DADOS ABERTOS GOVERNAMENTAIS: uma metodologia para publicar dados semânticos no estado de Minas Gerais

OPEN GOVERNMENT DATA: a methodology to publish semantic data in the state of Minas Gerais

Marcela Pires Estevanovic
Universidade Federal de Minas Gerais

Marcello Peixoto Bax
Universidade Federal de Minas Gerais

RESUMO

O avanço de tecnologias semânticas permite aplicar técnicas para anotar dados a fim de explicitar o conhecimento representado por eles. Para recuperar conhecimento, é possível enriquecê-los semanticamente para consumo humano e para processamento computacional. Assim, para publicar dados de qualidade na Internet é recomendável associá-los a um modelo conceitual como as ontologias que são representações do conhecimento compartilhado de uma área específica. O artigo apresenta uma metodologia para enriquecer e publicar dados abertos governamentais aplicada a um estudo de caso envolvendo Emendas Parlamentares Impositivas no estado de Minas Gerais. Apresenta-se um conjunto de dados anotados semanticamente e como estes são utilizados para responder consultas relacionadas ao domínio.

Palavras-Chave: Dados Abertos Governamentais. Enriquecimento Semântico. Ontologias.

ABSTRACT

The advancement of semantic technologies makes it possible to apply techniques to annotate data to make explicit the knowledge represented by them. To retrieve knowledge, it is possible to semantically enrich them for human consumption and for computational processing. Thus, to publish quality data on the Internet it is recommended to associate them with a conceptual model such as ontologies that are representations of the shared knowledge of a specific area. The article presents a methodology to enrich and publish open governmental data applied to a case study involving Impositive Parliamentary Amendments in the state of Minas Gerais. It presents a semantically annotated dataset and how it is used to answer queries related to the domain.

Keywords: Open Government Data. Semantic Enrichment. Ontologies.

1 INTRODUÇÃO

No contexto em que governos informatizam cada vez mais seus processos, surge a necessidade de gerir e analisar a grande quantidade de dados que são armazenados diariamente. Em consonância com esse movimento inicia-se outro que busca garantir a abertura dos dados, os Dados Abertos (ou *Open Data*). Os Dados Abertos Governamentais (DAG) focam em temas de interesse público e em conjuntos de informações controladas pela Administração Pública. Essa iniciativa ganha força quando é instituída a Parceria para Governo Aberto (*Open Government Patership)* que dispõe algumas diretrizes, difunde e incentiva práticas para aprimorar a transparência, acesso à informação e participação social. Em 2011, o Brasil[1], junto de sete países, endossou a criação dessa organização e é um de seus atuais membros (CGU, 2020).

Dados são qualificados como abertos quando qualquer pessoa pode acessar, utilizar, modificar, compartilhar (para qualquer finalidade), respeitando exigências básicas de referência (OPEN KNOWLEDGE FOUNDATION, 2021). Isso beneficia a transparência e permite que as ações de Governo possam ser acompanhadas, fiscalizadas e monitoradas. É comum que esses dados sejam publicados e disponibilizados para a população em portais na Internet. Como exemplos no Brasil, em nível federal existe o Portal de Dados Abertos[2]; em Minas Gerais há o seu análogo[3] e o Portal da Transparência[4].

Os meios de acesso aos dados são cruciais quando se leva em consideração a Lei de Acesso à Informação. Dentre diversas recomendações, essa lei obriga a publicação de informações produzidas ou de responsabilidade das entidades públicas em sítios eletrônicos. Além disso, recomenda disponibilizar os arquivos em formatos não proprietários (como . *.csv*, *.json* e *.xml*) para facilitar a leitura das informações e permitir o acesso externo (BRASIL, 2011). Idealmente, de acordo com Eaves (2009), as bases disponibilizadas devem ser disponibilizadas em *.csv* (ou similar) e sem licenças ou patentes.

Entretanto, um desafio para a abertura desses dados é garantir a compreensão das informações representadas por eles. Uma forma de publicar conjuntos de dados é utilizar um arquivo de dados e um dicionário de dados. A descrição do conhecimento nesses dicionários facilita a compreensão do conteúdo, mas não é considerada como suficiente em cenários que dados possam ser interligados (BAX; SILVA, 2020). Para isso, é possível utilizar uma metodologia baseada em ontologias para anotar dados com semântica, expressando-os formalmente para evitar interpretações diferentes do significado original. Uma ontologia, em poucas palavras, é uma formalização explícita de um conhecimento compartilhado (RECTOR et al., 2019).

Existem inúmeras técnicas de anotação semântica de dados (ROCHA, 2021) e uma delas, descrita em Rashid *et al.* (2017), foi utilizada na pesquisa relatada por este artigo. A anotação semântica facilita a homogeneização de dados entre bases e a recuperação da informação (BAX; SILVA, 2020; GONÇALVES, 2020) e garante que o significado daquele conjunto de colunas e linhas seja preservado ao anotar os dados com uma ontologia. Essa tecnologia formaliza os acordos e convenções (associações e relações) e estabelece um conjunto de regras para que aquela representação faça sentido para aquele domínio.

A questão examinada por esta pesquisa se localiza no contexto das Emendas Parlamentares Impositivas no estado de Minas Gerais. Por se tratar de um tema recente no contexto da Secretaria de Estado de Governo (SEGOV), constatou-se a necessidade de descrevê-lo em mais detalhes, com a finalidade de apresentar a informação de forma transparente e simplificada para os cidadãos. Além disso, existe o potencial de utilizar esse conhecimento para aprimorar a tomada de decisão dos atores envolvidos no processo de

[1] Disponível em: https://www.opengovpartnership.org/members/brazil/ . Acesso em: 26 abr. 2022.
[2] Disponível em: https://dados.gov.br/pagina/dados-abertos. Acesso em: 26 abr. 2022.
[3] Disponível em: https://dados.mg.gov.br/. Acesso em: 26 abr. 2022.
[4] Disponível em: https://www.transparencia.mg.gov.br/. Acesso em: 26 abr. 2022.

organizar e publicar dados. Nesse sentido é importante garantir que as análises desses conjuntos de dados sejam confiáveis e, para isso, é necessário utilizar um método que explicite e formalize a semântica daquele domínio.

Para isso, a pesquisa se concentrou em adaptar uma metodologia pré-existente para enriquecer e publicar dados abertos governamentais e aplicá-la, como prova de conceito, no estudo de caso das Emendas Parlamentares Impositivas. Dessa forma, o presente trabalho pode ser classificado como uma pesquisa de abordagem qualitativa, de objetivo explorativo. Após delineado o fluxo da metodologia, será realizado uma prova de conceito em um estudo de caso (SANTOS, 1999).

A organização deste texto se inicia com a Introdução, em seguida são apresentados dois tópicos teóricos – a Seção 2 "Representação de dados abertos governamentais" e a Seção 3 "Anotação e enriquecimento semântico". A Seção 4 apresenta os aspectos metodológicos do trabalho e na Seção 5 é descrita a aplicação da metodologia. Por fim, são apresentadas as considerações finais.

2 REPRESENTAÇÃO DE DADOS ABERTOS GOVERNAMENTAIS

O movimento de dados abertos tem por objetivo integrar e reutilizar dados, o que pode proporcionar análises mais profundas e abrangentes sobre diferentes temas. Para Eaves (2009), três pontos caracterizam esse tipo de dado: (1) Se o dado não pode ser encontrado na web, ele não existe; (2) Se ele não estiver aberto e disponível em formatos processáveis por máquina, não pode ser reaproveitado; (3) Se algum dispositivo legal impedir sua reutilização, ele não é útil. O setor público deve orientar-se por diretrizes que favoreçam a publicação e a disseminação de Dados Abertos Governamentais e que permita outros agentes sociais reutilizá-los, criando diversas aplicações. De forma complementar, oito princípios devem ser observados para que os dados sejam considerados abertos (OCDE, 2021; OPENGOVDATA.ORG, 2022; TAUBERER, 2014). Eles devem ser:

- **completos** (disponíveis sem limitações de privacidade, segurança ou privilégios de acesso);
- **primários** (publicados como coletados na fonte, de forma mais granular possível, sem agregação ou modificação)**;**
- **atuais** (atualizados recorrentemente para que seu valor não seja perdido)**;**
- **acessíveis** (acessados pelo maior número de usuários e processados de diferentes formas e softwares);
- **processáveis por máquina** (formalizados e estruturados para permitir o processamento automatizado);
- **não discriminatórios** (disponíveis a todos, sem exigência de pedido formalizado ou cadastro)**;**
- **não proprietários** (não devem estar em formatos de domínio exclusivo de empresas)**;**
- **licenças livres** (não devem estar sujeitos a direitos autorais, patentes, propriedade intelectual ou segredo industrial).

Além desses princípios, existe outra forma de caracterizar a qualidade dos dados publicados utilizando a classificação sugerida por Tim Berners Lee (2006). O chamado modelo de "Cinco Estrelas" propõe tornar os dados recuperáveis e exploráveis para gerar novos conhecimentos inferidos de conexões. Quando as informações se tornam *links* na Internet, é possível conectar diferentes temas (ou domínios) e automatizar inferências. O autor enumera quatro regras básicas:

1. utilizar o Identificador Uniforme de Recursos (URI) para nomear objetos na Internet;
2. utilizar o protocolo web para permitir a busca por esses objetos;

3. disponibilizar informações (metadados) que possam prover propriedades importantes em linguagens de representação padrão web;
4. incluir outros *links* já existentes na Web para que mais informações sejam cruzadas.

Considerando essas quatro regras, é possível então classificar a qualidade dos dados abertos em quantidade de estrelas (*cf.* Figura 1 e Quadro 1).

Figura 1 – Modelo Cinco Estrelas dos dados abertos conectados.

Fonte: 5 Star Data (2021).

Quadro 1 - Estrelas do padrão de dados abertos conectados e suas definições.

Quantidade de Estrelas	Definição
★	Disponível na Internet, em qualquer formato e com licença aberta.
★★	Disponível na Internet, em formato estruturado (como arquivo com extensão .xls).
★★★	Disponível na Internet, em formato estruturado e não proprietário (arquivos com extensão .csv).
★★★★	Disponível na Internet, em formato estruturado e não proprietário, dentro dos padrões estabelecidos pela W3C (RDF e SPARQL[5]) usando URLs para identificar entidades e propriedades.
★★★★★	Todas as regras anteriores com o adicional de conectar seus dados a outros, de forma a fornecer contexto.

Fonte: adaptado de 5 Star Data (2021).

Uma tecnologia citada no padrão de é o *Resource Descripition Framework* (RDF) (W3C, 2014). Considerado um dos elementos chaves da Web Semântica, o padrão permite descrever informação como recursos e suas propriedades organizados em triplas criando sentenças compostas de [SUJEITO] – [PREDICADO] – [OBJETO]. Esses elementos são recursos que podem possuir um identificador único (URI) ou um literal. Essas triplas fazem declarações sobre os recursos. Depois de gerar as triplas, essas podem ser persistidas em arquivos nos formatos RDF/XML, *Turtle* ou *N-Triples*. O formato *Trutle* é considerado o mais compreensível para leitura humana. A título de exemplo, a Figura 2, traz a representação de triplas que declaram o *Domain* (*rdfs:domain:Indicacao*) e o *Range* (*rdfs:range:Municipio*) da propriedade (*owl:ObjectProperty*) *temMunicípio* – que associa uma indicação a um município. A asserção na Figura 2 formaliza para o computador as restrições descritas (no caso, o domínio e o alcance da propriedade).

Figura 2 - Representação de triplas em formato *Turtle*.

```
@prefix : <http://www.semanticweb.org/user/ontologies/2021/10/freya> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
### http://www.semanticweb.org/user/ontologies/2021/10/freya#temMunicipio
:temMunicipio rdf:type owl:ObjectProperty ;
        rdfs:domain :Indicacao ;
        rdfs:range :Municipio .
```

Fonte: elaborado pelos autores.

Outro padrão relevante é o **SPARQL**, uma forma de consultar os documentos gerados em RDF. Essa linguagem pode gerar consultas com diferentes níveis de complexidade e pode abarcar uma grande quantidade de informações. As consultas podem ser realizadas em formulários web, chamados *endpoints*[5]. Dessa forma, ao disponibilizar conjuntos de dados é crucial anotar os seus significados, seus metadados (RASHID et al., 2017). A próxima seção apresenta a técnica de anotação utilizada neste estudo.

3. ANOTAÇÃO E ENRIQUECIMENTO SEMÂNTICO DE DADOS ABERTOS GOVERNAMENTAIS

Anotar algo implica, nesse contexto, a relacionar e estabelecer coerência entre os elementos de um domínio de conhecimento. No contexto de dados, implica descrever os metadados de uma base de dados de forma a esclarecer para o usuário o escopo e o significado de cada coluna que compõe o conjunto. De acordo com Belloze *et. al.*(2012), a anotação semântica é uma associação entre termos e expressões de um domínio descrito por uma ontologia. As ontologias provêm uma forma de explicitar e formalizar uma conceitualização compartilhada. A conceitualização é um modelo abstrato que descreve os conceitos e suas relações relevantes em um domínio de conhecimento. É explícito porque os conceitos mais relevantes devem ser definidos e descritos, do contrário o modelo estará incompleto. É formalizado para que aquela descrição possa ser legível e processável por máquinas - ou seja, o computador interpreta a semântica daquelas relações e representações do conhecimento. Por fim, é compartilhado pois aquele conhecimento deve ser um consenso entre pessoas, enfatizando a necessidade da participação de um grupo de pessoas para definir uma ontologia (RECTOR et al., 2019).

As ontologias são baseadas em Lógica de Descrição[6], o que permite formalizar uma série de restrições existentes no domínio representado e inferir novos conhecimentos (SANTOS et al., 2017). Formalizar o conhecimento é essencial para homogeneizar a informação uma vez que cada indivíduo pode interpretar a linguagem natural de forma diferente e isso pode interferir na análise dos dados. Portanto, quando uma base de dados é anotada formalmente, é possível que incorram menos erros de interpretação.

O Dicionário Semântico de Dados (*Semantic Data Dictionary* ou SDD) proposto por Rashid *et al.* (2017) é utilizado para anotar semanticamente dados. Trata-se de uma abordagem que emprega metadados pré-definidos fundamentados em ontologias. Para enriquecer os dados, o SDD os associa a conceitos da ontologia. Uma das recomendações para que a técnica seja aplicada com sucesso é a participação de diferentes atores no processo

[5] Um exemplo é o formulário de consulta ao Orçamento Federal Brasileiro, que conta com um tutorial sobre como realizar as pesquisas disponível em: http://orcamento.dados.gov.br/siopdoc/doku.php/acesso_publico:scripts_sparql (Acesso em 20 dez 2021)

[6] A teoria da Lógica de Descrição é baseada na definição de conceitos e relações. Dessa forma, é possível representar o conhecimento utilizando preceitos lógicos e formalizá-lo. Além disso, essa capacidade de descrever relações pode ser traduzida para a linguagem computacional (ALEXOPOULOS, 2020; OBITKO, 2007).

de anotação que deve ser orientada por especialistas de domínio e engenheiros do conhecimento. Isso se deve ao fato de que é necessário compreender profundamente o domínio a ser representado pelo modelo conceitual. Depois de selecionar o conjunto de dados a anotar, deve-se obter ou criar os seguintes artefatos:

- Ontologia de domínio: que formaliza os conceitos mais relevantes do problema de pesquisa. Deve-se reutilizar ontologias já existentes e consolidadas.
- *Dictionary Mapping* (DM): anota as relações semânticas do conjunto de dados em que cada coluna do conjunto de dados se torna uma linha do DM. Esse processo conceitualiza e formaliza as relações entre as classes do dataset, bem como a procedência do dado.
- *Codebook* (CB): estrutura os dados de colunas que possuem categorias agrupáveis e que são mapeados em conceitos correspondentes na ontologia.
- *Infosheet*: descreve os metadados do SDD como autor, contribuidores, data de criação, descrição do projeto, comentários, entre outras.

Assim, a representação proposta pelo SDD é processável por máquina, e uma vez que a anotação é separada da implementação da transformação (*script*) ela pode ser mais facilmente compreendida pelo especialista de domínio. Pode-se utilizar dicionários semânticos para harmonizar interpretações de conceitos em domínios correlatos. A Seção 5 apresenta os artefatos (ou *templates*) do SDD preenchidos.

Após o processo manual de anotação e preenchimento dos arquivos, é utilizado um *script* que interpreta o SDD e o converte em um grafo de conhecimento no padrão RDF que poderá ser consultado. Isso permite a formalização do vocabulário e possibilita a interoperabilidade e conexão a outros domínios (MOREIRA, 2021; TETHERLESS WORLD, 2021). Como exemplificado em Bax e Silva (2020) a aplicação do *script sdd2rdf* sobre os artefatos descritos acima, juntamente com o conjunto de dados a ser anotado, gera o arquivo RDF contendo o Grafo de Conhecimento.

Os grafos de conhecimento (*Knowledge Graphs* ou KG) são estruturas que permitem recuperar e acessar definições de domínio do conhecimento específicos. A semântica presente nesses grafos eleva a estrutura e seus recursos a um patamar em que é possível fazer análise de contextos complexos e até mesmo sugerir informações aos usuários de uma aplicação baseada nessa tecnologia (SANTOS et al., 2017).

De forma generalista, para definir um KG, de acordo com Paulheim (2017), é importante considerar um conjunto de quatro características: (1) descrever entidades e suas relações preservando os conceitos existentes no mundo utilizando um grafo (uma anotação formal e estruturada); (2) definir as classes e as relações entre entidades em um esquema (ou *schema* que é um conjunto de tipologias que podem ser relacionadas com um conjunto de propriedades); (3) permitir a conexão entre entidades de outros domínios do conhecimento; e por fim (4) não limitar a cobertura do grafo de conhecimento a somente um tópico do conhecimento. Com o propósito de utilizar-se dessa tecnologia é possível criar fragmentos de grafos de conhecimento para descrever um objeto de interesse específico que pode ser utilizado em outras aplicações (BAX; SILVA, 2020).

O padrão RDF pode ser utilizado para representar esses grafos de conhecimento, uma vez que permite a descrição em triplas que considera a linguagem e semântica daquelas relações. Portanto, com a ontologia construída, é possível gerar um RDF (seja no formato de um SDD ou outro) e gerar um fragmento do grafo. No próximo tópico, serão apresentados os aspectos metodológicos desta pesquisa.

4. METODOLOGIA

A pesquisa é aplicada e de abordagem qualitativa, uma vez que trata de adaptar e aplicar metodologia que busca transformar dados em formato relacional em grafos de conhecimento. Estes últimos servem de base para elaborar painéis de acompanhamento em ferramentas de *Business Analytics*. Os artefatos construídos, assim como outras informações, estão disponíveis no GitHub[7].

A metodologia aplicada é uma adaptação dos trabalhos de Gonçalves (2020) e de Bax e Silva (2020) que busca sistematizar uma forma de enriquecer dados com ontologias e publicar essas informações em painéis de acompanhamento. Baseada nos preceitos de Métodos Ágeis, a metodologia é cíclica e incremental. Isso favorece a construção de um modelo mais completo a cada ciclo. Um de seus fundamentos é a *Design Science Research* (DSR) que investiga a geração de conhecimento na construção de artefatos. Ao conciliar problemas práticos e teóricos, a DSR orienta a construção de conhecimento baseado em preceitos científicos para aprimorar práticas em áreas técnicas. Além disso, ao buscar respostas para problemas teóricos em casos reais, também lastreia soluções para situações práticas nas teorias científicas (BARBOSA; BAX, 2017; BAX, 2013; GONÇALVES, 2020). Esse processo iterativo, também chamado processo regulador, é descrito por Wieringa (2009).

Cada uma das etapas do ciclo (Figura 3) representa a análise de um problema prático à luz de correntes teóricas e a construção de um artefato que possa auxiliar a responder um problema teórico. Como descrito pelo autor, é crucial validar a produção em fases intermediárias do processo. Além disso, envolver pessoas que vivenciam o problema é essencial para validar os artefatos e gerar novas questões ou necessidades que possam ser respondidas pelo modelo a cada novo ciclo.

Figura 3 - Ciclo regulador de Wieringa (2009).

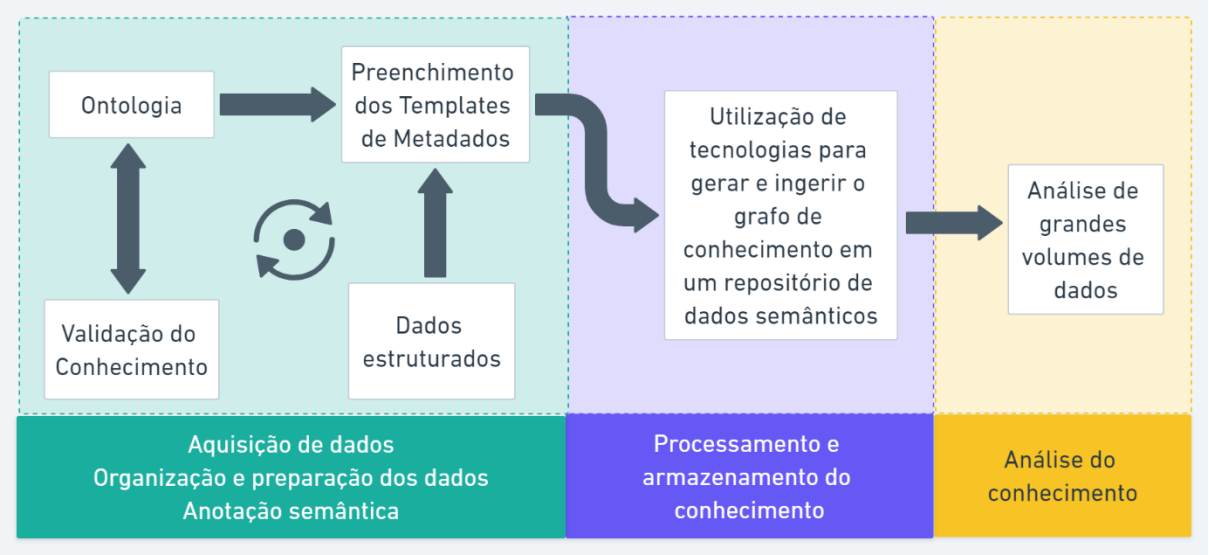
Fonte: Adaptado de Wieringa (2009).

Baseado nesse ciclo e no trabalho de Gonçalves (2020), propõe-se um outro formato simplificado para a metodologia (Figura 4). Como ilustra a Figura 4, cada uma das raias possui caixas que indicam os procedimentos para executar cada uma das fases. A primeira raia, "Aquisição de dados, organização e preparação dos dados e anotação semântica" aborda o tratamento da base de dados escolhida para iniciar o processo e como é realizada a anotação semântica, incluindo o desenvolvimento da ontologia. A segunda fase, "Processamento e armazenamento do conhecimento" representa as ferramentas que são

[7] Acesso em: https://github.com/marci-pires/EmendasParlamentaresMG

utilizadas para gerar e ingerir o grafo de conhecimento em um repositório de dados específico. Já a terceira "Análise do conhecimento" aborda a forma em que aquele conjunto de dados pode ser analisado utilizando uma ferramenta de *Business Analytics (*BI).

Figura 4 – Proposta de metodologia.



Fonte: Elaborado pelos autores, adaptada de Gonçalves (2020).

Nesta pesquisa a metodologia, sintetizada na Figura 4, é aplicada em um estudo de caso realizado sobre as Emendas Parlamentares Impositivas do estado de Minas Gerais, no ano de 2020. O propósito não é realizar uma análise das decisões legislativas e nem da efetividade da aplicação dessas emendas orçamentárias pelo Legislativo mineiro, mas propor uma forma de publicar esses dados seguindo os padrões internacionais de dados abertos governamentais. A aplicação da metodologia é descrita na Seção 5, onde discute-se cada tópico da pesquisa.

Para a primeira fase foram realizados grupos focais com atores chave da Secretaria de Estado de Governo do Estado de Minas Gerais (SEGOV), que tem como competência articular as necessidades do Governo e otimizar o processo de execução das emendas parlamentares (SEGOV, 2019), formaram o grupo de especialistas de domínio. De acordo com Gondim (2002), existem algumas modalidades para os grupos focais com diferentes objetivos: exploratórios, clínicos e vivenciais. Optou-se por utilizar a primeira modalidade. Os grupos exploratórios têm como foco a produção de conhecimento, a busca de novas ideias e necessidades de um determinado grupo para um contexto específico (*ibdem).*

Além disso, os grupos focais auxiliam na construção de instrumentos práticos e teóricos e, quando combinados com outras técnicas de pesquisa, podem trazer resultados positivos. A interação entre atores em torno da discussão do tema, pode propiciar um debate "aberto e acessível" que auxilia o processo de construção conjunta o conhecimento (TRAD, 2009, p. 792).

Na segunda fase "processamento e armazenamento do conhecimento" são utilizados *scripts* em linguagens para armazenamento de dados que persistem as informações em um *triplestore* para consulta. No caso, foi utilizado o Virtuoso[8] conectado ao Microsoft Power BI[9] via *Open Database Connectivity* (ODBC[10]), que cria uma ponte entre os dois softwares e permite a extração dos dados para o software de *Business Analytics*.

[8]Disponível em: https://virtuoso.openlinksw.com/ . Acesso em: 26 abr. 2022.
[9]Disponível em: https://powerbi.microsoft.com/pt-br/getting-started-with-power-bi/. Acesso em: 26 abr. 2022.
[10]Disponível em: https://www.openlinksw.com/odbc/. Acesso em: 26 abr. 2022.

A fase de análise do conhecimento envolve a elaboração de painéis de acompanhamento de indicadores criados a partir do trabalho de anotação semântica. Esse processo traz oportunidades de verificar a confiabilidade dos dados e pode ser criado por atores envolvidos diretamente com o domínio. Além disso, a formação desses painéis é uma forma de validar o conhecimento construído ao longo da aplicação da metodologia. A próxima seção apresenta a aplicação da metodologia.

5. ESTUDO DE CASO: EMENDAS PARLAMENTARES IMPOSITIVAS DO ESTADO DE MINAS GERAIS

Para iniciar a aplicação da metodologia deve-se definir qual será o uso da ontologia a ser construída e quais são os requisitos elementares que especificam o escopo do domínio do conhecimento (SEQUEDA; LASSILA, 2021). Essa fase de pré-modelagem conceitual é útil para evitar representar variáveis que não serão utilizadas, auxiliando na redução da complexidade do modelo, que sempre poderá ser revisitado ao longo dos ciclos futuros de desenvolvimento. Dessa forma, cada ciclo deve ser restrito a um conjunto de sentenças para que o conhecimento possa ser validado incrementalmente.

A primeira fase "**aquisição, organização e preparação dos dados e anotação semântica"** se inicia com a seleção da tabela (ou conjunto de tabelas) a ser analisada. Neste caso a tabela selecionada foi a de "Valores Indicados[11]" das emendas parlamentares impositivas do Estado de Minas Gerais, disponível no portal do sistema SIGCON-SAÍDA[12]. Foram selecionadas algumas colunas da tabela (Tabela 1) e os cabeçalhos e conteúdo foram normalizados.

Tabela 1 - Primeiras linhas do conjunto de dados a ser anotado.

ResponsavelNome	NumeroIndicacao	TipoIndicacao	DescricaoMunicipio	NomeConvenenteBeneficiado	NomeGrupoDespesa	ValorIndicado
AGOSTINHO PATRUS FILHO	52688	RESOLUCAO	SANTA MARIA DO SUACUI	HOSPITAL SANTA MARIA ETERNA	INVESTIMENTOS	400000
AGOSTINHO PATRUS FILHO	52703	RESOLUCAO	CAMPESTRE	FUNDO MUNICIPAL DE SAUDE DE CAMPESTRE	OUTRAS DESPESAS CORRENTES	100000
AGOSTINHO PATRUS FILHO	52750	RESOLUCAO	PEDRA AZUL	FUNDO MUNICIPAL DE SAUDE DE PEDRA AZUL	OUTRAS DESPESAS CORRENTES	100000

Fonte: Valores Indicados, "Planilha Relatório TCE MG - 2020 – Retificado", (SIGCON-SAÍDA, 2021).

A modelagem do domínio se iniciou com um levantamento bibliográfico e levou em consideração a experiência da primeira autora com o a área do governo em que atua com emendas parlamentares. Utilizou-se uma forma de organizar as sentenças em frases curtas que pudessem auxiliar a localização e definição das entidades, propriedades e restrições. Depois, disso, iniciou-se a modelagem visual do domínio para validação e, por fim, a ontologia.

A execução do grupo focal exploratório foi realizada de forma a validar as informações de cada uma das colunas da tabela escolhida e compreender como os conceitos se relacionavam. Isso foi adequado para guiar a construção da ontologia em sua primeira

[11] Na aba "Dados e Relatórios" da página https://sigconsaida.mg.gov.br/emendas-2020/ , é possível encontrar o arquivo "Relatório TCE MG - 2020 – Retificado" na integralidade (Acesso em 28 de abril 2022)

[12] O Sistema de Gestão de Convênios, Portarias e Contratos do Estado de Minas Gerais – Sigcon-MG – Módulo Saída tem como finalidade cadastrar, gerir e tramitar instrumentos jurídicos que possuem repasses de órgãos estaduais e entidades parceiras (OEEP) para convenentes, que podem ser Organizações da Sociedade Civil, Entes Federados ou pessoas jurídicas vinculadas, fundos municipais e Serviços Sociais Autônomos (SSA) (MINAS GERAIS, 2021).

versão e a produção dos *templates* do SDD. Foi de comum acordo iniciar o processo com a pergunta de competência "**Que responsável realizou o maior número de indicações em 2020?**" Cada encontro foi mediado pela primeira autora. No total foram realizados quatro encontros nos quais foram discutidos os conceitos presentes na tabela selecionada.

Foi possível correlacionar as necessidades discutidas no grupo focal e o cabeçalho da tabela, chegando a uma descrição: *Que responsável (ResponsavelNome) realizou mais indicações (NumeroIndicacao) em 2020?* A tripla formada aqui é a relação entre responsável e número da indicação, que representa um identificador único dos registros da tabela.

Para desenvolver a ontologia, denominada FREYA, foi necessário descrever os relacionamentos entre as classes encontradas, descrevendo suas propriedades e relações. Utilizou-se o sistema Protégé[13] que oferece suporte para criar, visualizar e modelar ontologias. Inicialmente, as colunas da tabela foram tratadas como classes.

Após elaborar a primeira versão da ontologia e validá-la com ajuda do grupo focal, passou-se ao processo de preenchimento dos *templates* de metadados do SDD. O primeiro passo é preencher a planilha de informações (*Infosheet*) com os metadados utilizados, assim como mostrado no Quadro 2 do Apêndice I. Como já visto, o *Dictionary Mapping* é um mapeamento de colunas do conjunto de dados para a ontologia. Isso envolve a explicitação de conceitos implícitos e a descrição dos seus atributos. No caso estudado, a Tabela 2 (Apêndice I) tem como primeira descrição o *NumeroIndicacao* que é o identificador único de cada indicação e que, pelo conhecimento das pessoas envolvidas é a Indicação em si. Por isso, é um conceito implícito naquele conjunto de dados, incluído como *??indicacao*. Então, é criada uma classe na ontologia, *freya:Indicacao*. Dessa forma, ocorre o enriquecimento semântico daqueles conceitos e relações antes implícitos de forma a gerar mais conhecimento.

O *Codebook*[14] (Tabela 3 – Apêndice I), descreve os dados categoriais do conjunto de dados e os mapeia para a ontologia FREYA. No caso, foi descrita a coluna de tipo de indicação e catalogados os códigos que aparecem descritos na tabela em classes da ontologia. Após preencher esses *templates*, iniciou-se a fase de processamento e armazenamento do conhecimento. Inicialmente, foi executado o script *sdd2rdf* que resulta em um fragmento de grafo de conhecimento persistido no Virtuoso em formato RDF.

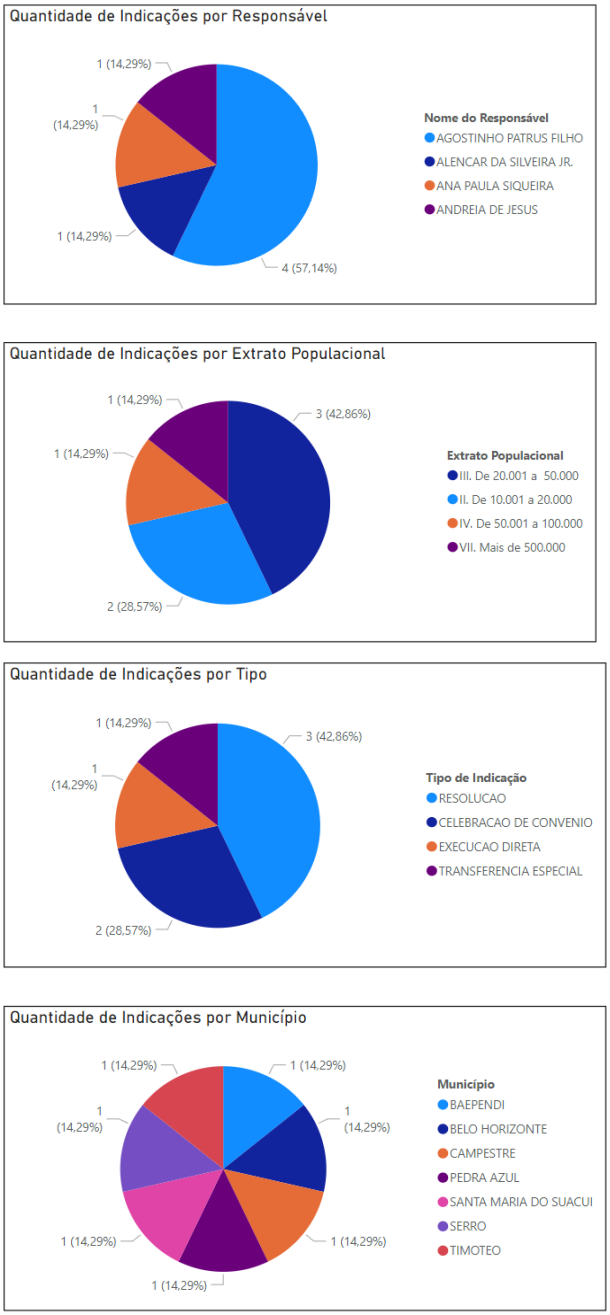
Então, é iniciada a fase de análise do conhecimento. Para que os dados sejam lidos no *Power BI*, é necessário configurar o ODBC de forma a conectar os dois sistemas. Após a carga dos dados é necessário ajustar as tabelas para criar os painéis. A Figura 5 mostra gráficos resultantes da junção de tabelas e informações inseridas diretamente no Power BI. Não é foco deste trabalho explorar as formas de transformação desses dados, mas mostrar uma das possibilidades de análise do conhecimento.

Espera-se que essa metodologia possa inspirar o desenvolvimento de um processo mais robusto e sistemático de preparação de dados e resultar em uma forma confiável de se basear as decisões em dados. Além disso espera-se que tal processo contribua para aumentar a transparência das informações disponibilizadas aos cidadãos.

Finalmente, reitera-se que este trabalho não busca julgar as decisões dos parlamentares nas alocações de seus recursos, mas objetiva apresentar uma forma de anotar e publicar dados semânticos.

[13] Disponível em: https://protege.stanford.edu/. Acesso em 30 de maio 2022.
[14] Em tradução literal "livro de códigos".

Figura 5 – Painel com as informações de uma indicação, de um Parlamentar, em 2020.



Fonte: elaborado pelos autores.

6. CONSIDERAÇÕES FINAIS

Este trabalho adaptou e aplicou em um estudo de caso uma metodologia para enriquecer e publicar dados abertos governamentais. A modelagem conceitual de dados envolveu diferentes atores que contribuíram para o modelo com diferentes percepções e interpretações. Isso auxiliou a criação de um conhecimento compartilhado e que pôde ser formalizado pela primeira versão simplificada de uma ontologia. Mesmo que o domínio pareça por demais complexo, ao iniciar de forma modular com poucas definições e adotando uma metodologia incremental e cíclica, é possível descrever objetos de estudo mais complicados.

Espera-se que este trabalho possa contribuir para anotar e formalizar o conhecimento sobre outros domínios específicos. Assim como demonstrado, isso pode trazer benefícios para a administração pública como forma de melhorar as políticas públicas e avançar com transparência. Além disso, existe um grande potencial para utilizar a tecnologia dos grafos para integrar dados no governo.

A metodologia apresentada visa organizar em etapas a anotação (preparar e organizar) e publicação de dados semânticos representados em RDF e fundamentados por uma ontologia. A utilização do SDD e de grafos de conhecimento permite, potencialmente, conectar diferentes fontes de dados. No caso estudado, foi possível condensar o conhecimento do grupo focal em uma ontologia, usada para anotar os dados e gerar um grafo de conhecimento que permitiu a construção de paneis de acompanhamento. Além disso, a abordagem descrita pode contribuir para integrar bases de dados heterogêneas e promover um alinhamento dos conceitos entre áreas correlatas.

Uma premissa do trabalho que pôde ser validada é que adotar os padrões descritos no contexto governamental tem potencial para aumentar a transparência e o reuso das informações. Por exemplo, o Terceiro Setor pode utilizar esses dados para construir plataformas que deem transparência às atividades de governo, beneficiando a população, já que o acesso à informação permite influência ativa dos cidadãos nas decisões de governo (SILVA, 2018).

Os artefatos descritos no trabalho, organizados pela metodologia apresentada, contribuíram para organizar o conhecimento do domínio e para a comunidade de dados abertos do estado de Minas Gerais. Como esforço futuro, é possível acionar agentes políticos envolvidos com a distribuição das Emendas Parlamentares para compreender a visão e as dificuldades que eles encontram em relação aos conceitos fundamentais do domínio e, assim, enriquecer o modelo. Uma outra possibilidade é testar a metodologia em domínio correlato para que os dados desta pesquisa possam ser interligados e as informações enriquecidas com outras classes e propriedades.

REFERÊNCIAS

5STARDATA. **As 5 estrelas dos Dados Abertos**. 2021. Disponível em: https://5stardata.info/pt-BR/. Acesso em: 6 ago. 2021.

ALEXOPOULOS, Panos. **Semantic Modeling for Data**. 1. ed. Sebastopol, CA: O'Reilly Media, 2020. Disponível em: https://books.google.es/books?id=MWH4DwAAQBAJ&dq=%22Semantic+Modeling+for+Data%22&lr=&hl=ca&source=gbs_navlinks_s%0Ahttps://learning.oreilly.com/library/view/semantic-modeling-for/9781492054269/. Acesso em: 14 jan. 2022.

BARBOSA, Daniel Mendes; BAX, Marcello. A Design Science como metodologia para a criação de um modelo de Gestão da Informação para o contexto da avaliação de cursos de graduação. **Revista Ibero-Americana de Ciência da Informação**, Brasília, DF, v. 10, n. 1, p. 32–48, 2017. DOI: 10.26512/rici.v10.n1.2017.2471.

BAX, Marcello Peixoto. Design science: filosofia da pesquisa em ciência da informação e tecnologia. **Ciência da Informação**, Brasília, DF, v. 42, n. 2, p. 298–312, 2013.

BAX, Marcello Peixoto; SILVA, Evaldo de Oliveira Da. Uso de Dicionário Semântico de Dados na anotação de modelos de dados dimensionais para geração de indicadores de desempenho. **Ciência da Informação**, Brasília, DF, v. 49, n. 3, p. 128–141, 2020.

BELLOZE, Kelle T.; MONTEIRO, Daniel Igor S. B.; LIMA, Túlio F.; SILVA-JR, Floriano P.; CAVALCANTI, Maria Cláudia. An Evaluation of Annotation Tools for Biomedical Texts. *In*: (Andreia

Malucelli, Marcello Peixoto Bax, Org.)V SEMINAR ON ONTOLOGY RESEARCH IN BRAZIL 2012, Recife - PE. **Anais** [...]. Recife - PE p. 108–119. Disponível em: http://ontobras-most.net84.net/%0A.

BERNERS-LEE, Tim. **Linked Data**. 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 28 fev. 2022.

BRASIL. LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011 - Lei de Acesso à Informação, LAI. . 18 nov. 2011.

CGU. **O que é a iniciativa — Português (Brasil)**. 2020. Disponível em: https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/o-que-e-a-iniciativa. Acesso em: 22 nov. 2021.

EAVES, David. **The Three Laws of Open Government Data**. 2009. Disponível em: https://eaves.ca/2009/09/30/three-law-of-open-government-data/. Acesso em: 30 jan. 2022.

GONÇALVES, José Eugênio de Assis. **Método Ágil de Integração Semântica de Dados Científicos Baseado em Ontologias**. 2020. Tese de Doutorado - Universidade Federal de Minas Gerais, *[S. l.]*, 2020. Disponível em: http://hdl.handle.net/1843/34013.

GONDIM, Sônia Maria Guedes. Grupos focais como técnica de investigação qualitativa: desafios metodológicos. **Paidéia (Ribeirão Preto)**, *[S. l.]*, v. 12, n. 24, p. 149–161, 2002. DOI: 10.1590/s0103-863x2002000300004.

MINAS GERAIS. **Decreto 48138, de 17/02/2021**. 2021. Disponível em: https://www.almg.gov.br/consulte/legislacao/completa/completa.html?tipo=DEC&num=48138&comp=&ano=2021. Acesso em: 21 abr. 2022.

MOREIRA, Felipe Lélis. **Impacto do uso de dados abertos sobre a assimetria de influência do lobby no Congresso Nacional**. 2021. *[S. l.]*, 2021. Disponível em: http://hdl.handle.net/1843/39130. Acesso em: 27 jan. 2022.

OBITKO, Marek. **Description Logics** . 2007. Disponível em: https://www.obitko.com/tutorials/ontologies-semantic-web/description-logics.html. Acesso em: 21 abr. 2022.

OCDE. **Open Government Data**. 2021. Disponível em: https://www.oecd.org/gov/digital-government/open-government-data.htm. Acesso em: 27 jan. 2022.

OPEN KNOWLEDGE FOUNDATION. **Open Definition - Defining Open in Open Data, Open Content and Open Knowledge**. 2021. Disponível em: http://opendefinition.org/. Acesso em: 22 nov. 2021.

OPENGOVDATA.ORG. **The 8 Principles of Open Government Data**. 2022. Disponível em: https://opengovdata.org/. Acesso em: 28 fev. 2022.

PAULHEIM, Heiko. Knowledge graph refinement: A survey of approaches and evaluation methods. **Semantic Web**, Amsterdã, Holanda, v. 8, n. 3, p. 489–508, 2017. DOI: 10.3233/SW-160218. Disponível em: http://www.semantic-web-journal.net/system/files/swj1167.pdf. Acesso em: 1 mar. 2022.

RASHID, Sabbir M.; CHASTAIN, Katherine; STINGONE, Jeanette A.; MCGUINNESS, Deborah L.; MCCUSKER, James P. The semantic data dictionary approach to data annotation & integration. **CEUR Workshop Proceedings**, Vienna, Austria, v. 1931, n. 1, p. 47–54, 2017. Disponível em: http://ceur-ws.org/Vol-1931/paper-07.pdf.

RECTOR, Alan; SCHULZ, Stefan; RODRIGUES, Jean Marie; CHUTE, Christopher G.; SOLBRIG, Harold. **On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL**. **Journal of Biomedical Informatics: X**Academic Press Inc., , 2019. DOI: 10.1016/j.yjbinx.2019.100002. Acesso em: 31 maio. 2020.

ROCHA, Rafael. **Integração Semântica de Dados Tabulares em CSV: proposta de arcabouço comparativo de ferramentas**. 2021. Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: http://hdl.handle.net/1843/36618. Acesso em: 24 abr. 2022.

SANTOS, Henrique; DANTAS, Victor; FURTADO, Vasco; PINHEIRO, Paulo; MCGUINNESS, Deborah L. From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards. *In*: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Portoroz, SLOVENIA. v. 10250 LNCSp. 94–108. DOI: 10.1007/978-3-319-58451-5_7. Disponível em: http://link.springer.com/10.1007/978-3-319-58451-5_7. Acesso em: 21 maio. 2020.

SEGOV. **Decreto 47.792 de 18/12/2019**. 2019. Disponível em: https://www.almg.gov.br/consulte/legislacao/completa/completa.html?tipo=DEC&num=47792&comp=&ano=2019. Acesso em: 19 jun. 2021.

SEQUEDA, Juan; LASSILA, Ora. Designing and Building Enterprise Knowledge Graphs. **Synthesis Lectures on Data, Semantics, and Knowledge**, Etronic, v. 11, n. 1, p. 1–165, 2021. DOI: 10.2200/S01105ED1V01Y202105DSK020. Disponível em: https://www.morganclaypool.com/doi/10.2200/S01105ED1V01Y202105DSK020.

SIGCON-SAÍDA. **Emendas 2020 – SIGCON-Saída**. 2021. Disponível em: https://sigconsaida.mg.gov.br/emendas-2020/. Acesso em: 3 mar. 2022.

SILVA, Patrícia Nascimento. **Dados governamentais abertos: métricas e indicadores de reúso**. 2018. Belo Horizonte, Brasil, 2018.

TAUBERER, Joshua. **Open Government Data: The Book**. 2014. Disponível em: https://opengovdata.io/. Acesso em: 24 out. 2021.

TETHERLESS WORLD. **SemanticDataDictionary - GitHub**. 2021. Disponível em: https://github.com/tetherless-world/SemanticDataDictionary/. Acesso em: 1 mar. 2022.

TRAD, Leny A. Bomfim. Grupos focais: conceitos, procedimentos e reflexões baseadas em experiências com o uso da técnica em pesquisas de saúde. **Physis: Revista de Saúde Coletiva**, [online], v. 19, n. 3, p. 777–796, 2009. DOI: 10.1590/S0103-73312009000300013. Disponível em: http://www.scielo.br/j/physis/a/gGZ7wXtGXqDHNCHv7gm3srw/?lang=pt. Acesso em: 4 out. 2021.

WIERINGA, Roel. Design science as nested problem solving. *In*: PROCEEDINGS OF THE 4TH INTERNATIONAL CONFERENCE ON DESIGN SCIENCE RESEARCH IN INFORMATION SYSTEMS AND TECHNOLOGY - DESRIST '09 2009, New York, New York, USA. **Anais** [...]. New York, New York, USA: ACM Press, 2009. p. 1. DOI: 10.1145/1555619.1555630. Disponível em: http://portal.acm.org/citation.cfm?doid=1555619.1555630.

APÊNDICE I – ARTEFATOS DO SDD

Quadro 2 - Especificação da Infosheet.

Attribute	**Value**
Type	http://purl.org/dc/dcmitype/Dataset
Title	EmendasParlamentares
Alternative Title	EmendasParlamentares
Comment	Criando um grafo de conhecimento usando a técnica SDD
Description	Os dados foram anotados do dataset EmendasParlamentares para demonstrar a tecnica SDD
Date Created	18/11/2021
Creators	Marcela Pires
Contributors	Marcello Bax
Publisher	Marcela Pires
Date of Issue	12/10/2021
Identifier	freya
Keywords	Emendas Parlamentares; atividade legislativa
Language	PT-BR
Version	1.0
Source	EmendasParlamentares/config/Infosheet.csv
File Format	csv
Dictionary Mapping	EmendasParlamentares/input/DM/sdd_emendasparlament.csv
Codebook	EmendasParlamentares/input/CB/Codebook.csv
Code Mapping	EmendasParlamentares/config/code_mappings.csv

Fonte: elaborado pelos autores

Tabela 2 - *Dictionary Mapping* para o domínio de Emendas Parlamentares Impositivas.

Column	Attribute	attributeOf	Entity	Relation	inRelationTo
NumeroIndicacao	:originalId	??indicacao			
ResponsavelNome	freya:Indicacao	??responsavel			
TipoIndicacao	freya:Indicacao	??execucaodaindicacao			
DescricaoMunicipio	freya:Municipio	??municipiobeneficiario			
NomeConvenenteBeneficiado	freya:Beneficiario	??nomedobeneficiario			
valorIndicado	freya:Indicacao	??valorindicacao			
??indicacao			freya:Indicacao		
??responsavel			freya:NomeResponsavel		
??indicacao			freya:Indicacao	freya:indicacaoPossui	freya:NomeResponsavel
??execucaodaindicacao			freya:TipoIndicacao		
??indicacao			freya:Indicacao	freya:indicacaoPossuiTipo	freya:TipoIndicacao
??municipiobeneficiario			freya:Municipio		
??nomedobeneficiario			freya:Beneficiario	freya:beneficiarioSituadoEm	freya:Municipio
??indicacao			freya:Indicacao	freya:indicacaoBeneficia	freya:Beneficiario
??valorindicacao			freya:ValorIndicacao		
??indicacao			freya:Indicacao	freya:possuiValorIndicacao	freya:ValorIndicacao
??nomedobeneficiario			freya:Beneficiario	freya:benefiadorPor	freya:TipoIndicacao

Fonte: elaborado pelos autores.

Tabela 3 – *Codebook* para o domínio de Emendas Parlamentares Impositivas.

Column	Code	Class
TipoIndicacao	APLICACAO DIRETA DOACAO DE BENS	:AplicacaoDiretaDoacaoBens
TipoIndicacao	CELEBRACAO DE CONVENIO	:CelebracaoConvenio
TipoIndicacao	EXECUCAO DIRETA	:ExecucaoDireta
TipoIndicacao	EXECUCAO DIRETA CAIXA ESCOLAR	:ExecuçãoDiretaCaixaEscolar
TipoIndicacao	OUTROS INSTRUMENTOS	:OutrosInstrumentos
TipoIndicacao	RESOLUCAO	:Resolucao
TipoIndicacao	TRANSFERENCIA ESPECIAL	:TransferenciaEspecial

Fonte: elaborado pelos autores.