Fator de Impacto
LevelFactor	kpiOnto:hasDimension	??kpiPublication	Nível do Fator de Impacto
TimeMonth	kpiOnto:hasDimension	??kpiPublication	Mês de Apuração do KPI
TimeYear	kpiOnto:hasDimension	??kpiPublication	Ano de Apuração do KPI
PubQuantity	kpiOnto:hasAggFunction	??kpiPublication	Quantidade de Publicação

Fonte: Elaborada pelos autores.

Tabela 2 − Especificação do DM para objetos implícitos

Column	**Entity**	**Relation**	**sio:InRelationTo**
??kpiPublication	kpiOnto:Indicator	kpiOnto:isUsedBy	??researchInstitute
??researchInstitute	sio:Institute	kpiOnto:hasKpi	??kpiPublication

Fonte: Elaborada pelos autores.

Tabela 3 − Codebook - dimensões DTempo, DFatorImpacto e DCentroPesquisa

Column	Code	Label	Class
DCentroPesquisa	1	CIÊNCIA DA INFORMAÇÃO	kpionto:researchField
DCentroPesquisa	2	CIÊNCIA DA COMPUTAÇÃO	kpionto:researchField
DCentroPesquisa	3	LINGUÍSTICA	kpionto:researchField
DFatorImpacto	4	IMPACTO ENTRE 2 E 4	kpionto:ImpactFactor
DFatorImpacto	5	IMPACTO ENTRE 5 E 7	kpionto:ImpactFactor
DFatorImpacto	6	IMPACTO ENTRE 7 E 10	kpionto:ImpactFactor
DNivelImpacto	7	4	kpionto:LevelFactor
DNivelImpacto	8	7	kpionto:LevelFactor
DNivelImpacto	9	5	kpionto:LevelFactor
DTempoMes	10	Janeiro	kpionto:TimeMonth
DTempoMes	11	Fevereiro	kpionto: TimeMonth
DTempoMes	12	Março	kpionto: TimeMonth
...			
DTempoAno	25	2000	kpionto: TimeYear
DTempoAno	26	2001	kpionto: TimeYear
DTempoAno	27	2002	kpionto: TimeYear
DTempoAno	28	2003	kpionto: TimeYear
DTempoAno	29	2004	kpionto: TimeYear
...			

Fonte: Elaborada pelos autores.

Tabela 4 − Especificação do Infosheet

Atributo	**Valor**
dct:creator	Marcello P. Bax e Evaldo de Oliveira da Silva
dct:contributor	Marcello P. Bax
dct:created	20/04/2019
dct:description	Anotação semântica do dicionário de dados para geração do KPI de Publicação
owl:imports	http://semanticscience.org/ontology/sio-subset-labels.owl
schema:keywords	KPI, Publicação
dct:publisher	Evaldo de Oliveira da Silva
dct:title	Geração de KPIs com base na anotação semântica de modelos de dados dimensionais

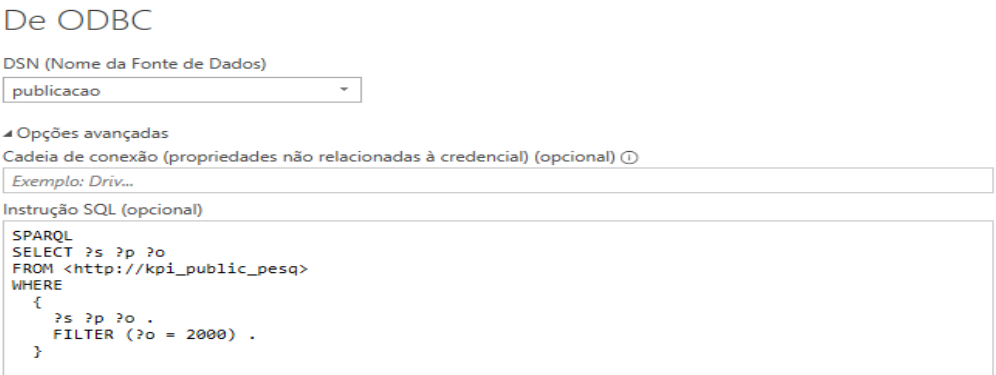
Fonte: Elaborada pelos autores.

```
example-kb:Id_Kpi-deb0 a example-kb:Id_Kpi , sio:Identifier ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Identificador do KPI"^^xsd:string ;
  sio:hasValue "1"^^xsd:integer .

example-kb:TimeMonth-50ac a example-kb:TimeMonth , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Mes"^^xsd:string ;
  sio:hasValue "JANEIRO"^^xsd:string .

example-kb:TimeYear-65ea a example-kb:TimeYear , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Ano"^^xsd:string ;
  sio:hasValue "2000"^^xsd:integer .

example-kb:ResearchField-896d a example-kb:ResearchField, kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Descricao do Centro de Pesquisa"^^xsd:string ;
  sio:hasValue "1"^^xsd:integer .

example-kb:ImpactFactor-3e96 a example-kb:ImpactFactor , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Descricao do Fator de Impacto"^^xsd:string ;
  sio:hasValue "1"^^xsd:integer .

example-kb:LevelFactor-e2d7 a example-kb:LevelFactor , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Nivel do Fator de Impacto"^^xsd:string ;
  sio:hasValue "4"^^xsd:integer .

example-kb:PubQuantity-8995 a example-kb:PubQuantity , kpiOnto:hasAggFunction ;
  sio:isAttributeOf example-kb:KpiPublication-ecc6 ;
  rdfs:label "Quantidade de Publicacao"^^xsd:string ;
  sio:hasValue "6"^^xsd:integer .

example-kb:Id_Kpi-f138 a example-kb:Id_Kpi , sio:Identifier ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Identificador do KPI"^^xsd:string ;
  sio:hasValue "2"^^xsd:integer .

example-kb:TimeMonth-50ac a example-kb:TimeMonth , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Mes"^^xsd:string ;
  sio:hasValue "JANEIRO"^^xsd:string .

example-kb:TimeYear-65ea a example-kb:TimeYear , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Ano"^^xsd:string ;
  sio:hasValue "2000"^^xsd:integer .

example-kb:ResearchField-08d4 a example-kb:ResearchField , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Descricao do Centro de Pesquisa"^^xsd:string ;
  sio:hasValue "3"^^xsd:integer .

example-kb:ImpactFactor-ed7c a example-kb:ImpactFactor , kpiOnto:hasDimension ;
  sio:isAttributeOf example-kb:KpiPublication-e9c1 ;
  rdfs:label "Descricao do Fator de Impacto"^^xsd:string ;
  sio:hasValue "2"^^xsd:integer .
```

- *Visualização dos dados.* Um painel (dashboard) construído em Ms-PowerBI (MICROSOFT, 2020) conecta-se ao Virtuoso via ODBC (Open Database Connectivity) e executa consultas SPARQL para ilustrar como os dados, extraídos do grafo, podem ser visualizados de diferentes formas. A consulta SPARQL apresentada na Figura 4 é utilizada para extrair os dados do grafo de conhecimento, representando os dados sobre as publicações realizadas no ano 2000, para serem carregados para um arquivo do Ms-PowerBI. A partir da carga dos dados, deve ocorrer a sua "transformação" em formato RDF, para visualização de indicadores e métricas por meio dos dashboards.

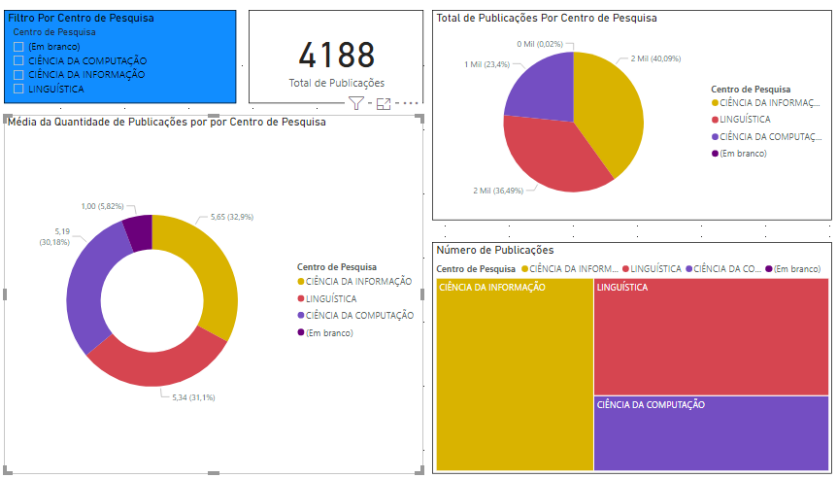
É importante ressaltar que é possível automatizar a carga e a transformação dos dados extraídos do Virtuoso quando o dashboard é criado no Ms-PowerBI, mesmo que novos dados sejam inseridos, alterados e excluídos nas fontes de origem. A figura 5 apresenta a visualização dos indicadores gerados, tais como o total geral de publicações, o percentual, a média e o número de publicações por centro de pesquisa. A figura 6 apresenta a visualização de outros indicadores, usando, como dimensão, o fator de impacto. Nesse caso, foram gerados os seguintes indicadores: mediana do número de publicações por fator de impacto e total de publicações por fator de impacto.

Figura 4 – Conexão via ODBC a partir de uma consulta SPARQL



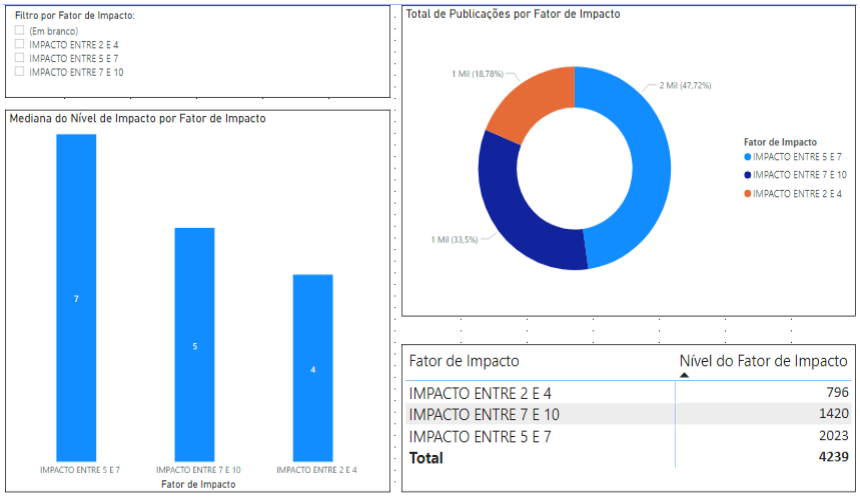
Fonte: Elaborada pelos autores.

Figura 5 – *Dashboard* gerado a partir dos RDFs com indicadores por Centro de Pesquisa



Fonte: Elaborada pelos autores.

Figura 6 - *Dashboard* gerado a partir dos RDFs com indicadores por Fator de Impacto



Fonte: Elaborada pelos autores.

5 – TRABALHOS CORRELATOS

Kritikos, Plexousakis e Woitsch (2017) afirmam que os dados conectados (*Linked Data*[15]) são um mecanismo para integrar fontes distintas, permitindo realizar inferências para derivar conhecimento novo. Eles utilizam essa ideia no contexto de negócios (BPaaS, *Business Process as a Service*), a fim de coletar e vincular informações originadas de diferentes sistemas. Propõem o uso de ontologias para melhorar a comparação de KPIs gerados dos dados integrados entre os sistemas. Wetzstein, Ma e Leymann (2008) indicam que KPIs sejam modelados por analistas de negócios que exploram anotações semânticas de processos de negócios. Os modelos de KPI são automaticamente calculados para serem geridos por meio de um painel de monitoramento em tempo real. Kourtesis, Alvarez-Rodrigues e Paraskakis (2014) sugerem uma estrutura semântica para gerenciamento de QoS (*Quality of Service*). Eles utilizam abordagens para o gerenciamento de QoS baseado em semântica, bem como os principais métodos e técnicas para explorar diversos dados. Silva *et al.* (2018) propõem um conjunto de funções para compor a estrutura semântica necessária à definição de dicionário de dados. Apresentam, ainda, como a estrutura semântica está relacionada à configuração sintática dos dicionários de dados, a fim de identificar padrões que possam ser usados no desenvolvimento de procedimentos para extração de informações e modelos semânticos.

O caso de uso apresentado neste artigo se diferencia dos trabalhos acima por utilizar a modelagem ontológica, aplicando a abordagem SDD para anotar dados de KPIs. A anotação é realizada manualmente, por especialistas de domínio. O resultado é a geração de grafos de conhecimento a partir de uma ontologia e de *templates* de metadados, sendo úteis para inferir novos conhecimentos, que podem melhorar a tomada de decisão dentro das organizações.

6 – CONSIDERAÇÕES FINAIS

No contexto organizacional, a modelagem conceitual adequada dos dados envolve a interpretação e negociação de significados sobre entidades, relacionamentos e regras de negócios, que ocorrem naturalmente na comunicação entre os vários atores (ou "partes interessadas"). As vantagens do SDD atingem pleno potencial em cenários onde a integração de fontes de dados diversas (internas ou externas) se faz necessária para enriquecer os dados que se quer analisar.

[15] O termo *Linked Data* refere-se a um conjunto de melhores práticas para publicar e conectar dados estruturados na Web (BIZER; HEATH; BERNERS-LEE, 2011).

O processo apresentado visa a organizar etapas para anotação com SDDs e geração do grafo de conhecimento (em RDF), representando formalmente o conjunto de fatos originados da combinação de dados de diferentes fontes. Um exemplo de geração de KPI, usando como fonte um modelo dimensional, para avaliar critérios de desempenho em função de publicações científicas produzidas por institutos de pesquisa ilustrou o processo, constituindo uma validação preliminar do método.

Argumentou-se que os SDDs contribuem para organizar e integrar dados oriundos de diferentes nichos da organização ou fora dela (por exemplo, via Web), gerando informações que estruturam conhecimentos sobre diversos indicadores empresariais (KPIs). Isso facilita os alinhamentos semânticos sobre os KPIs a partir de uma abordagem de modelagem de dados ampla, do tipo *top down*, e não apenas *bottom up*. O processo proposto permite associar dados de *datasets* a conceitos consensuados com a finalidade de gerar KPIs, enriquecendo-os e formalizando-os com ontologias. Além da simplificação da integração conceitual consensuada dos dados, outra contribuição é a estruturação formal (em lógica) dos KPIs em grafos de conhecimento fundamentados por ontologias.

Outras contribuições podem ser consideradas, como, a título de ilustração, para a comunidade da ontologia KPIOnto, que foi reutilizada no SDD. A abordagem descrita pode também contribuir para integrar fontes de dados heterogêneas, a exemplo de quando há necessidade de organizações diferentes trocarem informações sobre KPIs, como forma de monitoramento dos indicadores ou para atendimento a marcos regulatórios (estabelecidos por agências de regulação de diferentes mercados, como: energia, petróleo, saúde). Essa situação acontece frequentemente entre órgãos do governo, agências reguladoras e organizações empresariais, onde projetos de SDDs podem permitir o alinhamento conceitual de estados informacionais entre as organizações (por exemplo, sua situação financeira).

Nesse aspecto, o governo brasileiro já considera a dimensão semântica no desenvolvimento e na manutenção de ontologias, bem como em outros recursos de organização da informação, em vista de melhorar a interoperabilidade e a troca de informações, por meio do *e-Ping* (Interoperabilidade de Governo Eletrônico) (BRASIL, 2018). Finalmente, os SDDs contribuem para a curadoria dos dados, já que seguem boas práticas de modelagem (princípios FAIR).

Futuras pesquisas investigarão como a modelagem via SDD, tal como apresentada neste artigo, constituiria alternativa vantajosa à modelagem dimensional clássica (do tipo "*data mart*" ou "*data warehouse*"). Uma hipótese é que a flexibilidade de modelos conceituais ontológicos "livres de esquemas" (*schema free*) traria vantagens para a geração de KPIs no contexto analisado. Poderia, por exemplo, tornar a evolução do conhecimento sobre os indicadores de desempenho das organizações mais organizado, flexível, incremental e semanticamente enriquecido pela explicitação de sua semântica formal, advinda do uso de ontologias representadas em Lógica de Descrições (*Description Logic*) (KRÖTZSCH; SIMANCIK; HORROCKS, 2012).

REFERÊNCIAS

AGRESTI, A. *Categorical data analysis.* 2nd ed. New Jersey: Wiley, 2003.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: The story so far. *In:* SHETH, A. *Semantic services, interoperability, and web applications*: emerging concepts. Hershey, USA: IGI Global, 2011. p. 205-227.

BRASIL. Programa de Governo Eletrônico Brasileiro. *e-PING Padrões de Interoperabilidade de Governo Eletrônico*. Brasil, Comitê Executivo de Governo Eletrônico, nov. 2018. Disponível em: http://eping.governoeletronico.gov.br/. Acesso em: mar. 2021.

BUNEMAN, P.; KHANNA, S.; WANG-CHIEW, T. Why and where: A characterization of data provenance. *In:* VAN DEN BUSSCHE, J.; VIANU V. (ed.). *Database Theory*: ICDT 2001. Berlin: Springer, 2001. p. 316-330. DOI: https://doi.org/10.1007/3-540-44503-X_20.

DBPEDIA. *About:* Triplestore. [*S.l.*], 2020. Disponível em: http://dbpedia.org/page/Triplestore. Acesso em: 16 set. 2020.

DIAMANTINI, C.; POTENA, D.; STORTI, E. SemPI: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators. *Future Generation Computer Systems*, [*s.l.*], v. 54, p. 352-365, jan. 2016. DOI: https://doi.org/10.1016/j.future.2015.04.011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167739X1500103X. Acesso em: mar. 2021.

ERLING, O.; MIKHAILOV, I. RDF Support in the Virtuoso DBMS. *In:* PELLEGRINI, T.; AUER, S.; TOCHTERMANN, K.; SCHAFFERT, S (ed.). *Networked Knowledge-Networked Media:* integrating knowledge management, new media technologies and semantic systems. Berlin: Springer, 2009. p. 7-24.

FEW, S. *Information dashboard design:* The effective visual communication of data. [*S.l.*]: O'Reilly Media, 2006.

HOGAN, A. *et al.* Knowledge Graphs. *ArXiv preprint*, arXiv: 2003.02320, Mar. 2020. Disponível em: https://arxiv.org/abs/2003.02320. Acesso em: mar. 2021.

KIMBALL, R.; ROSS, M. *The data warehouse toolkit*: the definitive guide to dimensional modeling. 3rd ed. [*S.l.*]: Wiley, 2013.

KOLAR, J.; HARRISON, A.; GLIKSOHN, F. Key performance indicators of Research Infrastructures. *Central European Research Infrastructure Consortium*, Italy, ago. 2018. Disponível em: https://www.ceric-eric.eu/2018/08/30/key-performance-indicators-of-research-infrastructures/. Acesso em: 30 ago. 2018.

KOURTESIS, D.; ALVAREZ- RODRÍGUEZ, J. M. ; PARASKAKIS, I. Semantic-based QoS management in cloud systems: Current status and future challenges. *Future Generation Computer Systems*, [*s.l.*], v. 32, p. 307-323, 2014. DOI: https://doi.org/10.1016/j.future.2013.10.015. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167739X1300232X. Acesso em: mar. 2021.

KRITIKOS, K.; PLEXOUSAKIS, D.; WOITSCH, R. Towards Semantic KPI Measurement. *In:* INTERNATIONAL CONFERENCE ON CLOUD COMPUTING AND SERVICES SCIENCE, 7., 2017, Portugal. *Proceedings* […]. Portugal: CLOSER, 2017. p. 91-102.

KRÖTZSCH, M.; SIMANCIK, F., HORROCKS, I. A description logic primer. *ArXiv preprint*, arXiv:1201.4089, jan. 2012. Disponível em: https://arxiv.org/abs/1201.4089. Acesso em: mar. 2021.

MEDEIROS, C. B. Gestão de Dados Científicos: da coleta à preservação. *SciELO em Perspectiva*, [*s.l.*], 22 jun. 2018. Disponível em: https://blog.scielo.org/blog /2018/06/22/gestao-de-dados-cientificos-da-coleta-a-preservacao/#.XXZ82ChKjIV. Acesso em: 4 set. 2019.

MICROSOFT. *PowerBI*. [*S.l.*], 2020. Disponível em: https://powerbi.microsoft.com/pt-br/. Acesso em: 24 abr. 2020.

MUSEN, M.A. *The Protégé project:* A look back and a look forward. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, jun. 2015. DOI: 10.1145/2557001.25757003. Disponível em: https://protege.stanford.edu/about.php. Acesso em: 23 abr. 2020.

PAN, J. Z. *et al.* (ed.). *Exploiting linked data and knowledge graphs in large organisations*. Switzerland: Springer, 2017.

PARMENTER, D. *Key Performance Indicators:* developing, implementing, and using winning KPIs. 3rd ed. New Jersey: Wiley, 2015.

RASHID, S. M. *et al.* The semantic data dictionary: an approach for describing and annotating data. *Data Intelligence*, v. 2, n. 4, p. 443-486, 2020. Disponível em: https://doi.org/10.1162/dint_a_00058. Acesso em: mar. 2021.

RASHID, S. M. *et al.* The semantic data dictionary approach to data Annotation and integration. *SemSci@ ISWC,* Vienna, Austria, p. 47-54, 2017. Disponível em: https://dblp.org/db/conf/semweb/semsci2017.html. Acesso em: mar. 2021.

SEMANTIC DATA DICTIONARY. *SDD Specification*. [*S.l*], 2019. Disponível em: https://github.com/tetherless-world/SemanticDataDictionary. Acesso em: 22 set. 2019.

SEMANTICSCIENCE INTEGRATED ONTOLOGY. *SIO.* [*S.l*], 2020. Disponível em: https://bioportal. bioontology.org/ontologies/SIO. Acesso em: 16 set. 2020.

SILVA, V. S.; HANDSCHUH, S.; FREITAS, A. Categorization of semantic roles for dictionary definitions. *ArXiv preprint*, arXiv:1806.07711, 2018. Disponível em: https://arxiv.org/abs/1806.07711. Acesso em: mar. 2021.

VAUDANO, E. The innovative medicines initiative: a public private partnership model to foster drug discovery. *Computational and structural Biotechnology journal*, [*s.l.*], v. 6, n. 7, p. e201303017, 2013. Disponível em: https://doi.org/10.5936/csbj.201303017. Acesso em: mar. 2021.

W3C. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation, Feb. 2014. Disponível em: https://www.w3.org/TR/rdf11-concepts/ . Acesso em 16 de set de 2020.

WETZSTEIN, B.; MA, Z.; LEYMANN, F. Towards measuring key performance indicators of semantic business processes. *In:* ABRAMOWICZ, W.; FENSEL, D. (ed.). *Business Information Systems*. Berlin: Springer, 2008. p. 227-238.

WILKINSON, M. D *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, [*s.l.*], v. 3, n. 160018, 2016. DOI: 10.1038/sdata.2016.18. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: mar. 2021.

WISE, J. *et al.* Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug discovery today*, [*s.l.*], v. 24, n. 4, p. 933-938, Apr. 2019. Disponível em: https://doi.org/10.1016/j.drudis.2019.01.008. Acesso em: mar. 2021.

WISE, J. *et al.* The positive impacts of real-world data on the challenges facing the evolution of biopharma. *Drug discovery today*, [*s.l.*], v. 23, n. 4, p. 788-801, Apr. 2018. Disponível em: https://doi.org/10.1016/j.drudis.2018.01.034. Acesso em: mar. 2021.