UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO

MARCELA PIRES ESTEVANOVIC

Dados abertos governamentais: adaptação e aplicação de uma metodologia para publicar dados enriquecidos semanticamente no estado de Minas Gerais

BELO HORIZONTE
2022

MARCELA PIRES ESTEVANOVIC

Dados abertos governamentais: adaptação e aplicação de uma metodologia para publicar dados enriquecidos semanticamente no estado de Minas Gerais

Dissertação apresentada ao Programa de Pós-graduação em Gestão e Organização do Conhecimento da Escola de Ciência da Informação (ECI), da Universidade Federal de Minas Gerais (UFMG), como requisito para a obtenção do certificado de Mestre Gestão e Organização do Conhecimento.

Área de Concentração: Ciência da Informação

Linha de Pesquisa: Gestão & Tecnologia da Informação e Comunicação (GETIC)

Orientador: Marcello Peixoto Bax

BELO HORIZONTE

2022

E79d Estevanovic, Marcela Pires.
Dados abertos governamentais [recurso eletrônico] : adaptação e aplicação de uma metodologia para publicar dados enriquecidos semanticamente no estado de Minas Gerais / Marcela Pires Estevanovic. - 2022.
1 recurso online (93 f. : il., color.) : pdf.

Orientador: Marcello Peixoto Bax.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 81-87.
Apêndice: f. 88-93.
Exigência do sistema: Adobe Acrobat Reader.

1. Ciência da informação – Teses. 2. Ontologia - Teses. 3. Web semântica - Teses. 4. Representação do conhecimento (sistemas especialistas) – Teses. I. Bax, Marcello Peixoto. II. Universidade Federal de Minas Gerais. Escola de Ciência da Informação. III. Título.

CDU: 025.4.03

Ficha catalográfica: Maianna Giselle de Paula – CRB/6- 2642

Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

FOLHA DE APROVAÇÃO

Dados abertos governamentais: adaptação e aplicação de uma metodologia para publicar dados enriquecidos semanticamente no estado de Minas Gerais

MARCELA PIRES ESTEVANOVIC

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Mestre em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Gestão e Tecnologia da Informação e Comunicação.

Aprovada em 01 de novembro de 2022, por videoconferência, pela banca constituída pelos membros:

Prof(a). Marcello Peixoto Bax (Orientador)
ECI/UFMG

Prof(a). Webert Junio Araujo
CEFET-MG

Prof(a). Lucinéia Souza Maia
UFOP

Prof(a). Mauro Araújo Câmara
Fundação João Pinheiro

Prof(a). Henrique Oliveira Santos
RPI

Belo Horizonte, 01 de novembro de 2022.

Documento assinado eletronicamente por **Marcello Peixoto Bax**, **Professor do Magistério Superior**, em 18/11/2022, às 13:26, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Webert Júnio Araújo**, **Usuário Externo**, em 12/12/2022, às 16:05, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mauro Araújo Câmara**, **Usuário Externo**, em 12/12/2022, às 18:58, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lucinéia Souza Maia**, **Usuário Externo**, em 13/12/2022, às 17:53, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Henrique Oliveira Santos**, **Usuário Externo**, em 24/01/2023, às 23:15, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1870627** e o código CRC **FC32E75B**.

Referência: Processo nº 23072.264757/2022-03 SEI nº 1870627

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO - ECI
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO - PPGGOC

ATA DA DEFESA DA DISSERTAÇÃO DA ALUNA

MARCELA PIRES ESTEVANOVIC

Realizou-se, no dia 01 de novembro de 2022, às 16:00 horas, por videoconferência, da Universidade Federal de Minas Gerais, a defesa de dissertação, intitulada *Dados abertos governamentais: adaptação e aplicação de uma metodologia para publicar dados enriquecidos semanticamente no estado de Minas Gerais*, apresentada por MARCELA PIRES ESTEVANOVIC, número de registro 2020660339, graduada no curso de ADMINISTRAÇÃO PÚBLICA, como requisito parcial para a obtenção do grau de Mestre em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Marcello Peixoto Bax - ECI/UFMG (Orientador), Prof(a). Webert Junio Araujo - CEFET-MG, Prof(a). Lucinéia Souza Maia - UFOP, Prof(a). Mauro Araújo Câmara - Fundação João Pinheiro, Prof(a). Henrique Oliveira Santos - RPI.

A Comissão considerou a dissertação:

(X) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 01 de novembro de 2022.

Assinatura dos membros da banca examinadora:

Documento assinado eletronicamente por **Marcello Peixoto Bax**, **Professor do Magistério Superior**, em 18/11/2022, às 13:26, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Webert Júnio Araújo**, **Usuário Externo**, em 12/12/2022, às 16:05, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mauro Araújo Câmara**, **Usuário Externo**, em 12/12/2022, às 18:59, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lucinéia Souza Maia**, **Usuário Externo**, em 13/12/2022, às 17:54, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Henrique Oliveira Santos**, **Usuário Externo**, em 24/01/2023, às 23:15, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1870590** e o código CRC **1B7588E0**.

Referência: Processo nº 23072.264757/2022-03 SEI nº 1870590

AGRADECIMENTOS

Aceitar o desafio de realizar um Mestrado após sair da faculdade trouxe uma variedade de sentimentos, alguns velhos conhecidos outros novos, para o meu dia a dia. A premeditada dificuldade se expandiu com a pandemia e, finalmente, foi vencida.

Chegar ao final desta jornada só foi possível com a ajuda do meu orientador Marcello Peixoto Bax, que, além de mostrar os melhores caminhos, apresentou diferentes técnicas e ferramentas. Serei eternamente grata por todo conhecimento repassado e confiança na conclusão do texto. Agradeço por toda ajuda e acolhimento nas minhas limitações e dificuldades.

Estendo meus agradecimentos aos colegas do PPGGOC-ECI e do grupo *e-Semantic Science* que contribuíram para o enriquecimento deste trabalho. Em especial, gostaria de agradecer ao Evaldo sempre disposto a ajudar e contribuir da forma mais positiva possível para o meu crescimento como profissional, além de me auxiliar a executar a parte técnica deste trabalho.

Aos meus pais, Ricardo e Lucimar, agradeço a todo o apoio emocional que vocês sempre ofereceram e pela compreensão nessa fase difícil, mas muito recompensadora. Ao amor da minha vida, Alexandre, agradeço a força e por garantir que eu me mantivesse motivada a terminar essa jornada. Gostaria também de expressar minha eterna gratidão aos meus familiares que, mesmo sem entender os tópicos estudados, nunca deixaram de apoiar meus sonhos.

Agradeço à Tânia e Barjute Bacha que me acompanham desde 2013 e sempre me inspiram a ser a melhor versão de mim mesma, a buscar os meus desejos e conciliar a disciplina com momentos leves e produtivos.

Aos meus amigos, meu coração é de vocês e vocês sabem. Chegar aqui sem o companheirismo de todos seria inviável. Dos mais antigos aos mais recentes vocês não fazem ideia da diferença que fazem na minha vida. Ao Tiago Franco, que neste momento está em terras distantes, obrigada pelo carinho e pela inspiração em seguir uma carreira acadêmica.

Aos parceiros de trabalho na Secretaria de Estado de Governo de Minas Gerais, agradeço imensamente a oportunidade que todos me concederam de executar um trabalho acadêmico em um contexto prático. A ajuda de todos foi indispensável para que eu pudesse executar os objetivos deste mestrado. Ao Thiago, obrigada por compreender as necessidades que uma dissertação traz e por nunca deixar de apoiar e acreditar no meu trabalho.

Alone, it's just a journey. Now adventures, they must be shared

RESUMO

Com o avanço de novas tecnologias na *Web* e com o aumento da quantidade de dados e informações produzidos, surge a necessidade de aplicar técnicas que auxiliem na preparação e organização de dados para publicação. No âmbito da Administração Pública, por exemplo, conjuntos de dados podem ser publicados na *Web* com o propósito de analisar o perfil epidemiológico de uma população, ou para informar o número de homicídios por região, entre outros. É importante que os dados estejam acessíveis e devidamente anotados com metadados a fim de ampliar o seu reuso por diferentes tipos de pessoas e organizações. No entanto, é comum encontrar publicados na *Web* conjuntos de dados em formatos heterogêneos ou sem documentação que os descrevam. Situações deste tipo prejudicam o consumo de informações por diferentes aplicações, indo de encontro à necessidade da transparência na administração pública, melhoria dos processos administrativos, ou até mesmo, no auxílio a decisões do poder público baseado em evidências. Nesse contexto, torna-se importante aplicar técnicas que permitam anotar dados a fim de tornar explícito o conhecimento representado por eles. Não obstante, é necessário pensar em como enriquecê-los semanticamente tanto para consumo humano, quanto para processamento por meio de computadores. Assim, ao publicar um conjunto de dados torna-se necessário também associá-lo ao modelo conceitual que o defina. O uso de ontologias permite representar o conhecimento em artefatos de *software* que organizam classes, propriedades, objetos e restrições de um domínio de conhecimento específico. Essa discussão encontra-se no âmbito do tema Dados Abertos Governamentais, em que é necessário preparar os conjuntos em formatos que possibilitem o processamento automatizado. Consequentemente, abre-se caminho para processar e extrair conhecimento para analisar um grande volume de dados. Neste trabalho, é proposta uma abordagem metodológica para publicar dados abertos governamentais. Para isso, foi realizado um estudo de caso aplicado para anotar dados de emendas parlamentares impositivas em Minas Gerais, utilizando como prova de conceito. Para construir a ontologia foram realizados grupos focais com servidores mineiros para validar a representação do conhecimento e preencher *templates* de metadados para elaborar os artefatos. Então, são ingeridos os dados para gerar um grafo de conhecimento que é acessado via repositório pelo programa de *Bussiness Intelligence*. Finalmente, apresenta-se os dados anotados semanticamente e é explorada uma forma de "navegar" pelos conceitos por meio de questões de competência. Portanto, este trabalho contribui com uma forma de anotar e enriquecer, continuamente, dados e publicá-los na *Web*. Isso aumenta a confiança e escalabilidade de modelos que se beneficiam das Tecnologias da *Web* Semântica.

Palavras-chave: Dados Abertos Governamentais, Ontologia, Tecnologias da *Web* Semântica.

ABSTRACT

With the advance of new technologies on the *Web* and the increase in the amount of data and information produced, there is a need to apply techniques that assist in the preparation and organization of data for publication. In the scope of Public Administration, for example, data sets can be published on the *Web* with the purpose of analyzing the epidemiological profile of a population, or to inform the number of homicides per region, among others. It is important that the data is accessible and properly annotated with metadata to broaden its reuse by different types of people and organizations. However, it is common to find data sets published on the *Web* in heterogeneous formats or without documentation describing them. Situations of this type hinder the consumption of information by different applications, meeting the need for transparency in public administration, improvement of administrative processes, or even, in helping the government to make decisions based on evidence. In this context, it is important to apply techniques that allow the annotation of data to make the knowledge represented by them explicit. Nevertheless, it is necessary to think about how to enrich them semantically both for human consumption and for processing by computers. Thus, when publishing a dataset, it is also necessary to associate it with the conceptual model that defines it. The use of ontologies allows knowledge to be represented in *software* artifacts that organize classes, properties, objects, and constraints of a specific knowledge domain. This discussion is within the scope of Open Government Data, where it is necessary to prepare the sets in formats that allow automated processing. Consequently, it opens the way to process and extract knowledge to analyze a large volume of data. In this paper, a methodological approach to publish open government data is proposed. For this, a case study was conducted to annotate data from impositive parliamentary amendments in Minas Gerais, using it as a proof of concept. To build the ontology, focus groups were conducted with Minas Gerais servers to validate the knowledge representation and fill in metadata *template*s to elaborate the artifacts. Then, data is ingested to generate a knowledge graph that is accessed via a repository by the Business Intelligence program. Finally, the semantically annotated data is presented and a way to "navigate" the concepts through competency questions is explored. Therefore, this work contributes a way to continuously annotate and enrich data and publish it on the *Web*. This increases the reliability and scalability of models that benefit from Semantic *Web* Technologies.

Palavras-chave: Open Government Data, Ontology, Semantic *Web* Tecnologies.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ALMG	Assembleia Legislativa de Minas Gerais
Cagec	Cadastro Geral de Convenentes do Estado de Minas Gerais
CB	*Codebook*
CDUSP	Código de Defesa do Usuário de Serviços Públicos
CKAN	*Comprehensive Knowledge Archive Network*
CSV	*Comma Separated Values*
DM	*Dictionary Mapping*
DSR	*Design Science Research*
HADatAc	*Human-Aware Data Acquisition framework*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
IRI	*Internationalized Resource Identifier*
JSON	*JavaScript Object Notation*
KG	*Knowledge Graph*
LAI	Lei de Acesso à Informação
LD	*Linked Data*
LDO	Lei de Diretrizes Orçamentárias
LGPD	Lei Geral de Proteção de Dados pessoais
LOA	Lei Orçamentária Anual
LOD	*Linked Open Data*
ODBC	*Open Database Connectivity*
ODIN	*Ontology-based Data Integration*
OECD	Organização para a Cooperação e Desenvolvimento Econômico
OGD	Dados Abertos Governamentais
OGP	*Open Government Partnership*
OKF	*Open Knowledge Foundation*
OSC	Organizações da Sociedade Civil
OWL	*Ontology Web Language*
PDCA	*Plan-Act-Check-Act*

pLOA	Projeto de Lei Orçamentária Anual
PPAG	Plano Plurianual de Ação Governamental
RDF	*Resource Description Framework*
RDF-S	*Resource Description Framework Schema*
RPI	Rensselaer Polytechnic Institute
SDD	*Semantic Data Dictionary*
SEGOV	Secretaria de Estado de Governo do Estado de Minas Gerais
SEPLAG	Secretaria de Estado de Planejamento e Gestão
SIGCON-SAÍDA	Sistema de Gestão de Convênios, Portarias e Contratos do Estado de Minas Gerais - Módulo Saída
SPARQL	*SPARQL Protocol and RDF Query Language*
SSA	Fundos Municipais ou Serviços Sociais Autônomos
TCE-MG	Tribunal de Contas do Estado de Minas Gerais
TWC	*Tetherless World Constellation*
UNPAN	*United Nations Public Administration Network*
URI	Identificador Uniforme de Recursos
URL	Localizador Uniforme de Recursos
W3C	*World Wide Web Consortium*
XLSX	Arquivo do Microsoft Excel
XML	*eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

No contexto da informatização, os dados são gerados diariamente em grandes quantidades, em diferentes sistemas de informação. Surge, então, a necessidade de gerir, analisar e utilizar o conhecimento dessas bases de dados. Portanto, para usufruir desses conjuntos de informações é preciso disponibilizar acesso a eles. Os chamados dados abertos (ou *Open Data*) podem ser acessados por qualquer pessoa e devem ser distribuídos em formatos reutilizáveis (OPEN KNOWLEDGE FOUNDATION, 2021).

No cenário governamental, existe uma demanda específica por disponibilizar dados e informações para a população. Essa abertura de dados pode contribuir para o acesso à informação, para a construção de políticas públicas, promoção de inovação e criação de novos serviços para os cidadãos (OCDE, 2021). Como forma de especializar o conceito inicial, os dados abertos governamentais (ou OGD[1]) beneficiam a transparência e permitem que as ações de governo sejam rastreadas, fiscalizadas e monitoradas.

A iniciativa de disponibilizar dados públicos em formato aberto ganha força quando é instituída a Parceria para Governo Aberto (*Open Government Partnership*[2]), em 2011, que dispõe algumas diretrizes e boas práticas para fomentar a transparência, participação social e o acesso à informação. O Brasil foi um dos fundadores da organização e atua, até o momento dessa pesquisa, como membro (CGU, 2020). Alguns frutos dessa participação podem ser exemplificados pelo portal federal de dados abertos[3], os portais de transparência dos estados de Minas Gerais[4] e Alagoas[5].

Esses resultados são também influenciados pela Lei de Acesso à Informação que determina aos órgãos e entidades públicas, de todas as esferas de governo, o dever de publicar as informações produzidas por eles, com clareza, em sítios eletrônicos. Além disso, esse dispositivo legal prevê requisitos mínimos para a publicação de conjuntos de dados, que são: permitir baixar os arquivos, publicar em formato não proprietário e permitir o acesso externo e automatizado àquelas informações (BRASIL, 2011c).

No contexto brasileiro, a administração pública precisa disponibilizar seus dados e garantir a segurança e confiabilidade das informações. Para que essas condições sejam

[1] Do inglês, *Open Government Data.*
[2] Disponível em: https://www.opengovpartnership.org/members/brazil/ . Acesso em 25 dez 2022.
[3] Disponível em: https://dados.gov.br/pagina/dados-abertos . Acesso em 25 abr 2022.
[4] Disponível em: https://www.transparencia.mg.gov.br/ . Acesso em 25 abr 2022.
[5] Disponível em: http://transparencia.al.gov.br/ . Acesso em 25 abr 2022.

contempladas, é recomendável padronizar o processo de abertura de dados e, assim, reduzir problemas de qualidade, recuperação e integração .

Padronizar os significados dos conceitos e torná-los coesos entre si é crucial, sobretudo, quando existe a necessidade de processar as informações em unidades de processamento computacional. É pertinente buscar uma solução tecnológica que possa reduzir problemas conceituais e que possa descrever os conjuntos de dados em um formato legível por máquinas. Esses aspectos podem ser mitigados por um processo de anotação e enriquecimento semântico baseado em ontologias.

Uma ontologia pode ser definida como uma formalização explícita de um conhecimento compartilhado (RECTOR et al., 2019). Dessa forma, permite modelar um esquema conceitual, que representa um domínio de conhecimento específico, através da descrição das relações entre os objetos e das propriedades daquele contexto (*ibdem*). Uma outra característica relevante dessa tecnologia é o compromisso em envolver diferentes agentes para elaborar sobre o conhecimento e acordar, de forma consensual, a respeito dos conceitos envolvidos (NOY; MCGUINNESS, 2011). As bases da representação em ontologias como solução tecnológica são: a *Ontology Web Language* (OWL) e o *Resource Description Framework* (RDF).

A OWL expressa as relações semânticas entre as classes, propriedades e objetos de uma ontologia, além de explicitar relações de hierarquia e definir restrições entre as associações (W3C, 2012). Já o RDF, de forma complementar, permite interagir e recuperar informações de um banco de dados a partir de um modelo de conceito definido, além de descrever os metadados daquela representação do conhecimento. O uso dessas linguagens está relacionado à potencial interpretação de computadores sobre conjunto de dados e modelos conceituais. Além disso, é possível que sejam geradas inferências a partir das definições explicitadas.

Ao publicar um conjunto de dados, é importante que este esteja acompanhado de informações a respeito do conteúdo daquele conjunto, um dicionário de dados. É comum que esses pacotes de informação sejam úteis para descrever o material, mas podem não ser suficientes para um cenário em que vários dados precisem ser interligados (BAX; SILVA, 2020). Dentre as técnicas descritas na literatura (BRITTO; RUY; AZEVEDO, 2020; ROCHA, 2021; SILVA, 2008) que tratam de forma generalista temas semelhantes à anotação e integração semântica de dados, a postulada em Rashid *et al* (2017) apresenta a técnica do Dicionário Semântico de Dados (*Semantic Data Dictionary* ou SDD). A anotação semântica pode ser definida como uma associação entre termos, conceitos e expressões de uma área específica por meio de uma ontologia (BELLOZE et al., 2012). O SDD permite organizar e manipular dados

de diferentes fontes e formatos. Além disso, esse formato permite unir a parte conceitual, definida pelo ferramental semântico, com os dados relacionais, formando os grafos de conhecimento (*Knowledge Graphs* ou KG), sendo estas estruturas que, quando presentes em aplicações tecnológicas, permitem acessar informações de diferentes domínios, recuperar conhecimento e inferir informações (SANTOS et al., 2017).

Com base nessas possibilidades, **como a aplicação de uma metodologia de anotação e enriquecimento semântico pode contribuir para o processo de publicação de dados abertos governamentais?**

Para isso, esse trabalho se dispõe a adaptar as metodologias descritas em Gonçalves (2020) e Bax e Silva (2021). Os dois textos apresentam formas análogas de utilizar a técnica do SDD para anotar dados e formalizar ontologias para descrever aquele domínio. O primeiro método, denominado ODIN é baseado em ciclos curtos de desenvolvimento de modelos conceituais, anotação semântica de dados e geração de grafos de conhecimento. De forma semelhante, o artigo de Bax e Silva (2021) apresenta uma metodologia para implementar a técnica do SDD em um caso específico. A solução proposta neste trabalho é estabelecer um fluxo, adaptado dos trabalhos supracitados, com passos que permitam organizar, preparar e anotar dados, preparar e armazenar o conhecimento e, por fim, analisar o conhecimento. Dessa forma, é possível alinhar diferentes técnicas e contribuir para o processo de abertura de dados de uma organização específica.

Para aplicar a metodologia, foi realizado um estudo de caso das Emendas Parlamentares Impositivas do estado de Minas Gerais, no ano de 2020, que será utilizado como prova de conceito. Foi selecionado um conjunto de dados específico, organizou-se a primeira versão de uma ontologia, estabeleceu-se um grupo focal para validar a representação do conhecimento e foram preenchidos os arquivos de metadados para elaborar o SDD. Na sequência, utilizou-se de tecnologias para gerar e ingerir um grafo de conhecimento para implementar um repositório que é acessado pelo programa de análise de dados.

Cabe ressaltar que o propósito desta pesquisa não é realizar uma análise sobre as decisões legislativas e nem sobre a efetividade da aplicação dessas emendas orçamentárias realizadas pelos membros do Legislativo mineiro, e sim, propor uma forma de realizar a publicação desses dados em um formato que esteja dentro dos padrões internacionais de dados abertos governamentais. Em suma, o objetivo deste trabalho é adaptar uma metodologia para enriquecer e publicar dados abertos governamentais e aplicá-la, como prova de conceito, no estudo de caso das Emendas Parlamentares Impositivas do estado de Minas Gerais do ano de 2020. Para os objetivos específicos lista-se:

- Adaptar as metodologias propostas nos trabalhos de Gonçalves (2020) e Bax e Silva (2020);
- Propor uma ontologia base sobre o domínio de emendas parlamentares impositivas no estado de Minas Gerais, levando em consideração o escopo da anotação;
- Aplicar a metodologia resultante em um estudo de caso;
- Utilizar técnicas e propor uma infraestrutura para persistir os dados anotados;
- Apresentar uma maneira de utilizar os artefatos resultantes da anotação semântica para elaborar um painel de acompanhamento básico.

Espera-se que este trabalho possa contribuir para a abertura de dados em um contexto de governo, o que beneficia a transparência, e aprimora a gestão pública, fortalecendo o processo democrático e a prestação de políticas públicas. Além disso, o processo de enriquecimento semântico, utilizando ontologias e a técnica SDD, permitem interligar dados de diferentes fontes, o que favorece a geração de conhecimento.

Este trabalho está dividido em seis capítulos. O primeiro é esta introdução, seguida do referencial teórico apresentado no Capítulo 2 (Dados abertos governamentais e a *Web* Semântica) e no Capítulo 3 (Anotação e enriquecimento semântico de dados). O Capítulo 4 apresenta a metodologia e o Capítulo 5 apresenta a aplicação da metodologia ao estudo de caso das Emendas Parlamentares Impositivas. Por fim, o Capítulo 6 traz a conclusão, dificuldades e limitações e indicação de trabalhos futuros.

2 DADOS ABERTOS GOVERNAMENTAIS E A *WEB* SEMÂNTICA

As tecnologias da informação podem facilitar a busca por dados, auxiliar o compartilhamento e a compreensão de conceitos de diferentes áreas, inclusive as relacionadas às políticas públicas e gestão governamental (TAUBERER, 2014). Para isso, é recomendável compreender os conceitos que envolvem dados abertos governamentais, a *Web* Semântica e a relação entre eles.

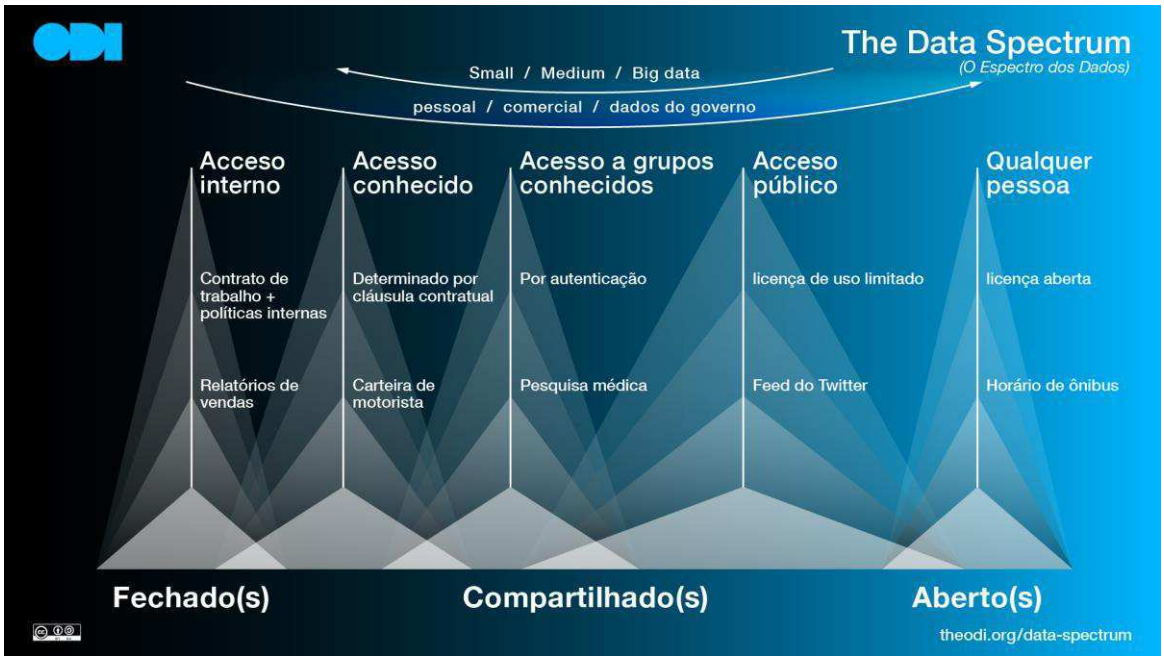
2.1 Dados Abertos Governamentais

No campo da Ciência da Informação, é possível discutir três conceitos principais a respeito de representações do mundo: dados, informação e conhecimento. O primeiro pode ser tratado como algo que simplesmente existe e não tem um significado além daquele número ou palavra. O segundo conceito amplia o anterior ao colocar esses dados em perspectiva a outros, acrescentando descrições a eles. O último requer entender como essas informações se conectam, com a intenção de torná-la útil. Esses conceitos são uma forma de descrever o processo da construção do conhecimento.

Segundo a *Open Knowledge Foundation* (OKF), dados abertos podem ser utilizados livremente, reutilizados e redistribuídos por qualquer pessoa – sujeitos a poucas regras e exigências de compartilhamento, mas sempre disponibilizando a fonte original. Portanto, essa definição carrega três pilares fundamentais: disponibilidade e acesso, reuso e distribuição e participação universal (OPEN KNOWLEDGE BRASIL, 2021). A Figura 1 demonstra e exemplifica o espectro de dados e os diferentes formatos de abertura. Quando se trata de dados compartilhados, é possível compreender que o acesso depende de uma validação ou identificação do usuário. Já os considerados abertos devem necessariamente estar disponíveis para qualquer pessoa e a licença deve ser aberta.

Um exemplo disso são os horários e rotas de ônibus, despesas públicas, gastos com cartões corporativos utilizados por membros do poder executivo entre outros assuntos de interesse público.

Figura 1 - O Espectro dos Dados.



Fonte:(OPEN DATA INSTITUTE, 2021).

A principal concepção desenvolvida foi a de que dados de governos são propriedades comuns e, portanto, devem ser acessíveis, legíveis por máquina e interoperáveis, para que a informação possa ser produzida por e para todos (ISOTANI; BITTENCOURT, 2015). Por definição, dados abertos governamentais (OGD) são dados públicos representados em meio digital; estruturados em formato aberto; processáveis por máquina; referenciados na *Web*; e disponibilizados sob licença aberta que permita sua livre reutilização, combinação ou consumo por aplicações digitais desenvolvidas pela sociedade.

De acordo com Eaves (2009), existem três leis principais que definem os OGD: (1) Se o dado não pode ser encontrado na *Web*, ele não existe; (2) Se 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. De forma complementar, são definidos oito princípios que devem ser seguidos para que aqueles dados possam ser considerados abertos (OCDE, 2021; OPENGOVDATA.ORG, 2022; TAUBERER, 2014). Eles devem ser:

- **completos** (sem limitações de privacidade, segurança ou privilégios de acesso);
- **primários** (publicados como coletados na fonte, da 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 que possa ser utilizado e processado 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 utilizar formatos de domínio exclusivo de empresas)**;**
- **licenças de licenças** (não estar sujeitos a direitos autorais, patentes, propriedade intelectual ou segredo industrial)

Essas definições vão de encontro à perspectiva de um governo digital, que busca melhorar a qualidade da prestação de serviços públicos e dos processos de gestão por meios digitais e inovadores. O processo de digitalização auxilia na integração entre os órgãos de forma mais eficiente (POSSAMAI, 2010). A Figura 2 apresenta alguns pontos da chamada transição do governo eletrônico para o governo digital.

Figura 2 - Transição do governo eletrônico para o governo digital.

Fonte: Elaborado pela autora (ESTEVANOVIC, 2019) adaptado de (OCDE, 2018), p.1.

Ao propor um setor público orientado a dados é levada em consideração a mudança do paradigma da informação para a gestão e utilização dos dados para prover políticas públicas com maior qualidade e eficiência. De forma complementar, um governo, que inclui no seu processo de transformação digital a abertura como padrão, compromete-se a manter uma postura ativa em prover os dados de forma acessível e aberta, apoiando-se em tecnologias, sem a necessidade de ser provocado a publicar esses conjuntos de dados (OCDE, 2018).

A importância de análises guiadas por dados é trazida no artigo "*Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis*" que

aborda a necessidade de o conhecimento fazer parte, não somente no momento da extração de dados, mas durante o processo de análise deles a fim de produzir resultados consistentes e reproduzíveis. Essa postura fornece insumos para que os agentes considerem os problemas semânticos e façam novas interpretações ao avaliar os diferentes contextos em que aqueles dados são trabalhados (SCHEIDER; OSTERMANN; ADAMS, 2017).

Essas análises sinalizam uma possibilidade de que ferramentas tecnológicas e digitais podem trazer benefícios para a melhoria da qualidade das informações geradas, pela semântica a ser levada em consideração, e da escalabilidade de projetos que estabelecem modelos conceituais. Além de utilizar os dados para a melhoria de processos, a publicação de dados abertos governamentais pode aprimorar a transparência ao otimizar o acesso aos dados e uma maior participação e colaboração dos cidadãos nas políticas públicas (ÁVILA, 2015; FANTINI, 2015; LIRA, 2014; MARTINS, 2018; PEREIRA, 2017; SANTAREM SEGUNDO, 2015).

Para corresponder às normativas e acompanhar o desenvolvimento tecnológico mundial é necessário estabelecer estratégias institucionais para padronizar a governança e a abertura desses dados. O Brasil como membro da Parceria de Governo Aberto (*Open Government Partnership* ou OGP) se compromete a "fornecer dados de forma oportuna, em formatos fáceis de localizar, compreender e utilizar, e que facilitem a reutilização" (BRASIL, 2011b), ou seja, o governo assume a responsabilidade por publicar dados abertos.

No contexto brasileiro, os movimentos mais expressivos que visam melhorar a utilização dos recursos digitais, proteger usuários e definir as regras para acesso, disponibilização e utilização da informação são: a Lei de Acesso à Informação Pública[6] (LAI); o Código de Defesa do Usuário de Serviços Públicos (CDUSP)[7]; a Lei de Governo Digital[8]; e a Lei Geral de Proteção de Dados Pessoais[9] (LGPD).

Em seus artigos, a LAI se propõe a regular[10] o acesso à informação dos órgãos públicos como um direito para todos os cidadãos e, inclusive, orienta que a gestão da documentação governamental é obrigação da respectiva administração pública (BRASIL, 2011c). Já o CDUSP, além de outras diretrizes, dispõe que os órgãos que devem integrar as bases de dados e compartilhar informações e documentos dos cidadãos, que já estejam em posse

[6] Lei Federal nº12.527, de 18 de novembro de 2011 (BRASIL, 2011c)
[7] Lei nº 13.460, de 26 de junho de 2017 (BRASIL, 2017)
[8] Lei Federal nº14.126, de 29 de março de 2021 (BRASIL, 2011a)
[9] Lei 13.709, de 14 de agosto de 2018 (BRASIL, 2018)
[10] Regras previstas no inciso XXXIII do art. 5ª, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal

daquela entidade (BRASIL, 2017). A LGPD dispõe sobre os tratamentos de dados pessoais, objetivando proteger os direitos fundamentais de liberdade, privacidade e livre desenvolvimento da personalidade (BRASIL, 2018). Por fim, a Lei de Governo digital, expõe explicitamente como diretriz a interoperabilidade de sistemas e a promoção de dados abertos (BRASIL, 2011a). Portanto, o governo brasileiro, seus estados e municípios são obrigados, por força de lei, a publicar seus dados de forma transparente para a população.

Diversos países que possuem iniciativas de publicação de dados abertos governamentais são apresentados no artigo *"Open Data Portal based on Semantic Web Technologies*" de Jovanovik, Trajanov e Kostovski (2012) que avaliam como as tecnologias da *Web* Semântica podem prover a portais que compartilham dados abertos conectados. Junto a isso, os autores buscam criar uma comunidade alinhada aos princípios comuns das tecnologias mencionadas para evitar a separação de informações publicadas na Internet.

Quando se trata de interoperabilidade, é importante prever a implementação de um ambiente que esteja preparado para acolher esses dados. Os artífices que compõem a *Web* de dados (da *Web* semântica) podem atuar como um meio de conectar esses dados. Para isso, é necessário buscar estabelecer um processo para que o conhecimento de um conjunto de dados possa ser compartilhado . Dessa forma, dados abertos governamentais devem, idealmente, utilizar tecnologias que permitam a interoperabilidade de dados. Essas tecnologias podem ser encontradas na *Web* Semântica.

2.2 *Web* Semântica

A *Web* Semântica[11] é um termo apresentado por Tim Berners-Lee, Hendler e Lassila para referir-se ao resultado de um processo de estruturação do conteúdo das páginas da *Web* de forma a acessá-lo e permitir a atuação de agentes informatizados (BERNERS-LEE; HENDLER; LASSILA, 2001). O principal objetivo dessa implementação é tornar o conhecimento amplamente acessível e útil. Para tanto, as chamadas Tecnologias de *Web* Semântica envolvem principalmente formatos de dados que auxiliam na codificação do conhecimento a ser processado por máquinas (HITZLER; KRÖTZSCH; RUDOLPH, 2010).

Desse modo, uma das formas de definir a formação da *Web* Semântica como conceito se divide em três tópicos: construir modelos, processar o conhecimento e trocar informações. A reunião desses conceitos constrói a base do que deve ser esperado ao tratar desse assunto em termos de complexidade de desenvolvimento e esforço conjunto (HITZLER; KRÖTZSCH; RUDOLPH, 2010).

Primeiramente, a construção de modelos auxilia na descrição de certos aspectos da realidade. Para isso, podem ser utilizadas ferramentas tecnológicas que permitam representar o conhecimento relacionado a um domínio de interesse. Uma dessas formas, que será abordada com mais detalhes no Capítulo 3 deste trabalho, são as ontologias.

Num contexto filosófico, uma ontologia pode ser definida como um estudo entre as relações e classes das coisas que existem no mundo. Esse conceito de modelagem, quando abordado por Aristóteles, deriva da observação da realidade e da estruturação do conhecimento (HITZLER; KRÖTZSCH; RUDOLPH, 2010). Hessen (1999) apresenta dois fatores dois fatores importantes que compõem a essência da filosofia: a "visão de si" e a "visão do mundo". No decorrer das épocas e dos diferentes pensadores é criado um movimento pendular entre os pontos elementares abordados por cada um. Em suma, "a filosofia é a tentativa do espírito humano de atingir uma visão de mundo, mediante a autorreflexão sobre suas funções valorativas teóricas e práticas" (1999, p. 9).

Ao longo dos anos, a definição de modelo receberam incrementos quanto a sua estrutura (como herança de conceitos e hierarquia de classes). Além disso, agregar o conceito de taxonomia (a ciência da representação) que auxilia a compreender a ordem natural dos objetos e possibilita uma forma de comunicação comum e a troca de informações (HITZLER; KRÖTZSCH; RUDOLPH, 2010). As taxonomias podem ser definidas como um vocabulário controlado de um domínio de conhecimento, representado em formato que permita recuperar

[11] Do inglês, *Semantic Web.*

informações e conhecimento. Isso auxilia o acesso à informação em um sistema eletrônico por se utilizar de lógica para construir as relações e hierarquias. Ademais, essa categorização taxonômica permite definir uma organização do conhecimento de um domínio específico em um modelo (TERRA et. al 1998 RAMIREZ, 2015; SILVA; SALES; DOS SANTOS, 2020).

A construção de modelos não possui somente um fim científico, mas busca sanar necessidades práticas. Dessa forma, a necessidade de construir sistemas de informação computadorizados traz requisitos de modelagem conceitual. Portanto, nesse contexto, o conceito de ontologia é uma descrição formal processável por máquina sobre o conhecimento em um domínio de interesse. Isso posto, a representação do conhecimento se torna mais complexa e ao se beneficiar das tecnologias da *Web* essas modelagens ganham escala e permitem desambiguar termos (HITZLER; KRÖTZSCH; RUDOLPH, 2010). A seção 3.1 explora esses conceitos mais profundamente.

O segundo tópico traz junto aos recursos de informática a capacidade de inferir ou deduzir, a partir de relações lógicas, conhecimento de um determinado domínio. A máxima da lógica ao ser representada em sistemas permite simular o comportamento lógico inerente aos seres humanos. Então, ao afirmar que: $Todo\ homem$ é $mortal \rightarrow Todos\ os\ brasileiros$ são $homens$ é possível inferir que $Todo\ brasileiro$ é $mortal$. Essa forma de inferir conhecimento a partir de fatos pode ser reproduzida em diferentes domínios. Esses axiomas declaram quais as regras e relações entre o conhecimento existente (HITZLER; KRÖTZSCH; RUDOLPH, 2010).

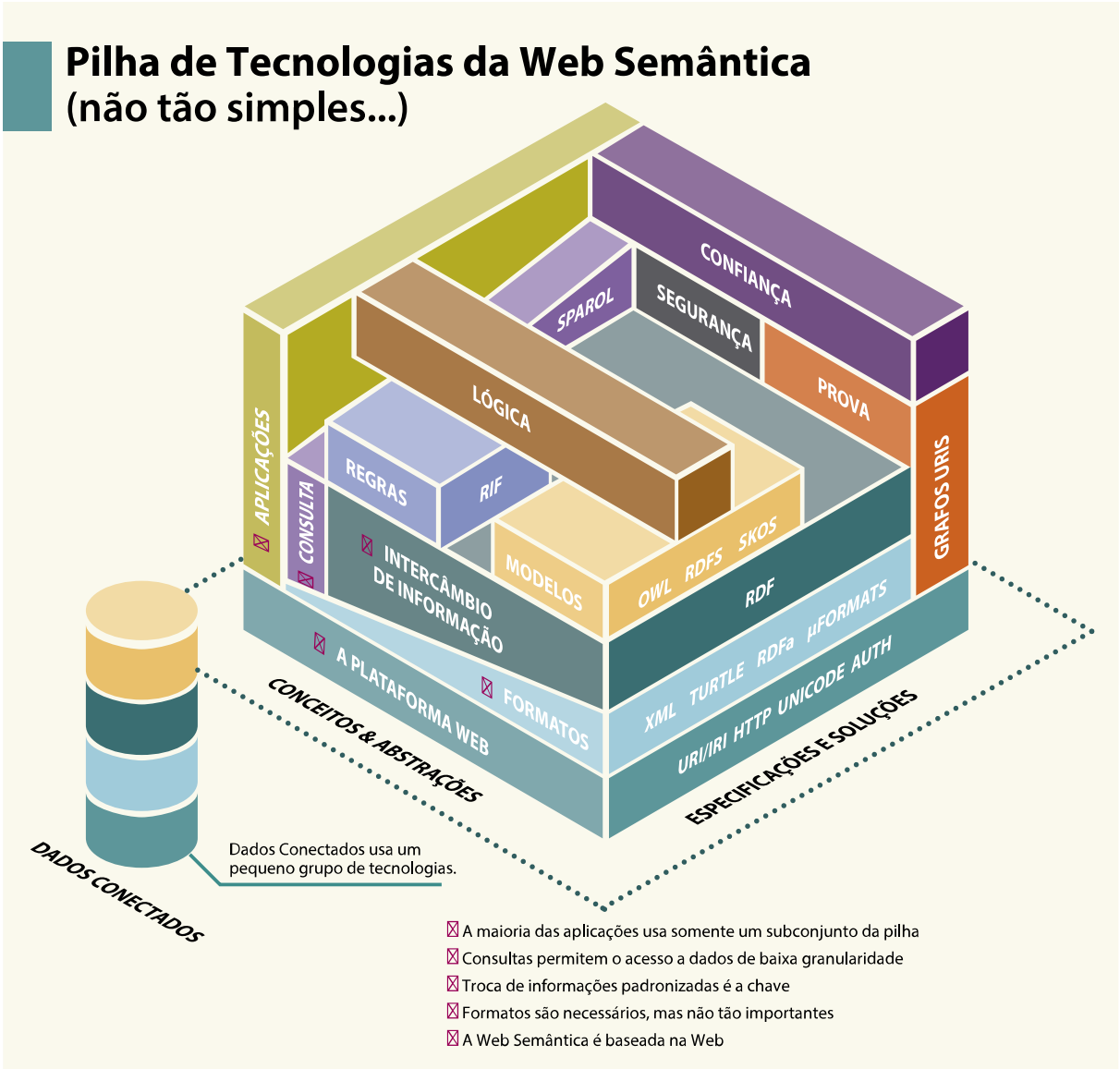
O terceiro tópico é a troca de informação. A comunicação é parte importante da construção do conhecimento. No contexto da *Web*, é possível reaproveitar definições de outras fontes através de protocolos e outras tecnologias que a infraestrutura da *Internet* possui. Isso é possível uma vez que, delimitam-se padrões para formato de dados e sejam utilizados *hyperlinks* que permitem a conexão entre vários conceitos representados (HITZLER; KRÖTZSCH; RUDOLPH, 2010). A ênfase na universalização do protocolo de troca de informação e da linguagem comum para obter dados em diferentes fontes é uma das bases para a busca de conhecimento em outras bases de dados.

A *Web* Semântica possibilita a troca de informações de forma mais eficaz, principalmente em razão da necessidade da formalização dos conceitos e das declarações de significado. Expressar o conhecimento em um formato específico e padronizado é uma tarefa de modelagem de domínio e que pode ser mais efetiva ao reduzir o escopo do modelo a ser construído. Em resumo, o modelo deve possuir um uso intencional, definido. Dessa forma, ao delimitar um contexto, é possível expressar as características semânticas (a interpretação formal

ou significado do texto) e inferir informações sobre aquelas representações. Essas padronizações são propostas pelo *World Wide Web Consortium* (W3C) e facilitam a troca de informações enriquecidas com semântica.

Como uma extensão *Web* clássica, a *Web* Semântica é composta de várias camadas com níveis de conceitos e abstrações que utilizam diferentes soluções tecnológicas para cada uma delas. A Figura 3 ilustra a complexidade e a formação ordenada da arquitetura tecnológica necessária para implementar as funcionalidades propostas a serem expressas por informações mais precisas e com a possibilidade de serem interpretadas por máquinas (ISOTANI; BITTENCOURT, 2015).

Figura 3 - Composição das camadas da *Web* Semântica.

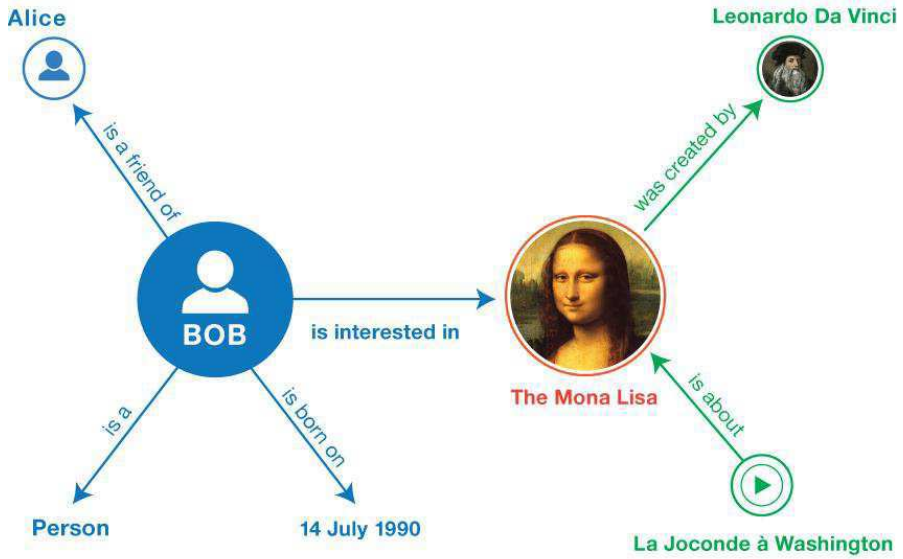


Fonte: Isotani e Bittencourt (2015, p. 30)

A base das tecnologias de *Web* Semântica é a plataforma da *Web* que possui especificações básicas que sustentam o restante das ferramentas, sendo as principais: URI ou *Internationalized Resource Identifier* (IRI), *Hypertext Transfer Protocol* (HTTP). A troca de informações é baseada no *Resource Description Framework* (RDF) e o acesso a esses dados pode ser utilizando a linguagem para acesso de dados, o *SPARQL Query Language for RDF* (SPARQL). A camada dos modelos da composição tecnológica da *Web* Semântica apresenta os tipos de linguagens utilizadas para representar os vocabulários com definições de conceitos e relacionamentos, com o objetivo de auxiliar na integração de dados e mitigar problemas como ambiguidade e repetição (W3C, 2015).

O RDF é um padrão utilizado para persistir[12] e compartilhar informações na *Web*. Essa linguagem é composta por triplas que relacionam [SUJEITO] – [PREDICADO] – [OBJETO]. Esses conjuntos de relações formam grafos que conectam dois pontos por uma aresta. Esses blocos formam relações entre conceitos que podem ser utilizados para fazer declarações sobre o domínio estudado. As chamadas triplas utilizam as tecnologias da *Web* Semântica (URIs e HTTP, principalmente) para criar conexões entre as classes, instâncias e propriedades (ALLEMANG; HENDLER, 2008; SAGI et al., 2022). A Figura 4 exemplifica como a representação em RDF relaciona as informações de classes e suas propriedades e como é possível inserir as entidades específicas de cada classe.

Figura 4 – Exemplo de um grafo RDF que representa algumas instâncias de um conjunto de informações.



Fonte:W3C (2014).

[12] Gravar em um banco de dados.

Como representado no grafo, ao relacionar a pessoa Bob com seu interesse na Mona Lisa, é criada uma tripla que relaciona esses conceitos por uma propriedade. Dessa forma, é possível especificar diferentes relações com as respectivas restrições utilizando a linguagem do RDF (W3C, 2014). Para representar a semântica contida nesses conjuntos, é necessário adotar a especificação do RDF-*Schema* (RDF-S) que viabiliza representar esses recursos por se tratar de um vocabulário que aprimora o RDF ao permitir a criação de hierarquias entre conceitos (ISOTANI; BITTENCOURT, 2015).

Para isso, são utilizadas *tags* que denotam as especificidades de cada uma dessas propriedades destacadas aqui:

- rdf:Property: utilizada para definir as relações entre objetos e sujeitos;
- rdfs:domain: atribui àquela classe uma propriedade de sujeito daquela relação;
- rdfs:range: define que aquele valor na outra ponta do relacionamento é o objeto daquela relação, realizada pelo sujeito por meio de uma ação, que é a propriedade que os conecta;
- rdfs:subClassOf: define qual é a superclasse de um conceito.

Contudo, essas propriedades não são suficientes para declarar todas as peculiaridades de cada domínio. Por isso, foi criada uma linguagem mais expressiva que permite representar o conhecimento por meio de descrições baseadas em ontologias. A *Web Ontology Language* (OWL) auxilia a consulta às informações descritas e suporta a inferência para gerar novos conhecimentos (W3C, 2012), a partir da lógica das ontologias que será explicada de forma mais profunda na Seção 3.1 deste trabalho. Alguns dos recursos mais utilizados na OWL são:

- owl:Class: utilizada para definir uma classe que permite agrupar indivíduos de forma hierárquica aos outros conceitos;
- owl:ObjectProperty: utilizado para definir relacionamento de objetos de um domínio;
- owl:DataProperty: utilizado para definir relacionamentos entre classes e dados que a representam.

Essas linguagens são elementos-chave das denominadas tecnologias da *Web* Semântica. Depois de gerar as triplas, representando o conhecimento por uma ontologia[13], é

[13] Conceito explorado em mais profundidade no Capítulo 3.

possível persistir essas informações em diferentes formatos serializados do RDF: RDF/XML, *Turtle* ou *N-Triples*.

O formato *Turtle* é considerado o mais compreensível para leitura humana por organizar o conteúdo em blocos que separam as propriedades (de objeto e dados) das classes, o que facilita identificar as triplas uma vez que os prefixos utilizados são resumidos nas primeiras linhas do código. Um exemplo da interação dessas duas linguagens de representação é apresentado na Figura 5. As sete primeiras linhas apresentam as IRIs e os prefixos que serão substituídos, dessa forma quando for apresentado "owl:" está implícito que se trata daquele link já descrito. A declaração do *Domain* (`rdfs:domain:ResponsavelNome`) e o *Range* (`rdfs:range:Indicacao`) da propriedade (`owl:ObjectProperty`) *fazIndicacao,* associa uma indicação a um responsável.

Figura 5 – Representação simplificada de uma tripla serializada em formato Turtle.

```
@prefix : <http://www.semanticWeb.org/freya#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@base <http://www.semanticWeb.org/freya> .

<http://www.semanticWeb.org/freya> rdf:type owl:Ontology .

###  http://www.semanticWeb.org/freya#fazIndicacao
:fazIndicacao rdf:type owl:ObjectProperty ;
              rdfs:subPropertyOf owl:topObjectProperty ;
              rdfs:domain :ResponsavelNome ;
              rdfs:range :Indicacao .
```

Fonte: elaborado pela autora.

Esses recursos, apresentados pelos prefixos acima, colaboram para trazer aspectos de semântica para os dados disponíveis e criam conexões entre eles. O crescimento das bases de dados e conteúdo na *Web* fortalece a necessidade da utilização desses padrões apresentados. Neste ponto, as noções de Dados Conectados ou *Linked Data* podem ser apresentados para conjugar as necessidades de publicação de dados abertos governamentais levando em consideração os aspectos semânticos deles (HITZLER; KRÖTZSCH; RUDOLPH, 2010; RECTOR et al., 2019).

2.3 Linked Open Data ou Dados Abertos Conectados

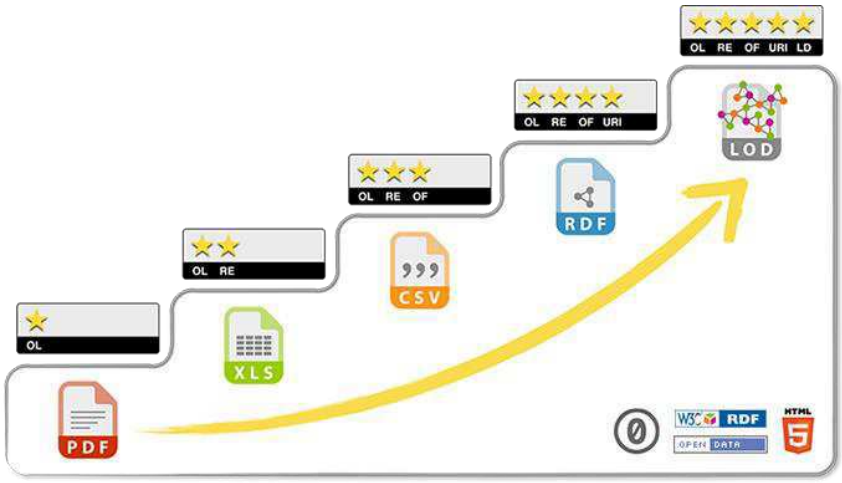
A expressão *Linked Data* (Dados Conectados ou LD) se refere a um conjunto de boas práticas para compartilhar e conectar dados estruturados na *Web*. Os recursos da *Web* Semântica são utilizados para representar formalmente o conhecimento de um conjunto de informações. As oportunidades criadas pelos padrões estabelecidos para os Dados Conectados têm o potencial de aprimorar a prestação de serviços públicos e, dessa forma, é possível automatizar o acesso a dados atualizados e integrados, formando grandes bases de conhecimento na *Web* de Dados (AUER, 2014; CARBONARO, 2021; ISOTANI; BITTENCOURT, 2015).

Quando as informações se tornam links na Internet, é possível conectar diferentes temas (também chamados de domínios) e potencialmente automatizar tarefas e processos, além de permitir o uso para usuários e máquinas. Então, Berners-Lee (2006) enumera quatro regras básicas:

1. Utilizar os *Uniform Resource Identifier*, ou Identificador Uniforme de Recursos (URI) para nomear objetos (classes ou propriedades) na Internet;
2. Utilizar o protocolo *Web*, HTTP, para permitir a busca por esses objetos;
3. Disponibilizar metadados que possam prover propriedades importantes em linguagens de representação na *Web*;
4. Incluir outros links que já existem na *Web* para que mais informações sejam conectadas.

A partir desses padrões definidos, existe a possibilidade de controlar a qualidade dos dados, criar rotinas de manutenção e governança de dados (ÁVILA, 2015). As iniciativas que objetivam disponibilizar as publicações de forma aberta embasam o conceito de Dados Abertos Conectados (*Linked Open Data* ou LOD). Os princípios que definem os principais padrões de desenvolvimento e publicação podem ser resumidos pelo “Sistema das cinco estrelas” (Figura 6) proposta por Berners-Lee (BERNERS-LEE, 2006).

Figura 6 – Padrão cinco estrelas dos dados abertos conectados



Fonte: (5STARDATA, 2021).

Nessa classificação é definida a quantidade de estrelas que compreendem o nível de evolução daquele conjunto de dados publicados no formato conectado. O Quadro 1 apresenta a relação entre a representação e sua definição. Quando se trata do primeiro nível aqueles dados foram disponibilizados na Internet em qualquer formato, legível ou não por máquinas, e são de uso livre. Já o segundo acrescenta, especificamente, que o formato da publicação deve estar estruturado, tabular. O terceiro inclui a necessidade desses arquivos estarem em um formato não proprietário, ou seja, que sejam livres de patentes de empresas para evitar limitações legais a respeito do seu uso. O quarto aponta que aqueles dados deverão ser apresentados de acordo com os padrões estabelecidos pela W3C que permitem a identificação das estruturas de dados utilizando URLs e as tecnologias da *Web* Semântica para reutilizar informações de outros conjuntos. O último define que, além das regras anteriores, aqueles dados devem se conectar a outros de forma a melhorar a rastreabilidade e interoperabilidade entre diferentes domínios.

Quadro 1 – Apresentação em formato tabular da quantidade de estrelas do padrão de dados abertos conectados e suas definições.

Quantidade de Estrelas	**Definição**
★	Disponível na Internet, em qualquer formato e com licença aberta (por exemplo em .pdf)
★★	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) 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.

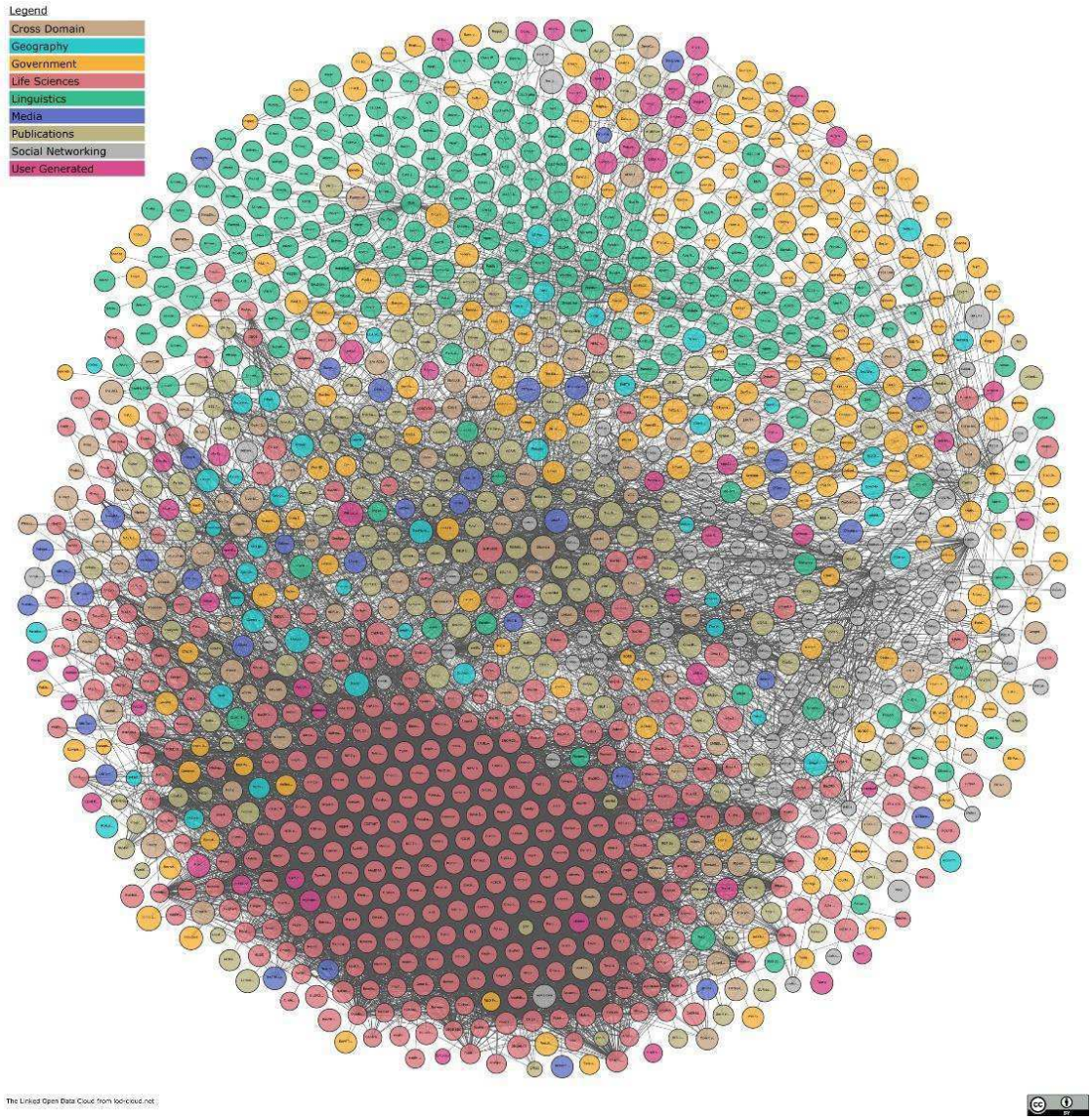
Fonte: adaptado de 5 Star Data (2021).

Cabe destacar que o RDF, de acordo com o artigo de Christoph Lange (2013) "*Ontologies and languages for representing mathematical knowledge on the semantic Web*", possui uma capacidade superior de conexão quando comparado a outras linguagens para representação do conhecimento. Isso se deve à possibilidade de reutilização de conceitos definidos em outro contexto sem impedir o processamento por máquinas.

Diferentes organizações (como Google, LinkedIn, BBC e Governos) utilizam essas tecnologias de dados conectados e, com o objetivo de tornar os dados disponíveis para todos, uma iniciativa da comunidade criou o *Linked Open Data Project*, que reúne diversos vocabulários RDF conectados na chamada "Nuvem LOD[14]" representada na Figura 7. Cada um dos círculos representa um vocabulário diferente que se conecta a outro e cada cor representa uma área diferente: Geografia, Governo, Ciências Biológicas, Linguística, Mídia, Publicações, Redes Sociais, Geradas por Usuários e Domínios Transversais. O portal do projeto disponibiliza "sub-nuvens" que apresentam as conexões pormenorizadas de cada domínio e, ao interagir com as imagens apresentadas é possível navegar pelo grafo gerado e observar as conexões apresentadas.

[14] No inglês, *Linked Open Data Cloud.*

Figura 7 - Nuvem de Dados Abertos Conectados.



Fonte: LOD Project (2022).

A DBpedia, representada nessa nuvem de dados abertos conectados, consiste em um vocabulário elaborado a partir do esforço conjunto de extrair e disponibilizar as informações da Wikipedia na *Web*. Essa iniciativa visa permitir a pesquisa acerca do conteúdo contido nas páginas da referida enciclopédia. O acesso a esse conhecimento estruturado, de acordo com as "Tecnologias de *Web* Semântica", permite interligar e reaproveitar conceitos tanto a partir da base de conhecimento criada, quanto de outras bases e, até mesmo, criar aplicações que utilizem essas informações (AUER et al., 2007).

Todas essas definições podem justificar a necessidade de adotar formatos para a abertura de dados no contexto governamental. Além de aumentar a transparência e o reuso das informações, também são fomentadas as atividades do terceiro setor que pode utilizar esses

dados para construir plataformas que beneficiem a população e, por fim, as definições corroboram com o aumento da participação da população nas atividades de governo, já que o acesso à informação permite uma influência ativa dos cidadãos nas decisões de governo (SILVA, 2018).

Com a necessidade de publicar dados abertos governamentais cresce o volume de informações disponibilizadas em diferentes formatos. Apesar do esforço na publicação de dados, o que favorece o *accountability* e traz benefícios para a transparência, faltam definições semânticas que tragam clareza àqueles dados representados. De maneira geral, um conjunto de dados e seus conceitos podem ser interpretados de forma equivocada ou diferente da aplicação naquele contexto (HOXHA; BRAHAJ, 2011). Mesmo com mecanismos da *Web* Semântica, ainda são necessárias ferramentas que garantam a representação do conhecimento e que sigam os princípios dos Dados Conectados. Assim, no próximo capítulo são apresentados os conceitos relacionados à anotação semântica de dados.

3 ANOTAÇÃO E ENRIQUECIMENTO SEMÂNTICO DE DADOS

A expressão Dados Conectados se refere a um conjunto de informações organizadas em triplas e compartilhadas em linguagem RDF. Entretanto, para garantir a interoperabilidade, integração e reutilização de conceitos é necessário avançar para abordagens que enfatizem a semântica dos dados. As ontologias, que utilizam OWL, possibilitam a representação formal do significado e o enriquecimento dos dados (ALMEIDA, 2021).

3.1 Ontologias

Dentre as tecnologias da *Web* Semântica, existem as ontologias que podem ser utilizadas para melhorar, integrar e encontrar dados em diferentes domínios. O raciocínio automático que é possibilitado pelas ontologias também pode ser chamado de inferência (ALMEIDA, 2021; BERNERS-LEE; HENDLER; LASSILA, 2001).

As ontologias são uma forma de explicitar e formalizar uma conceitualização compartilhada. A **conceitualização** se relaciona à necessidade de criar um modelo abstrato para descrever um modelo e as relações entre os conceitos relevantes para aquele contexto. **Explícito** porque todos os conceitos mais relevantes devem ser definidos e descritos, do contrário estará incompleto. **Formalizado** para que aquela descrição possa ser legível e processável por máquinas – ou seja, o computador precisa conseguir interpretar a semântica daquelas relações e representações do conhecimento. Por fim, **compartilhado,** pois aquele conhecimento deve ser um consenso entre pessoas, enfatizando a necessidade da participação de um grupo de pessoas para definir uma ontologia (GRUBER, 1995; RECTOR et al., 2019).

Normalmente, as ontologias são compostas de: indivíduos, classes, atributos e relacionamentos. A construção de ontologias é baseada em lógica de descrição, e por isso é possível anotar e categorizar o que pertence ou não àquele conjunto. Isso permite dizer que, quando interseccionadas duas classes que não possuem elementos em comum, ou seja, são disjuntas, é possível inferir que são classes diferentes. A lógica diz que classes também podem ser subclasses, que por consequência herdam as características e propriedades da sua classe pai (SEQUEDA; LASSILA, 2021). Além desses aspectos, essa tecnologia possui uma forma de inferir conhecimento entre as relações das classes, o que gera novos conhecimentos (SANTOS et al., 2017). A formalização do conhecimento é essencial para a homogeneização da informação uma vez que cada pessoa pode interpretar a linguagem natural de formas diferentes e isso pode interferir na análise dos dados. Portanto, quando uma base de dados é anotada, é possível que existam menos erros.

As ontologias são baseadas em Lógica de Descrição, o que permite descrever uma série de restrições existentes no domínio representado. Além de possuir uma forma de inferir novos conhecimentos (SANTOS et al., 2017). Dessa forma, é possível representar o conhecimento utilizando preceitos lógicos e formalizá-lo, além de traduzi-lo para uma linguagem computacional (ALEXOPOULOS, 2020; OBITKO, 2007).

Existem três componentes básicos para construir uma ontologia: Classe, Propriedade de Tipo de Dados, Propriedades do Objeto[15]. As classes representam os grupos abstratos ou coleções de objetos e são caracterizadas por atributos. Em uma ontologia, as classes podem ser relacionadas entre si, a partir de propriedades de objeto. De acordo com argumentos lógicos, essas também podem ser subclasses, que por consequência herdam as características e propriedades da sua classe pai (SEQUEDA; LASSILA, 2021). As propriedades de tipo de dados conectam classes a uma entidade instanciável, que pode representar uma medida ou um formato esperado.

Os tipos de ontologia mais comuns são (GUARINO, 1998 apud REIS JÚNIOR, 2020):

- Topo ou nível superior: descreve domínios genéricos e que podem servir para confeccionar novas ontologias.
- Tarefa: utiliza conceitos das ontologias de alto nível para declarar conceitos relativos a atividades diversas.
- Domínio: reutiliza definições das ontologias de topo para especializar a modelagem de conceitos a respeito de uma necessidade ou situação específica que se deseja representar
- Aplicação: trata-se de ontologias específicas que compreendem conceitos relacionados aos papéis que podem ser desempenhados por diferentes agentes do domínio que realizam uma tarefa

As ontologias são representações de conhecimento e permitem a inferência de dados sobre as proposições declaradas. A formalidade proporcionada por esse artefato reduz a ambiguidade semântica dos conceitos, o que melhora a comunicação daquele significado e aumenta a precisão das informações relacionadas. Além disso, as inferências processadas por máquinas permitem expandir o conhecimento explícito ao revelar as relações implícitas àquele domínio (ALMEIDA, 2021; SANTOS et al., 2017).

[15] É mais comum ver essas definições em inglês, que aqui foram traduzidas. Na língua inglesa são: *Class, Datatype Property* e *Object Property*

O autor Lamy Jean-Baptiste (2021) sugere dois propósitos principais para as ontologias: raciocínio automático e reuso de conhecimento. O primeiro retoma uma das principais características desse artefato que é a possibilidade de realizar inferências a partir da lógica. O segundo evoca os princípios da *Web* Semântica ao reutilizar *links* de outros recursos na *Web*.

3.2 O Dicionário Semântico de Dados

Anotar algo implica, de forma geral, relacionar e estabelecer coerência entre os elementos daquele domínio. No contexto de dados, descrever os metadados de uma base de dados é uma forma de deixar claro para o usuário qual é o escopo e quais as informações mais importantes de cada coluna. Quando se fala sobre anotação de dados o mais comum é se pensar em uma documentação da base de dados que normalmente é traduzida em um dicionário de dados – que pode ser definido como um repositório centralizado de informações sobre aqueles dados, seus significados, formato, origem, uso e relações que descreva o significado de cada variável. A utilidade dessa ferramenta é indiscutível e por isso a proposta deste trabalho não é ignorar esse potencial, mas aprimorá-lo utilizando tecnologias de *Web* Semântica (BELISÁRIO et al., 2020; RASHID et al., 2020).

Entretanto, como visto até aqui, no contexto de Dados Abertos Conectados na Internet, surge a necessidade de deixar esses aspectos explícitos e formalizados. De acordo com Belloze *et. al.* (2012), a anotação semântica é uma associação entre termos e expressões de um contexto que estão descritas em uma ontologia.

O exercício da anotação ou da modelagem podem servir para diferentes fins a partir de cada necessidade. Um modelo deve possuir uma motivação para sua concepção, seja para organizar recursos de valor para uma empresa ou representar um caso específico que merece a atenção. Para isso, é importante delimitar quais serão os parâmetros para analisar se aquele modelo construído atende aos requisitos planejados e se ele atinge o objetivo que levou à criação dele. Uma forma de balizar essas necessidades é utilizando perguntas de competência que possibilitam iniciar a modelagem semântica. Isso não significa que um modelo deve estar limitado a essas questões (ALLEMANG; HENDLER, 2008). A abordagem trazida pelos autores se alinha com as necessidades da conexão e reutilização de conhecimento que podem ser mitigadas por um processo de anotação semântica.

Aprimorar as anotações presentes nos dicionários de dados com semântica otimiza o esforço dos responsáveis pelos dados ao torná-las apropriadas para o consumo de máquinas,

o que permite combinar diferentes conjuntos de dados de diversas fontes. Mais do que somente descrever os objetos é necessário representá-los.

A integração de dados utiliza diferentes técnicas para combinar dados de fontes diversas com o objetivo de gerar informações com significados confiáveis para a organização. Em adição a isso, a integração semântica de dados combina conjuntos de dados de fontes distintas e os combina conceitualmente, utilizando ontologias e tecnologias de semântica. Cenários com níveis de complexidade elevados e de gestão descentralizada podem se beneficiar da criação de uma ontologia de domínio que possa organizar o conhecimento daquela organização e que permita a integração. Para aproveitar o máximo potencial desse tipo de conexão, é importante que exista uma metodologia que possa padronizar o desenvolvimento desse ferramental (GONÇALVES, 2020).

Nesse contexto, o Dicionário Semântico de Dados (*Semantic Data Dictionary* ou SDD), apresentado por Rashid et. al (2017), apresenta um conjunto de padrões de metadados baseados em ontologias que descrevem um conjunto de dados existentes. Essa técnica é realizada de forma manual, uma vez que a anotação parte de uma tabela de dados e representa suas colunas em classes e conceitos em uma ontologia. Ao formalizar a semântica dos dados utilizando esses artefatos, é possível interligá-los a outras fontes de dados.

Essas características favorecem o uso do SDD, uma vez que a representação alcançada é processável por máquinas, mesmo com a simplificação do processo. Isso se dá pelo fato de o processo de anotação ser separado da implementação tecnológica. Além disso, uma outra vantagem de utilizar esses dicionários semânticos em modelos dimensionais (ou relacionais) é a possibilidade de harmonizar interpretações de conceitos em diferentes domínios. Essa integração semântica entre dados de diferentes bases é um diferencial da ferramenta. Na prática, é necessário selecionar as classes de cada tabela e de maneira iterativa anotar os termos existentes com o uso de ontologias. No Capítulo 5 serão apresentados os *templates* do SDD preenchidos de acordo com o estudo de caso do trabalho.

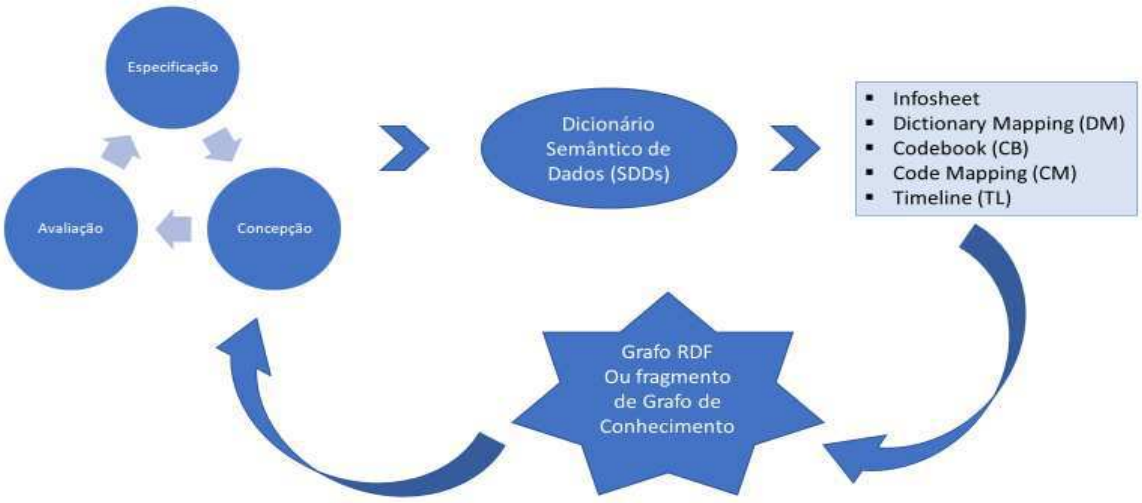
Após o processo manual de anotação e preenchimento dos arquivos, é utilizado um *script* que interpreta o SDD e o converte em um grafo de conhecimento no padrão RDF que poderá ser consultado. Isso permite a formalização do vocabulário e possibilita a interoperabilidade e conexão a outros domínios (MOREIRA, 2021; TETHERLESS WORLD, 2021).

A abordagem do SDD fornece uma metodologia que permite anotar as colunas de um conjunto de dados específicos utilizando uma (ou mais) ontologia(s) que representa(m) aquele domínio. O resultado da utilização dessas ferramentas pode ser traduzido em um grafo

de conhecimento (*Knowledge Graph* ou KG)[16] que pode ser submetido a *querys* e pode compor gráficos e relatórios sobre o conhecimento ali relacionado. O SDD é uma ferramenta que acelera o processo de engenharia de ontologias.

Essa perspectiva propõe a modularização do processo de anotação e divide o conteúdo em diferentes documentos em formato tabular (normalmente Excel ou *.csv*), que são: *Infosheet, Dictionary Mapping, Codebook, Code Mapping* e *Timeline.* A Figura 8 apresenta um resumo gráfico de como a metodologia ágil e o SDD se beneficiam do desenvolvimento em ciclos.

Figura 8 - Resumo gráfico simplificado do processo de anotação semântica de dados



Fonte: elaborada pela autora.

De maneira geral, existem ciclos separados de desenvolvimento que se complementam. Primeiramente, para este estudo de caso é necessário desenvolver uma ontologia para a base de dados selecionada. Em paralelo, inicia-se a construção das ferramentas do SDD e são realizados testes com as perguntas de competências e isso retorna para o desenvolvimento da ontologia e o ciclo segue dessa forma.

O método permite maior interação entre os atores (especialistas de domínio e especialistas do conhecimento) e o desenvolvimento fracionado das ferramentas se adapta à esfera complexa do domínio que envolve muitas variáveis, relações e propriedades definidas formalmente em diferentes legislações e sistemas informatizados. Esse esforço de "dividir para

[16] Para aprofundar nas possibilidades de construção e aplicação dos grafos de conhecimento, o estudo de Xiaohan Zou (2020) pode ser um passo inicial para compreender melhor o potencial dessa tecnologia.

conquistar" é primordial. Neste ponto, a modelagem conceitual ontológica é vantajosa uma vez que sua estrutura é dinâmica e flexível.

Os atores envolvidos no processo de governança de dados, na sua maioria são conhecedores do domínio, podem não possuir refinamento técnico para lidar com a construção de uma ontologia e isso pode tornar a preparação dos dados muito difícil ou até mesmo impossibilitá-la. É comum que os especialistas de domínio e outros participantes/pessoas envolvidos no processo de governança dos dados prefiram iniciar as modelagens com modelos gráficos, que podem ser rediscutidos e inseridos na Metodologia Ágil proposta por Bax e Gonçalves (2020). A Tabela 1 apresenta um exemplo de um conjunto de dados que pode ser representado com a técnica do SDD.

Tabela 1 - Exemplo de conjunto de dados a ser representado em um SDD.

Id	Nome	Precipitacao	Temperatura	Saneamento	TaxaDengue
2903508	Belo Campo	512	31.8	59.2	6.6
2927408	Salvador	1579	28.6	0.7	227.8

Fonte: Adaptado de Gonçalves (2020 p.76)

O *Dictionary Mapping* apresenta uma lista dos objetos que serão instanciados. Ou seja, no princípio ele reproduz na coluna "Rotulo" as colunas de uma tabela e os mapeia para a ontologia base. Na sequência, as entidades implícitas daquelas relações são mapeadas e definidos os atributos e propriedades delas Tabela 2. A Tabela 3 apresenta um exemplo de *Codebook (CB)* que permite categorizar e classificar as entidades do conjunto de dados que possuem diferentes classes.

Tabela 2 – Exemplo de *Dictionary Mapping* (DM).

Rotulo	Atributo	*isAttributeOf*	Entidade
Id	hasco:originalId	??Munic	
Nome	:NomeMunicipio	??Munic	
Precipitação	:Precipitacao	??Superficie	
Temperatura	:Temperatura	??Superficie	
Saneamento	:Saneamento	??Residencias	
TaxaDengue	:Taxa-Dengue	??Populacao	
??Munic			:Municipio

Fonte: Adaptado de Gonçalves (2020, p. 76)

Tabela 3 – Exemplo de um *Codebook* (CB).

Column	Label	Class
Genero	Masculino	:Masculino
Genero	Feminino	:Feminino
Genero	Não Binário	:NaoBinario

Genero	Outros	:Outros

Fonte: Elaborado pela autora.

Após o processo manual de anotação e preenchimento dos arquivos apresentados, é utilizado um *script* que interpreta o SDD e o converte em um grafo de conhecimento no padrão RDF (MOREIRA, 2021; TETHERLESS WORLD, 2021). Como exemplificado no trabalho de Bax e Silva (2020) a aplicação do *script sdd2rdf*[17] sobre os artefatos descritos acima, juntamente com o conjunto de dados a ser anotado, gera o arquivo RDF contendo o Grafo de Conhecimento.

3.3 Grafos de conhecimento

Os grafos de conhecimento (no inglês, *Knowledge Graphs* ou KG) são comumente definidos como uma grande coleção de dados estruturados unidos de forma significativa. Por extrapolarem os conceitos de Dados Conectados e modelos RDF, os KG permitem realizar tarefas e inferências particulares da técnica. 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; SEQUEDA; LASSILA, 2021).

De forma generalista, a definição de um KG, de acordo com Paulheim (2017), considera quatro características: (1) descreve entidades e suas relações considerando conceitos existentes utilizando um grafo (uma anotação formal e estruturada); (2) define classes e relações entre entidades em *schema* (conjunto de tipologias que podem relacionar-se um conjunto de propriedades); (3) permite conectar entidades de outros domínios; e por fim (4) não limita a cobertura do grafo de conhecimento a somente um tópico do conhecimento. Além disso, é possível utilizar-se dessa tecnologia para criar fragmentos de grafos que descrevem objetos de interesse e finalidades específicos (BAX; SILVA, 2020).

Dessa forma, o padrão RDF (W3C, 2014) pode ser utilizado para representar esses grafos de conhecimento tendo em vista 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.

[17] Disponível em: https://github.com/tetherless-world/SemanticDataDictionary/tree/master/sdd2rdf . Acesso em 29 ago 2022.

3.4 Trabalhos correlatos

O trabalho apresentado por Hoxha e Brahaj (2011) trata questões que envolvem a publicação de dados abertos governamentais de acordo com os princípios do *Linked Data*. A proposta do artigo é analisar dados estatísticos brutos de diferentes fontes e publicar esses dados em um conjunto único, acessível pela *Web*. O processo descrito perpassa a construção de uma ontologia de domínio e permite, de acordo com os resultados apresentados, a integração de dados o que permite a análise combinada com outras fontes, além de facilitar a criação de aplicações inovadoras e criativas.

O enriquecimento de dados conectados apresentado por Buchmann e Karagiannis (2016) em "*Enriching linked data with semantics from domain-specific diagrammatic models*" é baseado em um modelo chamado RDFizer e provê um exemplo com argumentos generalizáveis para propor uma visão de modelo conceitual ontológico em um sistema de informação. O artigo demonstra que existem benefícios e limitações quanto a esse desenvolvimento. Os pontos positivos da pesquisa são a possibilidade de criar um modelo de dados legível por máquina e que pode tratar tanto os dados existentes no sistema fora do modelo quanto os gerados pela interação com a representação. Além disso, a visão de um modelo conceitual que cresce à medida que as necessidades surgem é trazido como uma oportunidade do modelo. Apesar disso, um ponto negativo trazido é a lentidão para implementar processos que adotam as metodologias de *Web* Semântica que necessitam de um esforço contínuo para continuarem funcionais.

O artigo "*From data to city indicators:a knowledge graph for supporting automatic generation of dashboards*" apresentado por Santos *et al*. (2017) apresenta uma forma de gerar painéis de acompanhamento automaticamente utilizando grafos de conhecimento como base para a construção dos indicadores de Cidades Inteligentes. O trabalho foca em descrever operacionalmente o processo para isso. O processo de serialização proposto é tecnicamente complexo e a implementação dele no estudo de caso deste trabalho se tornaria inviável pelo tempo a ser gasto. O estudo desenvolvido permite vislumbrar o potencial de metodologias que utilizam anotação semântica e podem até mesmo automatizar processos, depois de maduros o suficiente.

No texto de Martins, Craveiro e Alcázar (2013), apesar de descrever com detalhes a criação de uma ontologia para o orçamento público federal brasileiro, não disponibiliza um formato de acesso legível por máquina, apenas dispõe como imagens as descrições de conceitos, classes e propriedades, ao trabalho produzido. Ademais, uma crítica ao esforço dos autores é

justamente não buscar, em outras ontologias, conceitos que pudessem ser reaproveitados o que seria de grande proveito para enriquecer o trabalho e compartilhar o conhecimento.

Entretanto, apesar do texto de Martins, Craveiro e Alcázar (2013) não disponibilizar os dados e modelos, foram encontrados no portal do SIOPDoc[18] recursos em RDF para acesso aos dados do orçamento público federal que utilizam da ontologia desenvolvida para conectar dados da execução do orçamento do ano de 2001 a 2020. Um ponto de destaque é o *Endpoint* de consulta SPARQL[19] da base de dados e exemplos para pesquisas e recuperação de dados, inclusive sobre as emendas parlamentares federais[20]. Não foram encontrados outros portais semelhantes a esse, mas, pelo tipo de estruturação e atualização, é possível avaliar que existe uma equipe que é responsável e tem interesse em mantê-lo funcionando.

[18] Disponível em: http://orcamento.dados.gov.br/siopdoc/doku.php/acesso_publico:dados_abertos/ . Acesso em: 20 jan 2022.
[19] Disponível em: http://www1.siop.planejamento.gov.br/sparql/ . Acesso em: 20 jan 2022.
[20] Disponível em: http://orcamento.dados.gov.br/siopdoc/doku.php/acesso_publico:scripts_sparql?&#consultas_sobre_emendas_parlamentares Acesso em: 20 jan 2022.

4 METODOLOGIA

O objetivo deste trabalho é adaptar uma metodologia para enriquecer semanticamente e publicar dados abertos governamentais e aplicá-la a um estudo de caso. No decorrer deste capítulo serão apresentados os métodos e materiais utilizados para atingir os objetivos e, assim, responder o problema de pesquisa.

A metodologia a ser adaptada provém dos trabalhos de Gonçalves (2020) e de Bax e Silva (2020) que buscam sistematizar uma forma de enriquecer dados semanticamente com ontologias e publicar essas informações. Os dois métodos são baseados nos preceitos de Métodos Ágeis e apresentam fases cíclicas e incrementais, o que contribui para construção dos artefatos que, ao serem revisitados, favorecem o aumento da complexidade do modelo.

A base da metodologia utilizada nesse trabalho é a *Design Science Research* (DSR). Esse método contribui ao alinhar o rigor científico com necessidades práticas. Ao não desassociar os dois tópicos, a técnica auxilia o processo de pesquisa a encontrar soluções de problemas teóricos em situações práticas que, por sua vez, podem contribuir com respostas que solucionem as questões de conhecimento postas (BAX, 2013). A *Design Science* auxilia na concepção de artefatos que possam contribuir com a geração de conhecimento e beneficiar a comunidade científica.

Como estratégia para uma pesquisa a DSR fornece insumos que podem orientar a construção do conhecimento em diferentes áreas de aplicação. Para isso, a DSR propõe a iteração entre os artefatos produzidos. Essa relação parte tanto do eixo teórico quanto do eixo prático. Isso permite avançar para além da pesquisa tradicional, adaptando-se às necessidades de uma pesquisa que pode ser aplicável em um problema do mundo real (HEVNER, 2007).

O processo de iteração entre problemas teóricos e práticos pode ser resumido pelo ciclo regulador proposto por Wieringa (2009), ilustrado na Figura 9. A primeira etapa (Problema) aborda a investigação de um problema e busca compreender, descrever e analisar a situação prática que será impactada por uma possível solução. A segunda fase (Projeto) descreve o plano que deverá ser aplicado para solucionar aquele problema. Já a terceira (Validação) inclui as tarefas necessárias para avaliar se o projeto trouxe algum tipo de resposta aos problemas colocados na primeira fase – portanto, essa fase auxilia na previsão dos possíveis resultados da construção dos artefatos. Já a quarta fase (Implementação) pretende analisar a forma de colocar em prática as soluções propostas e executar o planejamento. Por fim, a fase de Avaliação, busca validar e analisar os artefatos que foram implementados e qual será o próximo problema a ser compreendido.

Figura 9 - Ciclo regulador de Wieringa (2009).

Fonte: Adaptado de Wieringa (2009).

Essa característica cíclica da DSR se aproxima das denominadas Metodologias Ágeis, que podem contribuir para a elaboração de diferentes objetos de pesquisa. Adaptadas da Engenharia de *Software*, as metodologias ágeis vêm sendo implementadas em processos de empresas e governos. Essas inovações se beneficiam de um modelo com ciclos curtos de planejamento, desenvolvimento, validação e ajustes (ou PDCA).

Assim como é descrito no manifesto ágil (Quadro 2), é considerado mais importante ter um produto funcional, garantir a interação entre os indivíduos envolvidos no processo e a possibilidade de resposta às mudanças, para além do resultado. Então, é possível adaptar o produto às necessidades ao longo do tempo de execução de um projeto. Isso é relevante quando se leva em consideração a geração de ontologias e outros instrumentos de gestão e organização do conhecimento. Dessa forma, analisar o problema e propor um fluxo guiado por perguntas de competência pode proporcionar uma solução consistente com as necessidades descritas.

Quadro 2 - Comparativo entre valores da metodologia Ágil e tradicional.

ÁGIL		TRADICIONAL
Interação entre os indivíduos	mais que	Processos e ferramentas
Produto em funcionamento	mais que	Documentação abrangente
Colaboração com o cliente	mais que	Negociação de contratos
Responder às mudanças	mais que	Cumprir prazos
VALORES		

Fonte: Adaptado de "Manifesto para Desenvolvimento Ágil de *Software*" (2001).

Para este trabalho, a metodologia de Gonçalves (2020), que foi desenvolvida para suportar a integração semântica de dados no contexto de projetos de pesquisa científica, será adaptada para um cenário governamental que trata das emendas parlamentares impositivas. O estudo em relação a Bax e Silva (2020) foi focado em adaptar os pontos apresentados à uma realidade específica. Essas adaptações estão alinhadas a princípios de aquisição do conhecimento e dos processos iterativos necessários para a construção do conhecimento, assim como descrito na teoria de aprendizado conectivista (FEKETTIA; FARIAS; MUSTARO, 2013; SIEMENS, 2005), em que o aprendizado se dá de forma coletiva e não individualmente.

4.1 A metodologia de Gonçalves (2020)

O trabalho desenvolvido por Gonçalves (2020) apresenta um processo de publicação de dados científicos com foco na integração semântica de dados utilizando ontologias, denominado ODIN. Por se tratar de um nicho específico, o objetivo desse subtópico é explorar os aspectos básicos do que foi desenvolvido a fim de mostrar o que será adaptado para esse trabalho.

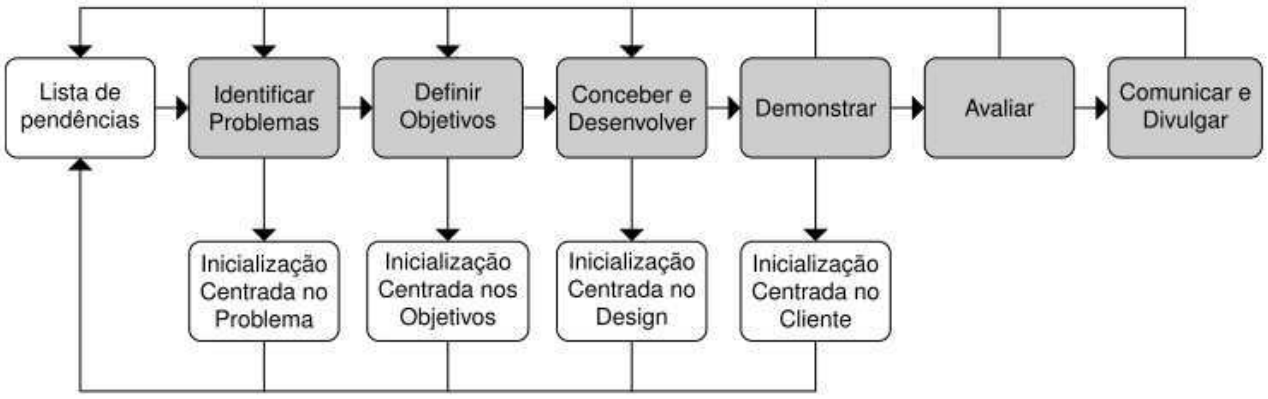
A integração proposta sugere a necessidade de elaborar arquivos com descrições de metadados que qualificam os conjuntos de dados. Esses arquivos são mencionados no texto como artefatos ou *templates* que possibilitam unir os dados a um modelo conceitual, que é a ontologia. Esse formato beneficia a metodologia pois auxilia a definir o domínio estudado, além de organizar e identificar os conceitos mais relevantes para aquele contexto.

A metodologia proposta pelo autor supracitado propõe conectar os arquivos de dados, metadados e ontologias em um *software* onde é realizada a ingestão de dados e geração de grafos de conhecimento. Esses conjuntos de dados enriquecidos poderiam, de acordo com as hipóteses apresentadas nesta pesquisa, ser reutilizados e reproduzidos. Isso é fruto da aquisição de conhecimento formal, o que permite reaproveitar definições de diferentes estudos.

Por utilizar preceitos ágeis, a metodologia descreve um processo que executa passos determinados repetidamente e ao longo dos ciclos aprimora os resultados. Essas atividades são realizadas por um grupo de atores que possuem papéis específicos. Os especialistas de domínio são pesquisadores que possuem interesse no processo de integração e análise dos dados de uma pesquisa. Os ontologistas são os responsáveis por criar e evoluir as ontologias utilizadas no projeto. Já os desenvolvedores são aqueles que colaboram na criação dos artefatos e auxiliam a solucionar problemas de ordem tecnológica. Por fim, os cientistas de dados utilizam seu conhecimento em técnicas estatísticas e de análise de dados para tratar e extrair informações

daqueles conjuntos de dados. As etapas do método, apresentadas na Figura 10, ocorrem com a participação desses atores.

Figura 10 - Etapas da metodologia desenvolvida por Gonçalves.



Fonte: Gonçalves (2020, p. 60).

O fluxo de iteração tem como objetivo gerar um grafo de conhecimento a cada rodada, o que possibilita a integração entre os dados e os artefatos. Uma observação colocada pelo autor da metodologia é que no decorrer do ciclo, caso algo não possa ser realizado, essa atividade retorna para a lista de pendências (ou *Backlog*), que centraliza as prioridades de execução. Cada atividade pode ser descrita como questões de competência que guiam a modelagem ontológica e, portanto, definem a produção dos artefatos.

A primeira etapa é a **identificação do problema**. Nesse momento são selecionadas as perguntas que guiarão a iteração. No primeiro ciclo, os atores se concentram em gerar uma primeira versão dos *templates* para criar a versão inicial do grafo. Para isso, é pressuposta a existência de uma ontologia de domínio que guiará a anotação dos dados. É recomendado mapear os conceitos existentes no cabeçalho do conjunto de dados para uma ontologia base e, quando possível, reaproveitar definições de outras ontologias. Essa "base" deverá representar conceitos provenientes dos dados e de outros indicadores de interesse. Nas outras iterações, as atividades são retiradas da **Lista de Pendências** que é alimentada à medida em que as atividades ocorrem e novas necessidades surgem.

A **definição dos objetivos da solução** se concentra em encontrar respostas às hipóteses de pesquisa definidas pelos especialistas de domínio. Nesse momento são acertadas as características desejadas dos artefatos e como eles poderão responder ao conjunto de perguntas de competência selecionadas.

A etapa seguinte, de **design e desenvolvimento,** inicia-se com a definição e especificação do que será construído em consonância com os problemas e objetivos

estabelecidos anteriormente. Nessa fase, cada perfil de atores possui tarefas específicas para serem desenvolvidas em razão das suas especialidades. Isso é fundamental para o primeiro ciclo do método em que é realizada a primeira implementação da ontologia base e dos *templates*. Esse mapeamento inicial é mais complexo e demanda maior esforço para executar[21]. Aqui destaca-se a construção da primeira versão da ontologia que busca representar um conjunto de dados específicos. Isso auxilia a controlar a complexidade do modelo conceitual, o que por sua vez torna mais clara as respostas às perguntas iniciais. Ao final dessa etapa, são implementados os grafos de conhecimento referentes à ingestão de dados daquele ciclo.

A fase de **demonstração** apresenta os resultados desenvolvidos utilizando consultas SPARQL no grafo RDF e programas de análise e apresentação de dados, como aplicativos de *Business Intelligence.* Essa visualização preliminar pode mostrar pontos de ajustes necessários no planejamento e nas definições conceituais, o que contribui para o desenvolvimento ágil do modelo.

Na sequência, a **avaliação** busca examinar se o que foi construído nas fases anteriores responde de forma satisfatória as necessidades dos problemas definidos. Além disso, é ponderado os potenciais de integração com dados ingeridos em outros momentos e quais pontos do processo podem ser revistos e ajustados. Por fim, a **divulgação** envolve disponibilizar ao público os artefatos construídos para receber considerações de diferentes pontos de vista e, caso necessário, ajustar o processo ou os produtos.

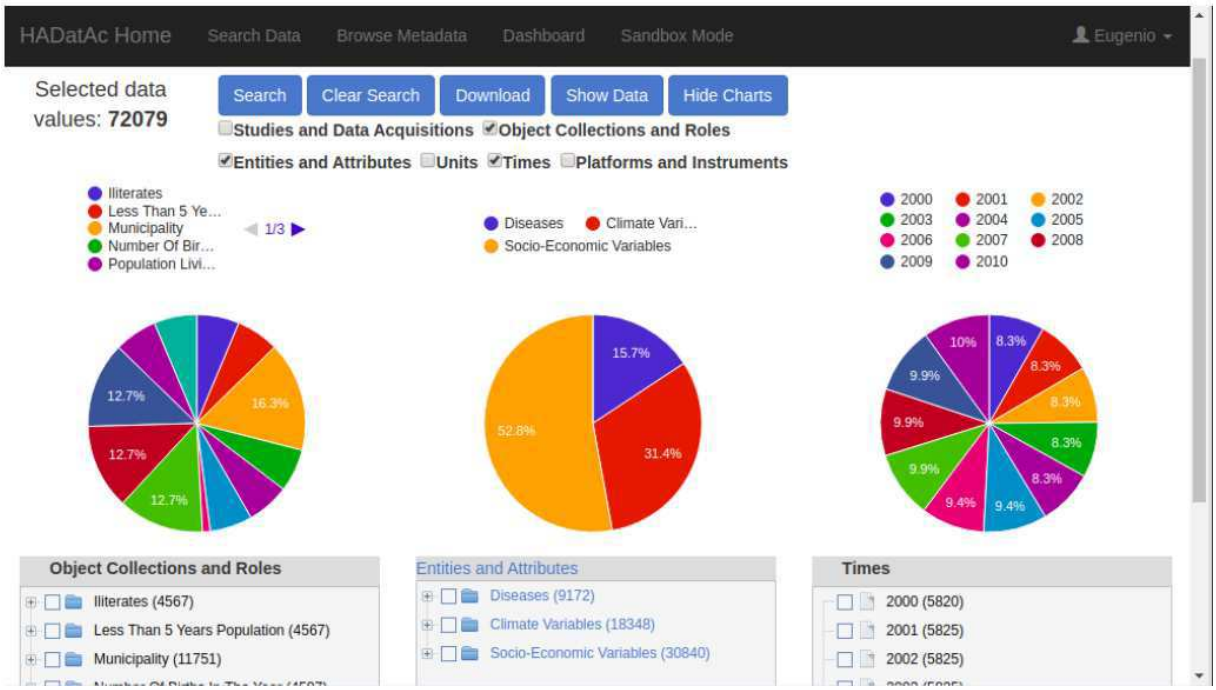
O método ODIN tem como mecanismo chave a ingestão semântica de dados. Esse procedimento envolve conectar os artefatos – dados, metadados e modelos conceituais – de forma a gerar um grafo de conhecimento. Esse processo ocorre durante o ciclo de iteração e o fluxo de trabalho descrito por Gonçalves (2020) explicita os pontos mais relevantes da implementação dessas tarefas.

Conceitualmente, a técnica envolve preparar os arquivos em *.csv*, e a partir do cabeçalho das tabelas montar a ontologia base. Em seguida, inicia-se o processo de mapeamento para os *templates* utilizados no sistema HADatAc[22]: o SDD e o SSD. Então, esses arquivos são inseridos no framework, junto com os arquivos de dados que realiza o processo de ingestão e criação do grafo RDF, apresentando os dados em formato facetado, assim como mostrado na Figura 11. Outros pontos apresentados pelo autor são referentes à: transformação de registros

[21] Por ser uma atividade manual e que depende de informação e validação em grupo.
[22] O *Human-Aware Data Acquisition framework* é uma infraestrtura que permite combinar metadados e conjuntos de dados de forma a enriquecer, utilizando uma coleção de ontologias, as informações disponíveis. Disponível em https://www.hadatac.org/ . Acesso em: 24 out 2022.

nos arquivos *.csv* (para adequar às necessidades de ingestão do HADatAc e normalizar as bases); definição de escopo de conceitos; e orientações gerais sobre o preenchimento do SDD e SSD.

Figura 11 – Tela de busca facetada do HADatAc com os dados ingeridos do estudo de Gonçalves (2020).



Fonte: Gonçalves (2020, p. 69).

4.2 O método Bax e Silva (2020)

O texto "Uso de Dicionário Semântico de Dados na anotação de modelos de dados dimensionais para geração de indicadores de desempenho" (2020) apresenta uma forma de utilizar a técnica do SDD para anotar dados que propiciem a geração de indicadores de performance para organizações. Essa abordagem de anotação semântica, baseada em ontologias, propõe relacionar dados com um modelo conceitual, enriquecendo, assim, aquelas informações.

Os autores utilizam um *script* em linguagem Python, o *sdd2rdf*, que permite interpretar os *templates* do SDD e o conjunto de dados selecionado. Além disso, ele converte a anotação em um grafo de conhecimento RDF (TETHERLESS WORLD, 2022). Esse formato permite interoperar dados uma vez que foram formalizados em uma ontologia, que pode reutilizar conceitos de diferentes fontes.

O processo de anotação semântica descrito se inicia com a seleção de um *dataset* que será trabalhado e a criação de uma ontologia de domínio que descreva os dados. Então, o

Dictionary Mapping (DM) é preenchido com as classes e relações do conjunto de dados utilizados na ontologia. O *Codebook* é preenchido com os dados categóricos da base de dados e mapeados para a ontologia. Na sequência, a *Infosheet* deve conter as informações referentes ao projeto de anotação. Por fim, o *script* reúne esses artefatos e gera um grafo de conhecimento dos dados e metadados anotados, além de representar as possíveis inferências mapeadas no decorrer do processo.

Os autores constatam que utilizar a técnica do SDD permite representar o conhecimento em formato recuperável, interoperável e aberto. Isso favorece positivamente metodologias que utilizam um processo de anotação semântica semelhante.

Na próxima seção relacionaremos as metodologias apresentadas nos subtópicos 4.1 e 4.2 adaptando-as para o contexto específico do presente trabalho.

4.3 Metodologia adaptada

Em síntese, o método ODIN é baseado em ciclos iterativos de ingestão semântica de dados baseados em ontologias. De forma complementar, o método de Bax e Silva (2020) apresenta uma forma de utilizar um *script* que permite utilizar os artefatos do SDD para gerar um grafo de conhecimento. Para a execução deste trabalho, foi elaborado o Quadro 3 que compara o ciclo regulador de Wieringa (2009), o método ODIN e o apresentado por Bax e Silva (2020). Então, para cada fase descrita no ciclo regulador de Wieringa, foram relacionadas etapas descritas pelos autores referenciados e, na última coluna do quadro, são apresentados processos adaptados.

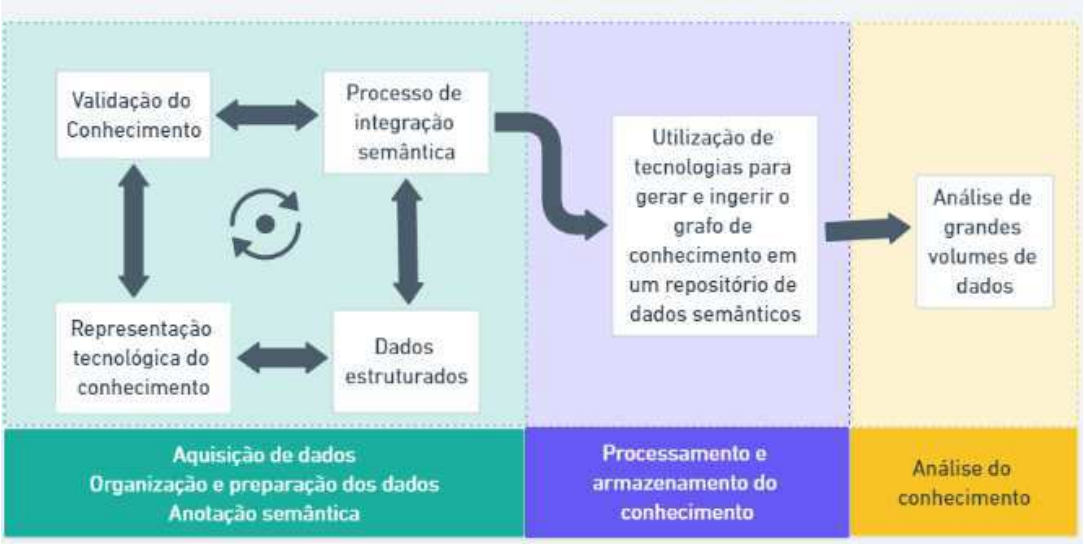
Quadro 3 – Quadro comparativo dos conceitos do ciclo regulador de Wieringa (2009) com a adaptação da metodologia de Gonçalves (2020) e Bax e Silva (2020).

Wieringa (2009)	Gonçalves (2020)	Bax e Silva (2020)	Adaptada
Problema	Identificar problemas	Conjunto de dados relacionados à KPIs	1 - Dados estruturados
Projeto	Definir objetivos	Ontologia de domínio	1 - Representação tecnológica do conhecimento (ontologia)
Validação	Demonstração	-	1 - Validação do conhecimento (grupo focal)
Implementação	Design e Desenvolvimento	*Dictionary Mapping*, *Codebook*, *Infosheet* e *Sdd2rdf*	1 - Processo de integração semântica (SDD) 2 - Processamento e armazenamento do conhecimento (*sdd2rdf*, Virtuoso)
Avaliação	Avaliação; Divulgação	ODBC e Power BI	3 - Análise do conhecimento (ODBC e Power BI)

Fonte: elaborado pela autora.

Como resultado desse quadro, elaborou-se um fluxo a ser seguido para executar a metodologia adaptada dos autores. A Figura 12 apresenta as três fases da metodologia proposta e os processos serão descritos na próxima sessão deste trabalho. A primeira trata os assuntos de aquisição de dados, organização e preparação dos dados e a anotação semântica. A segunda descreve o processo de armazenamento do conhecimento em uma infraestrutura que permite a recuperação das informações ali ingeridas. Por fim, a análise de dados corresponde ao fluxo de avaliação em que é possível visualizar grandes volumes de dados em formatos de fácil compreensão do público.

Figura 12 – Fluxo da metodologia proposta.



Fonte: Elaboração da autora no *software* Whimiscal[23], adaptado de Gonçalves (2020) e Bax; Silva (2021).

A primeira fase da metodologia adaptada é a visualização do problema prático que pode ser identificado em uma base de dados e formalizado em uma ontologia de domínio, validado por atores relevantes e anotados semanticamente. Para formalizar o domínio, foram feitas investigações bibliográficas (com o objetivo de avaliar o estado da arte do domínio, que no caso são as Emendas Parlamentares em Minas Gerais), e de campo (que conta com a observação direta da autora, que exerce atividade profissional em uma das áreas responsáveis pela articulação das Emendas Parlamentares no governo do estado). Cabe ressaltar que essa primeira fase pode se repetir até que o modelo atinja maturidade suficiente para realizar os próximos passos, isso se deve à necessidade de nivelar o conhecimento dos atores em relação às necessidades e especificações do processo.

Então, pode ser selecionada uma tabela com informações sobre um domínio específico relacionado às emendas parlamentares. Antes de iniciar os procedimentos, essa base deve ser normalizada para que as informações sejam processadas corretamente pelos programas. Portanto, deverá ser construída uma ontologia de domínio, baseada em questões de competência relativas aos dados selecionados.

Como forma de validar e aprimorar a construção do conhecimento – codificado em uma ontologia – escolheu-se a metodologia do grupo focal. A justificativa para isso é que a construção de uma ontologia deve ser coletiva e consensual, ou seja, um grupo de pessoas

[23] Disponível em: https://whimsical.com/. Acesso em: 04 de mai 2022.

(também chamados de especialistas de domínio) precisam chegar a um acordo sobre os conceitos e relações daquele assunto.

Para essa fase foram realizados grupos focais com atores chave da Secretaria de Estado de Governo do Estado de Minas Gerais (SEGOV), que têm como competência articular as necessidades do Governo e otimizar o processo de execução das emendas parlamentares com o objetivo de apoiar a articulação política entre as esferas de poder interessadas nos repasses orçamentários (SEGOV, 2019). Este é 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. Para o propósito desta pesquisa, 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 (*ibidem).*

Além disso, os grupos focais auxiliam a 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" (TRAD, 2009, p. 792) que agrega ao processo de construção conjunta o conhecimento. Utilizar esse método de coleta de dados é justificável nesse caso na medida que ela é considerada uma fonte primária de dados para estudos qualitativos. Isso pode ajudar a compreender e examinar o contexto em análise pela perspectiva das pessoas que vivem o processo diariamente.

Destaca-se que a quantidade de encontros de um grupo focal não é delimitada. Essa flexibilidade é útil para adaptar o roteiro das reuniões de forma a nivelar o conhecimento do domínio e das técnicas aplicadas. Para que um grupo focal seja realizado com sucesso, é necessário que o objetivo esteja claro e em consonância com a pesquisa. Ou seja, ele deve servir para um propósito específico e que esteja de acordo com os temas e questões que precisam ser respondidas pelo pesquisador(TRAD, 2009) .

Por essa razão, é importante delimitar a quantidade e perfil dos participantes, o meio a ser utilizado, o tempo dispendido por reunião, a quantidade de encontros e o roteiro a ser seguido. A metodologia (com início, meio e fim) deve estar clara tanto para o pesquisador – que conduzirá os encontros – quanto para os participantes (GONDIM, 2002). Para atingir esse objetivo, o roteiro das entrevistas deve ser curto e básico, com perguntas simples e gerais para que o entrevistado possa expor suas percepções a respeito dos conceitos.

As quatro reuniões foram gravadas com o consentimento dos participantes, que tiveram suas identidades preservadas para os fins dessa pesquisa. Isso é garantido uma vez que

todas as conversas gravadas ficarão em posse da autora desse trabalho e as transcrições não serão publicizadas sem garantir a anonimização dos participantes – se necessário, serão identificados por nomes fictícios. Então, pode-se preencher os modelos de documentos do SDD.

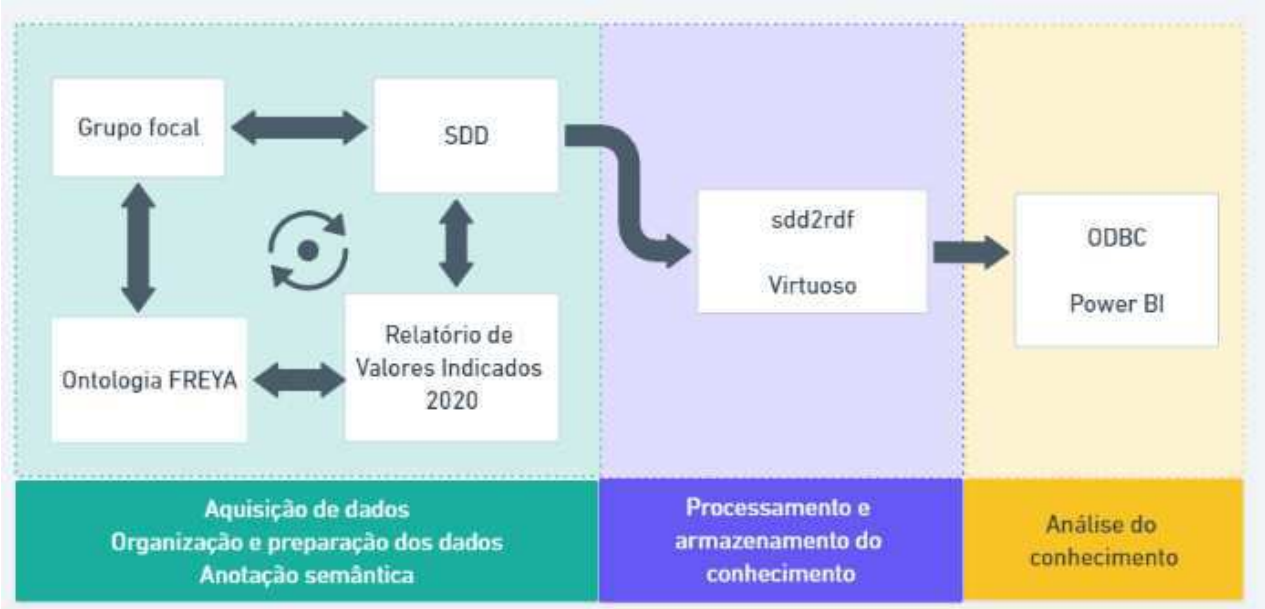
A segunda fase da metodologia, "Processamento e armazenamento do conhecimento", transforma, por meio de *scripts,* os dados e artefatos do SDD em um grafo de conhecimento que poderá ser persistido em uma infraestrutura que permita a consulta e recuperação das informações. Dessa forma, os modelos serão armazenados e ingeridos em um repositório de dados semânticos.

A última fase, "análise do conhecimento", deve buscar criar uma ponte entre o repositório definido e um programa que permita a extração dos dados para o *software* de *Business Analytics*. Essa avaliação do problema permite gerar painéis de acompanhamento para visualizar aquele conjunto de dados anotados. Utilizando essa visualização, pode ser possível responder às perguntas de competências definidas e que guiaram a construção dos artefatos da metodologia. O processo de anotação semântica traz confiabilidade para os dados e pode ser criado por atores envolvidos diretamente com o domínio. Além disso, a formação desses artefatos é uma forma de avaliar o conhecimento construído ao longo da aplicação da metodologia.

5 APLICAÇÃO EM UM ESTUDO DE CASO: EMENDAS PARLAMENTARES IMPOSITIVAS DO ESTADO DE MINAS GERAIS

Para aplicar a metodologia foi definido o uso da ontologia desenvolvida bem como os requisitos elementares para aquele domínio do conhecimento. Essa fase de pré-modelagem conceitual é útil para evitar representar variáveis que não serão utilizadas, o que auxilia a reduzir a complexidade do modelo, que 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 (SEQUEDA; LASSILA, 2021). A Figura 13 apresenta quais técnicas foram utilizadas para executar este trabalho.

Figura 13 - Materiais e métodos utilizados para a execução da pesquisa.



Fonte: elaborado pela autora.

A execução do processo explicado na metodologia foi realizada de forma manual utilizando um conjunto de dados específico da tabela de "Valores Indicados", na planilha de "Relatório TCE MG – 2020 Retificado"[24]. Portanto, a construção da ontologia e os grupos focais se basearam em um recorte desses dados. Esse procedimento foi necessário para delimitar o domínio de avaliação. Dessa forma, a metodologia adaptada neste trabalho foi executada completamente uma vez, com o objetivo de fazer uma prova de conceito. A organização deste capítulo está organizada em três tópicos, um para cada fase do método e em subtópicos que descrevem cada um dos passos.

[24] Disponível em https://www.emendas.mg.gov.br/dados-de-emendas-2020/ . Acesso em 28 jun 2022.

5.1 Aquisição de dados, organização e preparação, e anotação semântica

Para iniciar a metodologia adaptada neste trabalho, selecionou-se um conjunto de dados relacionais que dizem respeito ao tema tratado. A partir dessa escolha, é possível iniciar a primeira versão da ontologia de domínio. Então, iniciou-se o processo de validação do conhecimento que consiste em compreender a extensão das análises necessárias para aquele conjunto de dados. Por fim, será possível executar a primeira versão dos artefatos de integração semântica.

5.1.1 Dados estruturados

Para iniciar a construção, foi selecionada uma base de dados disponibilizada no sítio eletrônico[25] no menu de "Execução de Emendas 2020" que apresenta um relatório contendo o *status* e a execução das emendas parlamentares daquele ano. O arquivo "TCE-MG – 2020 – Retificado" é acompanhado de diversas tabelas com as informações a respeito do ciclo das emendas. Para os fins deste estudo, somente a planilha de "Valores Utilizados" foi utilizada para dar início à construção da ontologia.

Então, foram selecionadas algumas colunas da planilha (Tabela 4) para delimitar o volume de informação a ser anotada. Depois, os cabeçalhos e conteúdo foram normalizados (

Quadro 4) para retirar caracteres especiais que poderiam causar erros na execução das fases seguintes.

Tabela 4 – Quatro linhas iniciais 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: Portal de Emendas da Segov (2021a). Elaborado pela autora.

[25] Disponível em https://www.emendas.mg.gov.br/dados-de-emendas-2020/ . Acesso em 28 jun 2022.

Quadro 4 - De-para da tabela de Valores Indicados.

DE	PARA
Responsável (Nome)	ResponsavelNome
Número Indicação	NumeroIndicacao
Tipo de Indicação	TipoIndicacao
Descrição Município	DescricaoMunicipio
Nome Convenente Beneficiado	NomeConvenenteBeneficiado
Nome Grupo de Despesa	NomeGrupoDespesa
Valor Indicado	ValorIndicado

Fonte: Emendas 2020 (SEGOV, 2021a), elaborado pela autora.

O tratamento realizado envolveu retirar os espaços entre as palavras e os caracteres especiais do cabeçalho e das linhas da tabela. Como forma de representar a execução desse passo, o Quadro 4 apresenta o resultado desse procedimento. Foram excluídos espaços, parênteses, cedilhas, acentos e artigos, o que auxilia nos momentos seguintes. Após esse passo, foi iniciada a construção da primeira versão da ontologia.

5.1.2 Ontologia FREYA

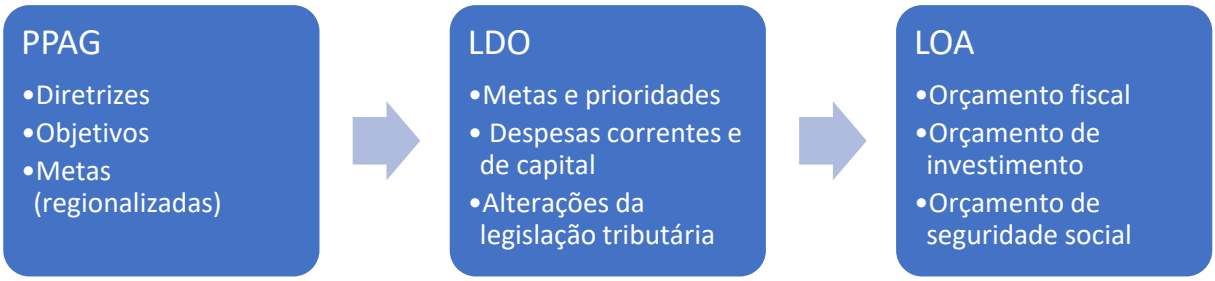
Ao analisar o contexto da tabela selecionada e organizar os dados, iniciou-se processo de criar a primeira versão da ontologia. De acordo com a literatura (GRUBER, 1995; NOY; MCGUINNESS, 2011; RECTOR et al., 2019), uma ontologia é um modelo conceitual que pode ser utilizado para representar a realidade. Para isso, foi necessária uma pesquisa bibliográfica prévia para auxiliar a estruturar os conceitos e delimitar a relevância do assunto para a Administração Pública. Para o caso específico de Minas Gerais é importante compreender como o ciclo das indicações ocorre, desde o momento em que a emenda parlamentar é definida até quando é indicado um valor. A seguir, é apresentado na Figura 14 o fluxo que ocorre desde o planejamento do orçamento mineiro até as indicações no Sistema de Gestão de Convênios, Portarias e Contratos do Estado de Minas Gerais – Módulo Saída (SIGCON-SAÍDA)[26].

O processo orçamentário no Brasil, previsto na Constituição Mineira (MINAS GERAIS, 1989), reúne a redação e aprovação de três leis fundamentais para o estudo em tela: o Plano Plurianual de Ação Governamental (PPAG), a Lei de Diretrizes Orçamentárias (LDO) e a Lei Orçamentária Anual (LOA). Essas legislações são compostas por marcos e itens

[26] 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, 2021a).

específicos que iniciam de maneira ampla no PPAG e vão ficando mais específicas (em detalhes de execução orçamentária e financeira) até a LOA. Existem extensos detalhes sobre esse processo, mas, para cumprir os objetivos deste trabalho, a Figura 14 apresenta os tópicos mais importantes de cada um desses instrumentos.

Figura 14 – Representação simplificada do processo orçamentário mineiro.



Fonte: Constituição do Estado de Minas Gerais, seção II "Dos Orçamentos", (1989). Elaborado pela autora.

O último instrumento é produzido pelo poder executivo (então chamado de projeto de LOA ou pLOA) e então encaminhado para o poder Legislativo, para apreciação e discussão dos parlamentares. Nesse momento, o orçamento proposto pela equipe do Executivo pode sofrer mudanças, ou emendas, pela casa legislativa. Durante esse período, os membros da Assembleia Legislativa de Minas Gerais podem alterar algumas definições orçamentárias para que os deputados possam participar, por meio de indicações, de projetos e programas no Estado. Como exemplo, a Figura 15 apresenta como essas emendas são apresentadas no Anexo V, da LOA 2020.

Figura 15 - Demonstração da maneira como as emendas são representadas na LOA mineira.

INCISO: 342 (Emenda nº 390)
1 491 04 122 024 2 007 0001 3 3 99 10 8 0 A 500.000,00
1 991 99 999 999 9 999 0001 9 9 99 10 1 0 D 500.000,00
Unidade Orçamentária Beneficiada: Secretaria de Estado de Governo
Objeto do gasto: Execução do Programa de Apoio ao Desenvolvimento Municipal - Padem (despesas correntes)
Dedução: Reserva de Contingência
Autor: Deputado Alencar da Silveira Jr.

Fonte: Lei Orçamentária Anual (MINAS GERAIS, 2020).

Esse padrão de representação possui informações essenciais para a movimentação das dotações e a forma que esses recursos serão executados de acordo com a denominação. Quando o projeto de LOA é enviado do executivo para a Casa Legislativa, é separada uma dotação orçamentária específica para que os parlamentares possam fazer suas modificações: a

reserva de contingência (representada na Figura 15 pela dotação em que é feita a dedução). O parlamentar e sua equipe de assessores definem junto às secretarias de estado quais os objetos de gasto e os valores a serem destinados nas futuras indicações. Portanto, ao verificar a LOA aprovada, é possível aferir que a emenda de número 390, possui um ou mais incisos – no caso da Figura 15 o de nº342 – que movimentou da dotação de reserva de contingência para a de ação "Execução do Programa de Apoio ao Desenvolvimento Municipal – Padem", no grupo de despesas correntes, e para a unidade executora SEGOV. Dentro dos conceitos de orçamento público, essa memória de cálculo é importante para localizar a origem daqueles recursos e a forma de utilização.

As dotações orçamentárias possuem vários códigos e definições para indicar como aquele dinheiro público será executado. Todas essas informações servem para que o caminho da dotação seja mais transparente, passível de tomada de contas e que as definições em lei sejam cumpridas. Por exemplo, quando se trata de execuções de despesas correntes, não se pode gastar essa verba na construção ou reforma de alguma edificação. O valor dessa dotação será utilizado por meio de algum instrumento jurídico de saída e repassará o recurso Concedente para o Convenente[27], sempre levando em consideração as regras dispostas na Lei de Diretrizes Orçamentárias (LDO).

Essa forma de transferência voluntária possui caráter discricionário, pautados na conveniência e disponibilidades das partes envolvidas no processo, dentro dos limites permitidos pela lei. Portanto, o instrumento a ser celebrado permite que o ente estadual possa celebrar algum tipo de relação convenial com outro, sempre observando a anuência das duas partes. No cenário brasileiro, é comum que existam associações entre a União e Municípios, em que o primeiro - o Concedente - envia algum tipo de recurso (podendo ser financeiro ou não financeiro) para o segundo - o convenente - que normalmente confirma seu compromisso enviando uma contrapartida. Em Minas Gerais, esses dispositivos vêm como forma de descentralizar o orçamento de um órgão ou entidade pública parceira para entes federados ou pessoas jurídicas a ele vinculadas, Organizações da Sociedade Civil (OSCs), Fundos Municipais ou Serviços Sociais Autônomos (SSA).

O Concedente ou Órgão Estadual ou Entidade Parceira (OEEP)[28] apresentado é quem irá receber os recursos daquela indicação específica. De acordo com a Resolução SEGOV nº001/2021 são considerados como

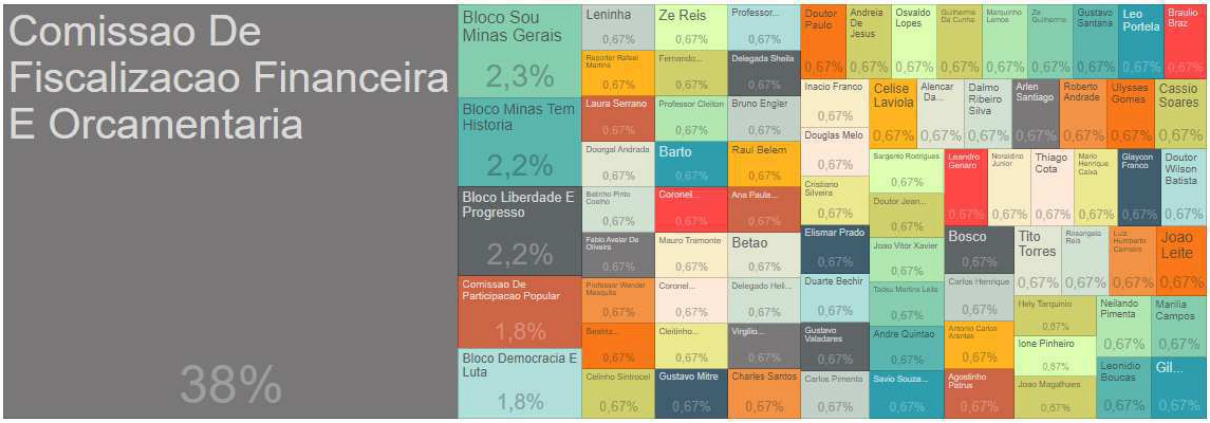
[27] Essas definições podem ser consultadas na página do GitHub dessa pesquisa https://github.com/marci-pires/EmendasParlamentaresMG

[28] No momento anterior ao SIGCON-SAÍDA, eles são chamados de Beneficiários.

> órgão ou entidade da Administração Pública do Poder Executivo estadual ou fundo municipal de saúde, caixa escolar da rede pública estadual, município, União, Estado ou entidade da administração pública indireta dos entes federados ou organização da sociedade civil – OSC – com cadastro completo no Cagec, **indicados por autores de emendas parlamentares individuais, de blocos ou de bancadas para fins de recebimento de recursos do orçamento fiscal do Estado de Minas Gerais** (art. 2o, V SEGOV, 2021b).

Para o ano de 2020 foram disponibilizados R$984.795.242,00, que foram divididos para parlamentares (cada um dos integrantes tem R$6.635.204,00 destinados para indicações individuais), blocos e bancadas, e comissões. Essas emendas são uma ferramenta para que o Poder Legislativo participe ativamente do orçamento anual e que possam por meio delas, aprimorar a alocação dos recursos públicos de acordo com suas preferências políticas e regionais. Essa distribuição (Figura 16) pode ser averiguada com mais detalhes no Portal da Transparência de Minas Gerais.

Figura 16 -Distribuição percentual do orçamento destinado a emendas parlamentares em 2020 por autores de cada emenda.



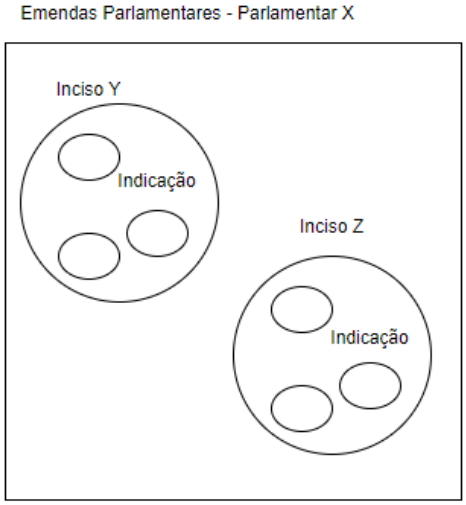
Fonte: Portal da Transparência de Minas Gerais, Emendas Orçamentárias[29].

As emendas parlamentares são divididas em três tipos principais: individual, bloco ou bancada e comissão. Os dois primeiros são de execução orçamentária obrigatória (impositivas) e o último não. Portanto, o total que é apresentado e o percentual no gráfico é equivalente ao valor total das emendas parlamentares. Os recursos destinados para as comissões não têm obrigação de ser executado no ano orçamentário corrente, entretanto é um valor expressivo (R$372.492.790,00), que é distribuído por definições políticas. Portanto, para este trabalho foi escolhido analisar as emendas impositivas pela obrigatoriedade de execução daquele orçamento.

[29] Para consultar o gráfico acesse: https://www.transparencia.mg.gov.br/planejamento-e-resultados/proposta-lei-orcamentaria/emenda-orcamentaria/emenda-deputados/2020/

A composição de uma emenda parlamentar pode ser representada como na Figura 17, em que uma emenda parlamentar pode ter mais de um inciso, e cada inciso pode ter uma ou mais indicações (realizadas no SIGCON-SAIDA).

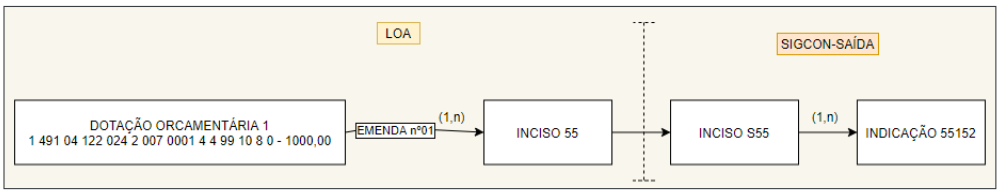
Figura 17 - Composição das emendas parlamentares e a relação delas com as indicações.



Fonte: Elaborado pela autora.

Para utilizar esse saldo proveniente de um inciso, é necessário que o parlamentar realize uma indicação formalizada. Essa ação é realizada no sistema eletrônico, gerido pela SEGOV, o SIGCON-SAÍDA - que subsidia o controle da execução orçamentária e financeira das emendas parlamentares, inclusive as de execução obrigatória, também chamadas de impositivas (MINAS GERAIS, 2021b). Portanto, quando os dados da LOA são carregados no sistema, cada um dos incisos recebe um identificador S, que identifica quais deles vieram diretamente da primeira carga da lei no sistema – quando existe remanejamento[30](R), saneamento ou remanejamento constitucional (P). Esse processo pode ser esclarecido ao observar o fluxo descrito na Figura 18.

Figura 18 – Sistematização do processo de *upload* dos dados da LOA para o SIGCON-SAÍDA.

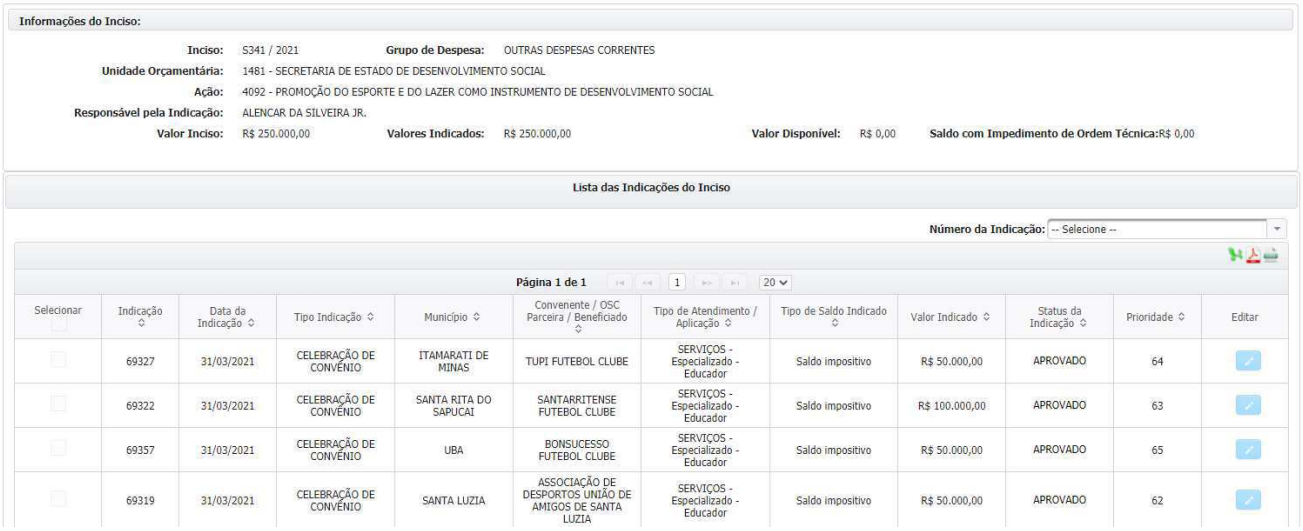


Fonte: Elaborada pela autora.

[30] No remanejamento podem ser alterados a ação, unidade executora e grupo de despesa.

Então, uma emenda pode ter um ou mais incisos que podem ter diversas indicações. Mas, uma indicação é proveniente de somente um inciso, que vem de uma emenda e que tem sua origem em uma dotação orçamentária. O SIGCON-SAÍDA simplifica esse processo e permite que as indicações sejam realizadas de acordo com diferentes formatos de aplicação – previamente acordados entre a área de normatização, secretarias de estado e parlamentares – para identificar qual tipo de instrumento jurídico será formalizado e qual convenente será beneficiado com o repasse dos recursos. A Figura 19 apresenta como as informações são organizadas no sistema e como o inciso é dividido em diferentes indicações, com os dados de cada uma delas.

Figura 19 - Informações do inciso no sistema SIGCON-SAÍDA.



Informações do Inciso:

Inciso: S341 / 2021 **Grupo de Despesa:** OUTRAS DESPESAS CORRENTES
Unidade Orçamentária: 1481 - SECRETARIA DE ESTADO DE DESENVOLVIMENTO SOCIAL
Ação: 4092 - PROMOÇÃO DO ESPORTE E DO LAZER COMO INSTRUMENTO DE DESENVOLVIMENTO SOCIAL
Responsável pela Indicação: ALENCAR DA SILVEIRA JR.
Valor Inciso: R$ 250.000,00 **Valores Indicados:** R$ 250.000,00 **Valor Disponível:** R$ 0,00 **Saldo com Impedimento de Ordem Técnica:** R$ 0,00

Lista das Indicações do Inciso

Número da Indicação: -- Selecione --

Página 1 de 1 1 20

Selecionar	Indicação	Data da Indicação	Tipo Indicação	Município	Convenente / OSC Parceira / Beneficiado	Tipo de Atendimento / Aplicação	Tipo de Saldo Indicado	Valor Indicado	Status da Indicação	Prioridade	Editar
	69327	31/03/2021	CELEBRAÇÃO DE CONVÊNIO	ITAMARATI DE MINAS	TUPI FUTEBOL CLUBE	SERVIÇOS - Especializado - Educador	Saldo impositivo	R$ 50.000,00	APROVADO	64	
	69322	31/03/2021	CELEBRAÇÃO DE CONVÊNIO	SANTA RITA DO SAPUCAI	SANTARRITENSE FUTEBOL CLUBE	SERVIÇOS - Especializado - Educador	Saldo impositivo	R$ 100.000,00	APROVADO	63	
	69357	31/03/2021	CELEBRAÇÃO DE CONVÊNIO	UBA	BONSUCESSO FUTEBOL CLUBE	SERVIÇOS - Especializado - Educador	Saldo impositivo	R$ 50.000,00	APROVADO	65	
	69319	31/03/2021	CELEBRAÇÃO DE CONVÊNIO	SANTA LUZIA	ASSOCIAÇÃO DE DESPORTOS UNIÃO DE AMIGOS DE SANTA LUZIA	SERVIÇOS - Especializado - Educador	Saldo impositivo	R$ 50.000,00	APROVADO	62	

Fonte: sistema SIGCON-SAÍDA, acesso 03/05/2021.

Avaliando as informações disponibilizadas no sistema e na planilha de "Valores Indicados", é possível iniciar a construção da ontologia que será apresentada ao grupo focal. Para guiar a primeira modelagem define-se a seguinte pergunta de competência: *Qual responsável (`:ResponsavelNome`) realizou (`:fazIndicacao`) mais indicações (`:NumeroIndicacao`) em 2020?* A tripla formada representa a relação entre responsável e número da indicação, que representa uma chave primária da tabela. Para desenvolver a ontologia, denominada FREYA, é necessário descrever os relacionamentos entre as classes, propriedades de dados e objetos. Para isso, foi construído o Quadro 5, que organiza os dados das triplas formadas.

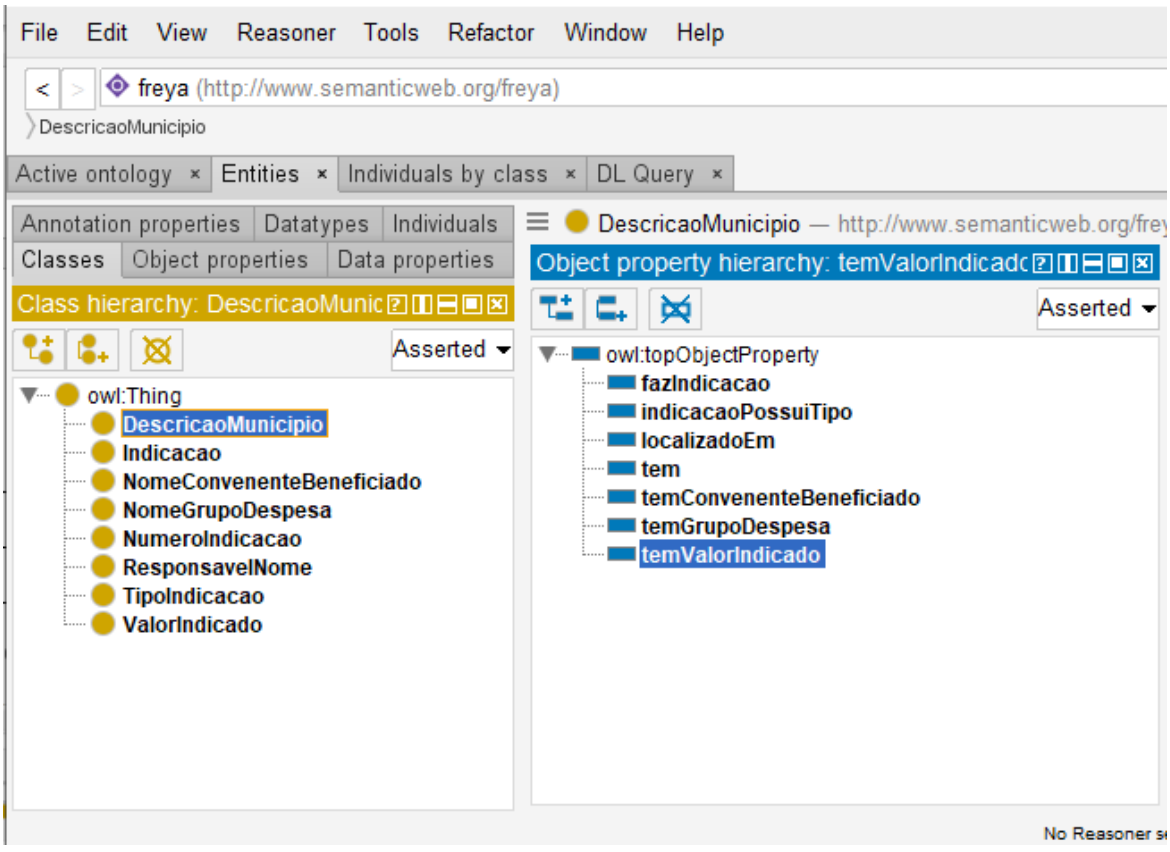
Quadro 5 – Organização das triplas da ontologia.

#	Sujeito	Predicado	Objeto
1	:ResponsavelNome	:fazIndicacao	:Indicacao
2	:Indicacao	:tem	:NumeroIndicacao
3	:Indicacao	:indicacaoPossuiTipo	:TipoIndicacao
4	:NomeConvenenteBeneficiado	:localizadoEm	:DescricaoMunicipio
5	:Indicacao	:temGrupoDespesa	:NomeGrupoDespesa
6	:Indicacao	:temValorIndicado	:ValorIndicado

Fonte: elaborada pela autora.

Depois dessa primeira organização, essas relações foram repassadas para o programa Protégé[31] que oferece suporte para criar, visualizar e modelar ontologias. A Figura 20 apresenta como o programa organiza as classes e as propriedades do objeto. Preliminarmente, os cabeçalhos da tabela são tratados como classes. Caso necessário, é possível refinar essas relações criando diferentes propriedades para adequar a modelagem da ontologia.

Figura 20 – Ontologia FREYA modelada no Protégé.



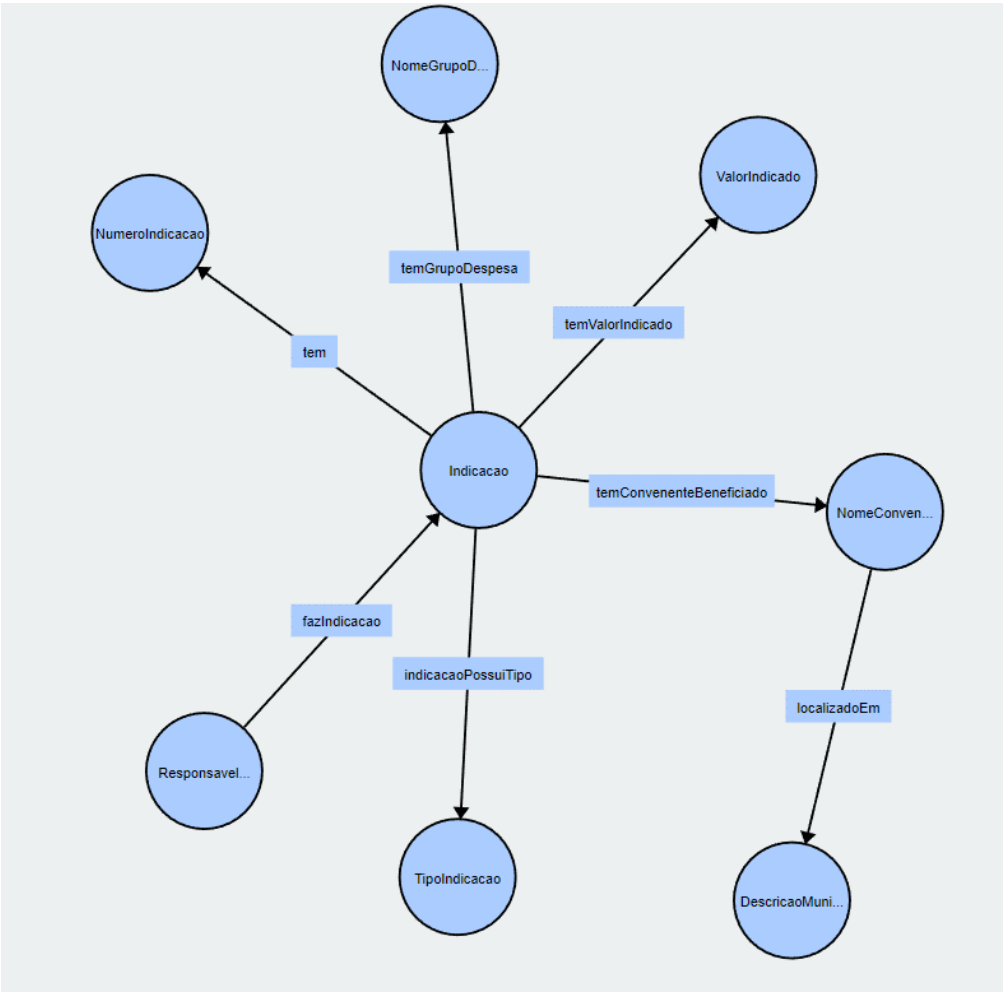
Fonte: elaborado pela autora.

Após a construção do artefato, foi gerado um arquivo *.owl* que foi inserido em um visualizador de ontologias. Esses sistemas permitem renderizar as relações definidas de uma

[31] Disponível em: http://protege.stanford.edu/ . Disponível em 20 abr 2022.

ontologia em formato de nós, que representam as triplas formadas. A Figura 21 apresenta como a aplicação "*Web-based Visualization of Ontologies"* (*Web*VOWL)[32] que é um visualizador de ontologias baseadas que pode ser utilizado *online*, gera uma versão da ontologia elaborada.

Figura 21 – Representação da primeira versão da ontologia FREYA.



Fonte: elaborada pela autora.

Na Figura 21 é possível verificar as relações entre as classes. Por exemplo, é possível aferir a tripla "`Indicacao-> temConvenenteBeneficiado -> NomeConvenente`" o que denota que uma indicação, realizada por um responsável, é destinada a um convenente específico. De forma complementar, o Convenente é localizado em um município que, para este caso, está descrito na coluna de `DescricaoMunicipio`. Com esse protótipo definido, foi possível dar início ao grupo focal que validou essas relações e incrementou as informações levantadas.

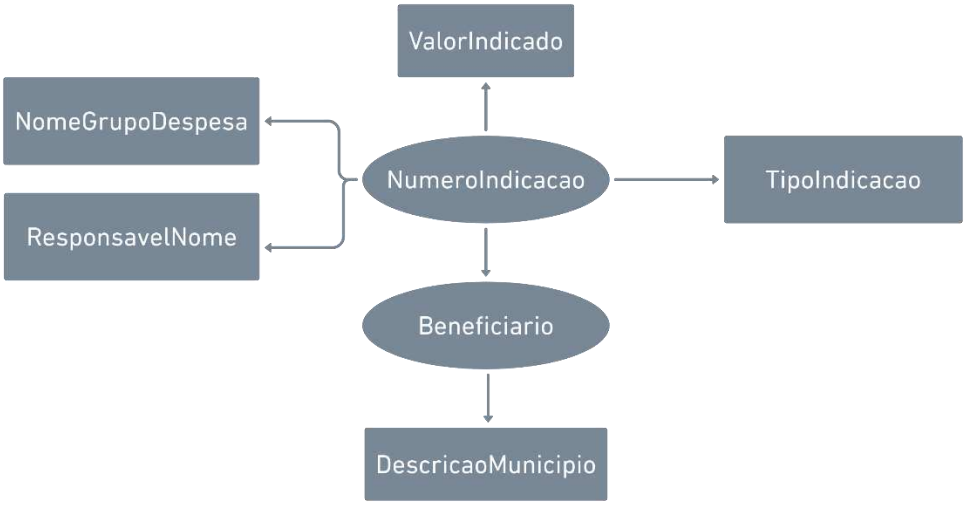
[32] Disponível em: http://vowl.visualdata*Web*.org/*Web*vowl.html . Acessado em 22 jun 2022.

5.1.3 Execução do grupo focal

Os encontros foram mediados pela autora que atuou, de acordo com a descrição da metodologia de Gonçalves (2020), como cientista do conhecimento. Em quatro reuniões, foram discutidos os conceitos e as questões de competência que foram utilizadas para construir o SDD. Foi explicado ao grupo, composto inicialmente por seis pessoas, o processo de construção de uma ontologia, as melhores práticas e o contexto do trabalho a ser executado posteriormente. Portanto, a execução do grupo focal exploratório foi realizada de forma a validar as informações de cada uma das colunas e compreender como os conceitos se relacionavam, de acordo com a experiência dos participantes com o tema.

Essa fase da metodologia foi essencial para validar a importância da discussão sobre a anotação dos dados em um contexto complexo. Além disso, em razão da pouca experiência dos participantes com sistemas e temas afins, foi necessário preparar uma visão simplificada da ontologia. A Figura 22 apresenta um modelo que foi apresentado no primeiro encontro e serviu para exemplificar como relacionar os cabeçalhos da tabela selecionada, ainda que de forma direta e sem propriedades que qualificam essa relação. Então, foi sugerida a pergunta de competência que norteou a criação da primeira versão da ontologia: Qual responsável (`:ResponsavelNome`) realizou (`:fazIndicacao`) mais indicações (`:NumeroIndicacao`) em 2020?

Figura 22 – Modelagem utilizada no grupo focal para exemplificar a ontologia.



Fonte: elaborado pela autora.

Então, para o primeiro encontro buscou-se estabelecer uma relação de familiaridade entre os participantes e os conceitos necessários para formular dúvidas e ideias sobre o tema de discussão. Para incentivar a participação, utilizou-se uma folha de tamanho A1 para que todos

pudessem contribuir com o desenho das relações, simulando a formação de uma ontologia. As figuras estão disponíveis no "APÊNDICE I - ROTEIRO E IMAGENS DO GRUPO FOCAL". Isso foi interessante pois reduziu uma barreira de insegurança com a discussão, que trata de métodos complexos. A Figura 35 (Apêndice I), apresenta o desenho obtido ao final do encontro. Cabe ressaltar que foram extrapolados alguns dos conceitos presentes na tabela escolhida para esta prova de conceito. Entretanto, isso não foi prejudicial para o andamento dos encontros, uma vez que ao compreender o potencial dos artefatos a serem construídos, os participantes se tornaram mais participativos e inclusive sugeriram outras perguntas de competência que envolvem diferentes restrições do domínio, mas que seriam de complexidade elevada e que não poderiam estar no escopo desta pesquisa.

O segundo encontro, que ocorreu na semana seguinte, focou em compreender como a indicação centraliza os conceitos do conjunto de dados. Buscou-se avaliar e acordar definições para algumas classes e, reduzir o escopo do primeiro encontro. Nesse ponto, os participantes, ao construírem a Figura 36 (Apêndice I), estabeleceram propriedades mais concisas e representativas da realidade. Esse refinamento é crucial e atua como uma forma de refinar, de forma coletiva, as ideias daquele assunto. Isso está em consonância com a literatura (GONDIM, 2002; TRAD, 2009) que apresenta a alternativa do grupo focal como uma forma de construir conhecimento.

O terceiro encontro contou com a presença de atores que executam tarefas relacionadas à gestão das emendas parlamentares no governo de Minas Gerais. Mesmo que a temática seja de entendimento do grupo como um todo, esses participantes trouxeram outra visão e ampliaram algumas definições e conceitos. De forma geral, foi incluído na Figura 37 (Apêndice I), termos relacionados à execução orçamentária e financeira (empenho, liquidação e pagamento) dos recursos de emendas parlamentares e como isso influencia nos *status* das indicações. Uma das perguntas sugeridas pelo grupo foi "Em qual fase a indicação está?" o que exigiria do modelo e da tabela proposta mais classes e relações a serem definidas.

O último encontro teve como objetivo validar a pergunta de competência inicial e a sua relevância, como prova de conceito. O grupo se mostrou interessado no resultado dos mapas construídos e como aquilo poderia se tornar um artefato tecnológico. Foi explicado que as ideias discutidas serviriam como validação do conhecimento sobre o assunto discutido e que isso tornaria o painel de acompanhamento mais confiável e escalável. Então, os participantes concordaram que seria um ponto de partida factível para seguir com a metodologia e que resolveria um problema prático e recorrente. Isso está em consonância com a DSR (BAX, 2013; WIERINGA, 2009) que busca solucionar problemas teóricos com soluções práticas e vice-

versa; e com as diretrizes da Metodologia Ágil que busca adaptar o objetivo de um projeto às necessidades mais latentes de um grupo. Após o consenso apresentado pelo grupo, o próximo passo é preencher as planilhas do SDD para a integração semântica.

5.1.4 Processo de integração semântica

Assim como apresentado, o *Semantic Data Dictionary* (SDD) é um conjunto de ferramentas que são utilizadas para anotar dados semanticamente. Essa abordagem utiliza-se de modelos de metadados pré-definidos baseados em ontologias. Essa técnica recomenda a participação de atores especialistas de domínio e engenheiros do conhecimento no processo de anotação (RASHID et al., 2020). Este trabalho optou por separar o momento da validação do conhecimento, em que há interação com os especialistas, do preenchimento dos *templates* de metadados. Isso se deve à dificuldade de aplicação da técnica e não seria produtivo ou rápido de realizar. Portanto, a autora (por se encaixar nos dois grupos) preencheu os arquivos do SDD com o auxílio do grupo de pesquisa "Semantic eScience"[33]. Cabe ressaltar que todos os arquivos utilizados na implementação do SDD estão em formato *.csv* e podem ser encontrados no *GitHub*[34].

Enquanto os encontros do grupo focal aconteciam no mês de novembro de 2021, foram preparados os arquivos para a execução do SDD. Após a elaboração da primeira versão da ontologia, o primeiro documento preenchido foi o *Infosheet*. Essa planilha de informações, apresentada no Quadro 6 contém os metadados da pesquisa e os vocabulários utilizados no SDD. De maneira geral, ele contém a localização das outras planilhas (*Dictionary Mapping*, *Codebook*, *Timeline* e *Properties*) e outras informações - tais como língua, versionamento, formato dos arquivos - que servem como um arquivo de configuração que une o SDD.

Na sequência, é preenchido o *Dictionary Mapping* (DM) que é composto das colunas da tabela de dados relacionais selecionada (também chamada de entradas explícitas) e os elementos para anotação semântica daquele conjunto. A estrutura do DM permite, de forma singular, anotar as entidades implícitas daqueles conceitos apresentados na tabela de "Valores Indicados".

A "*Column*" é preenchida com os cabeçalhos da tabela relacionada e deve seguir com as entidades implícitas encontradas no processo de enriquecimento semântico. A coluna "*Attribute*" deve descrever uma característica da coluna do *dataset* e deve ser preenchida com

[33] Disponível em: http://dgp.cnpq.br/dgp/espelhogrupo/7918248711733865. Acesso em: 22 mai 2022.
[34] Disponível em: https://github.com/tetherless-world/SemanticDataDictionary . Acesso em: 22 mai 2022.

uma propriedade de uma ontologia apropriada a fim de atribuir uma entrada semanticamente significativa. Através da coluna "*attributeOf*" é colocada uma anotação do conjunto de conceitos do domínio referente ao atributo do conjunto de dados selecionado. Cada uma das entradas da coluna anterior deve ser mapeada para a ontologia e relacionada na coluna "*Entity*".

Quadro 6 - 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 técnica SDD
Date Created	18/11/2021
Creators	Marcela Pires
Contributors	Evaldo da Silva e Marcello Bax
Publisher	Marcela Pires
Date of Issue	12/10/2021
Identifier	freya
Keywords	Emendas Parlamentares; atividade legislativa
Language	PT-BR
Version	2.0
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 pela autora.

No caso, a Tabela 5 apresenta o DM elaborado. A primeira descrição o *NumeroIndicacao* que é a chave primária da indicação e que é a Indicação em si. Por isso, é um conceito implícito naquele conjunto de dados e que é incluído como *??indicacao*. Então, é criada uma classe na ontologia *freya:Indicacao* na coluna *Entity*. Dessa forma, ocorre o enriquecimento semântico daquelas relações antes implícitas, transformando os dados anotados de forma a gerar mais conhecimento. A sequência lógica desse processo se repetiu até que fossem contemplados os conceitos mais relevantes para aquele conjunto de dados.

Tabela 5 - *Dictionary Mapping* para o domínio de Emendas Parlamentares Impositivas.

Column	Attribute	attributeOf	Entity	Relation	inRelationTo
NumeroIndicacao	hasco:originalId	??indicacao			
ResponsavelNome	freya:Indicacao	??responsavel			
TipoIndicacao	freya:Indicacao	??execucaodaindicacao			
DescricaoMunicipio	freya:Municipio	??municipiobeneficiario			
NomeGrupoDespesa	freya:GrupoDespesa	??grupodespesa			
NomeConvenenteBeneficiado	freya:Beneficiario	??nomedobeneficiario			
ValorIndicado	freya:Indicacao	??valorindicacao			
??indicacao			freya:Indicacao		
??responsavel			freya:ResponsavelNome		
??indicacao			freya:Indicacao	freya:indicacaoPossui	freya:ResponsavelNome
??execucaodaindicacao			freya:TipoIndicacao		
??indicacao			freya:Indicacao	freya:indicacaoPossuiTipo	freya:TipoIndicacao
??grupodespesa			freya:GrupoDespesa		
??indicacao			freya:Indicacao	freya:indicacoPossuiGrupoDespesa	freya:GrupoDespesa
??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 pela autora.

O *Codebook* descreve os dados categoriais do conjunto de dados e os mapeia para a ontologia FREYA. A tabela "Valores Indicados" possui duas colunas que podem ser categorizadas: Tipo de Indicação (TipoIndicacao) e Grupo de Despesa (NomeGrupoDespesa). Cada um dos códigos da Tabela 6 na coluna "*Code*" representam as categorias de dados da tabela e a "*Class*" apresenta quais classes foram criadas para a ontologia.

Tabela 6 – *Codebook* para o domínio de Emendas Parlamentares Impositivas.

Column	Code	Class
TipoIndicacao	APLICACAO DIRETA DOACAO DE BENS	:AplicacaoDiretaDoacaoBens
TipoIndicacao	CELEBRACAO DE CONVENIO	:Convenio
TipoIndicacao	EXECUCAO DIRETA	:ExecucaoDireta
TipoIndicacao	EXECUCAO DIRETA CAIXA ESCOLAR	:ExecuçãoDiretaCaixaEscolar
TipoIndicacao	OUTROS INSTRUMENTOS	:OutrosInstrumentos
TipoIndicacao	RESOLUCAO	:Resolucao
TipoIndicacao	TRANSFERENCIA ESPECIAL	:TransferenciaEspecial
NomeGrupoDespesa	INVESTIMENTOS	:Investimentos
NomeGrupoDespesa	OUTRAS DESPESAS CORRENTES	:OutrasDespesasCorrentes

Fonte: elaborado pela autora.

Os outros arquivos que compõem o SDD (*Code Mapping, Properties e Timeline*) não foram alterados para a execução deste trabalho, e, portanto, não serão apresentados neste texto. Finalizado o preenchimento dos *templates* é necessário verificar o formato dos arquivos seja *.csv* e a codificação de separação sejam vírgulas. Então, será iniciada a fase de "processamento e armazenamento do conhecimento" que irá utilizar um *script* para transformar esses arquivos em um grafo de conhecimento expresso no formato RDF.

5.2 Processamento e armazenamento do conhecimento

Após preencher os arquivos que compõem a técnica do SDD, é aplicado um *script* Python, o *sdd2rdf*, para gerar um fragmento de grafo de conhecimento. A serialização dos metadados descritos no SDD e na ontologia formam uma sequência de linhas que os descrevem em RDF. Dessa forma, é possível que um programa utilize aquelas informações e o conhecimento descrito por elas[35]. Então, esses dados são persistidos e armazenados no Virtuoso em formato manipulável SPARQL.

Para executar o *script*, é necessário baixar o projeto no repositório do "*Semantic Data Dictionary*[36]" disponível no *GitHub* e realizar as configurações de ambiente necessárias

[35] Como este trabalho é uma prova de conceito a transformação do arquivo resultante da aplicação do sdd2rdf foi realizada manualmente. Para fazer isso de forma automatizada, é necessário estabelecer um algoritmo que considere as necessidades do processo.
[36] Disponível em https://github.com/tetherless-world/SemanticDataDictionary . Acesso em 18 out 2022.

para executar a ferramenta. Dessa forma, é possível acionar na máquina local os artefatos necessários. Para isso, deve ser instalado[37] o Python 3.7 e as bibliotecas *Pandas* e *Numpy*. Depois, é preciso localizar, via *prompt* de comando, o caminho da pasta local em que se está o *script* a ser executado. Então, após configurar o arquivo "config.ini.example" com as variáveis correspondentes ao caminho definido, retorna-se ao *prompt* de comando e executa-se o comando "python *sdd2rdf*.py config.ini.example" que gera uma série de *outputs*.

A execução do *script* gera arquivos em diferentes formatos, entre eles o mais importante é o de formato *.trig,* uma vez que é o mais comum na leitura de grafos de conhecimento. A

[37] Script: \appdata\local\programs\python\python37;
\AppData\Local\Programs\Python\Python37\Lib\site-packages\pandas;
\AppData\Local\Programs\Python\Python37\Lib\site-packages\numpy.

Figura 23 apresenta um recorte do código gerado, que é apresentado em forma gráfica no Apêndice II. Os três blocos de código apresentados mostram como foi gerado a conexão entre os arquivos inseridos nas pastas. Essa saída confere com as asserções realizadas nos arquivos de *input.* Observa-se o relacionamento criado entre os conceitos, descritos no DM, de indicação e seu código identificador: $NumeroIndicacao \rightarrow hasco{:}\, originalId \rightarrow isAttributeOf \rightarrow indicacao \rightarrow$ "52688". De forma semelhante, também foi rastreada a relação entre classes e instâncias, deixando explícito que Agostinho Patrus é o responsável pela indicação de número 52688:

$$ResponsavelNome \rightarrow freya{:}\, Indicacao \rightarrow isAttribuiteOf \rightarrow responsavel \rightarrow "AGOSTINHO\ PATRUS\ FILHO"$$

Figura 23 – Parte do arquivo gerado pelo script (*freya-kg.trig*).

```
<http://www.semanticWeb.org/freya#NumeroIndicacao-96a986849900433b1006dcc95b09bc06>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.semanticWeb.org/freya#NumeroIndicacao> ;
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>    hasco:originalId ;
        <http://semanticscience.org/resource/isAttributeOf>
<http://www.semanticWeb.org/freya#indicacao-e581a563cc183eea2ef5b3896823502e> ;
        <http://semanticscience.org/resource/hasValue>    "52688"^^xsd:integer .

    <http://www.semanticWeb.org/freya#ResponsavelNome-944f0bcf5ae15c93b3c583c5f1776e0d>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.semanticWeb.org/freya#ResponsavelNome> ;
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>    freya:Indicacao ;
        <http://semanticscience.org/resource/isAttributeOf>
<http://www.semanticWeb.org/freya#responsavel-656efe1683ffd43f6102f9d4cd33b232> ;
        <http://semanticscience.org/resource/hasValue>    "AGOSTINHO PATRUS
FILHO"^^xsd:string .

<http://www.semanticWeb.org/freya#TipoIndicacao-85da08736956ece4cce6037de401c54d>
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.semanticWeb.org/freya#TipoIndicacao> ;
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>    freya:Indicacao ;
        <http://semanticscience.org/resource/isAttributeOf>
<http://www.semanticWeb.org/freya#execucaodaindicacao-9597584ab68638db5a40ae9f3a3cb81d> ;
        <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>    ":Resolucao"^^xsd:string ;
        <http://semanticscience.org/resource/hasValue>    "RESOLUCAO"^^xsd:string .
```

Fonte: elaborado pela autora, grifo nosso.

Para tornar mais simples a leitura dos códigos gerados, utilizou-se o conversor ":*isSemantic*"[38] para transformar o arquivo *.trig* no formato Turtle (*.ttl*). Essa extensão permite visualizar as informações de forma mais direta, sem a necessidade de apresentar as URIs de cada conjunto. A Figura 24 apresenta as linhas que podem ser comparadas à

[38] Disponível em: https://issemantic.net/rdf-converter. Acessado em 26 ago 2022.

Figura 23, apresentando o número e o responsável pela indicação.

Figura 24 – Linhas que representam a indicação 52688 em formato *.ttl*.

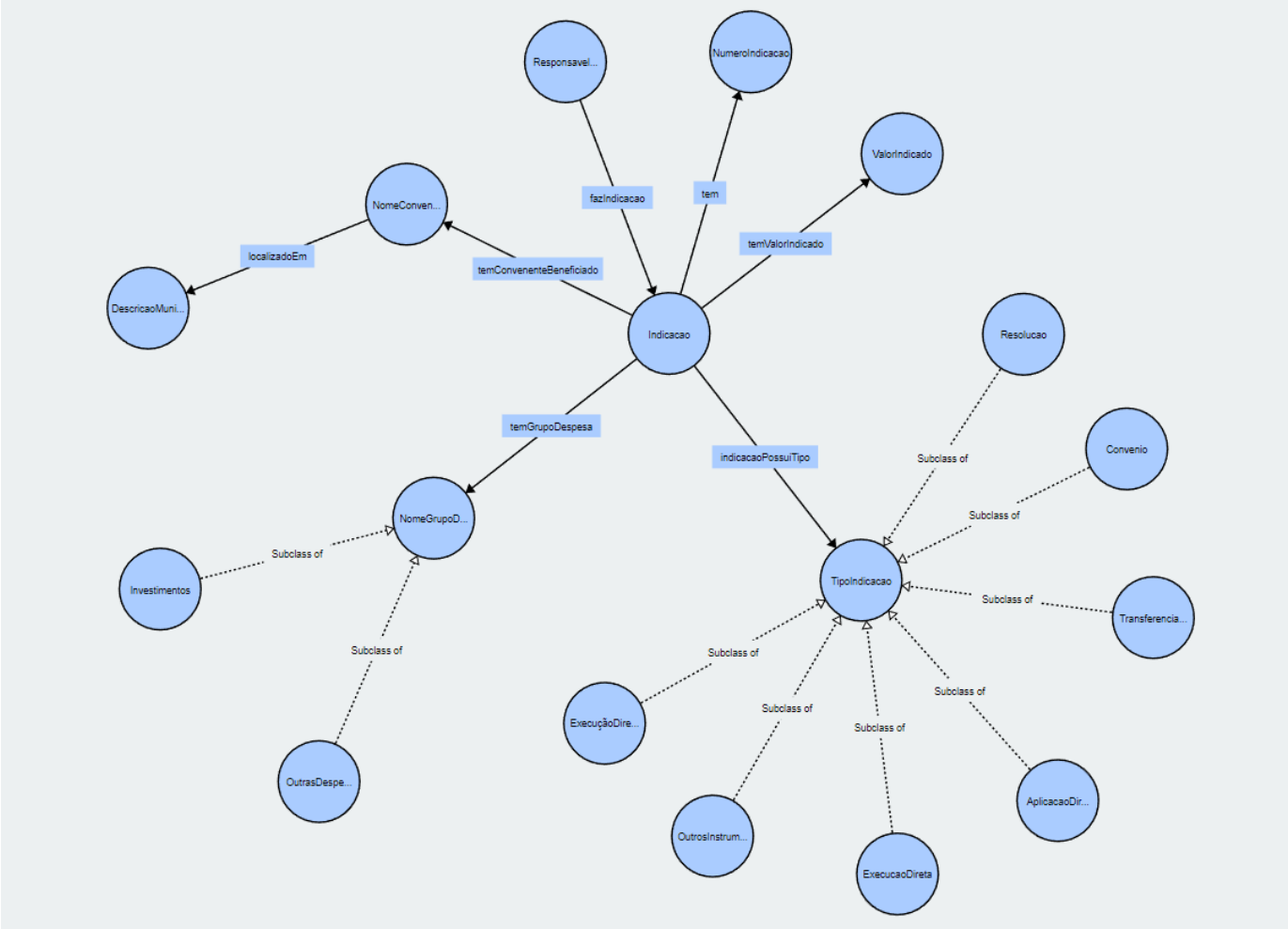
```
freya:NumeroIndicacao-96a986849900433b1006dcc95b09bc06 a
hasco:originalId,
        freya:NumeroIndicacao ;
    sio:hasValue 52688 ;
                                [...]
freya:ResponsavelNome-944f0bcf5ae15c93b3c583c5f1776e0d a
freya:Indicacao,
        freya:ResponsavelNome ;
    sio:hasValue "AGOSTINHO PATRUS FILHO"^^xsd:string ;
```

Fonte: elaborado pela autora.

Para obter uma visualização semelhante à Figura 20, foi gerado um arquivo RDF (pelo mesmo processo utilizado na transformação do *.trig* para *.ttl*) e transformado no "*OWL Syntax Converter*"[39] em *.owl*.

Figura 25 – Grafo de conhecimento gerado após utilização do *sdd2rdf*.

[39] Disponível em: http://mowl-power.cs.man.ac.uk:8080/converter/ . Acesso em 25 set 2022.

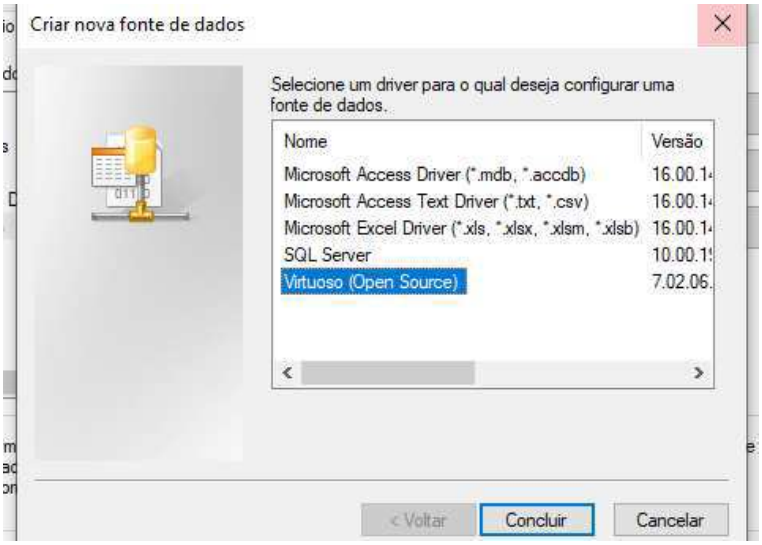


Fonte: elaborado pela autora.

De forma concomitante a esse processo, deve ser instalado localmente o "*OpenLink Virtuoso*[40]"onde se cria o repositório de dados semânticos. Depois de baixar a versão de instalação, é necessário configurar um (ODBC) que é um padrão de acesso a sistemas que permitem gerenciar bancos de dados. Ao criar uma fonte de dados (ver Figura 26 e Figura 27), é possível criar um fluxo de informações entre o Virtuoso e um programa que possa consumir essa fonte de dados. Após a instalação do servidor é necessário iniciá-lo[41].

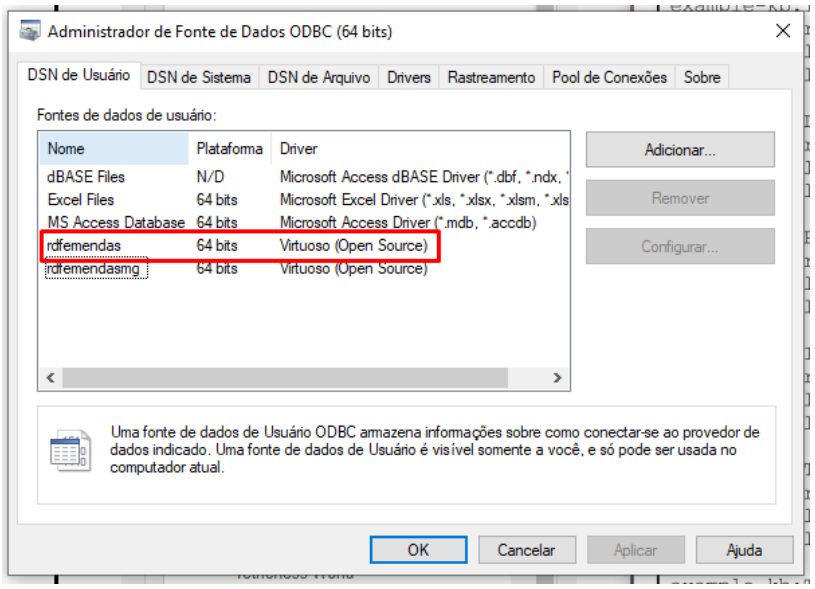
Figura 26 – Instrução para criar uma fonte de dados para o Virtuoso e configurar o ODBC.

[40] No decorrer do texto, será chamado de Virtuoso. Disponível em https://vos.openlinksw.com/owiki/wiki/VOS . Acesso em 27 set. 2022.
[41] Caso ocorra algum problema para iniciar o servidor, delete na pasta de instalação o arquivo virtuoso.lck.



Fonte: elaborada pela autora.

Figura 27 - Instrução para selecionar a fonte de dados criada no ODBC.



Fonte: elaborada pela autora.

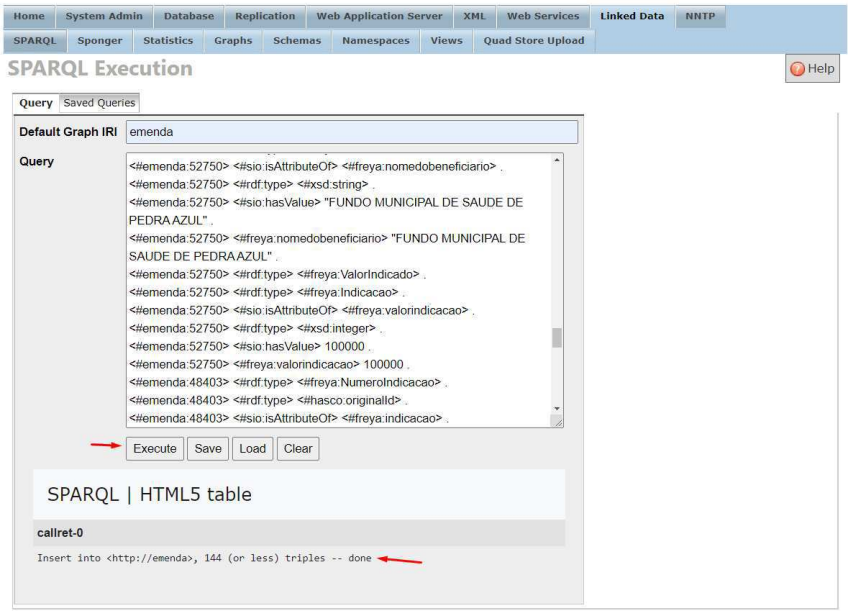
Então, a página da instalação local do servidor[42] deve ser acessada e acionado o "*Conductor*" para proceder com o login. A página apresentada na Figura 28 será visualizada. Para criar o repositório, clica no menu "*Linked Data*". Então é necessário inserir um nome desejado para o grafo e colar o texto do *script* (Figura 30) no campo de *query* do Virtuoso, assim como apresentado na Figura 29. Esse banco de dados de triplas (*triple store*) permite realizar consultas SPARQL.

Figura 28 – Página inicial do Virtuoso *Conductor* após fazer o login.

[42] http://localhost:8890/ , login e senha do administrador: dba/dba.

Fonte: elaborada pela autora.

Figura 29 - Criação do repositório semântico no Virtuoso



Fonte: elaborada pela autora, grifo nosso.

Figura 30 - Extrato do SPAQRL inserido no Virtuoso

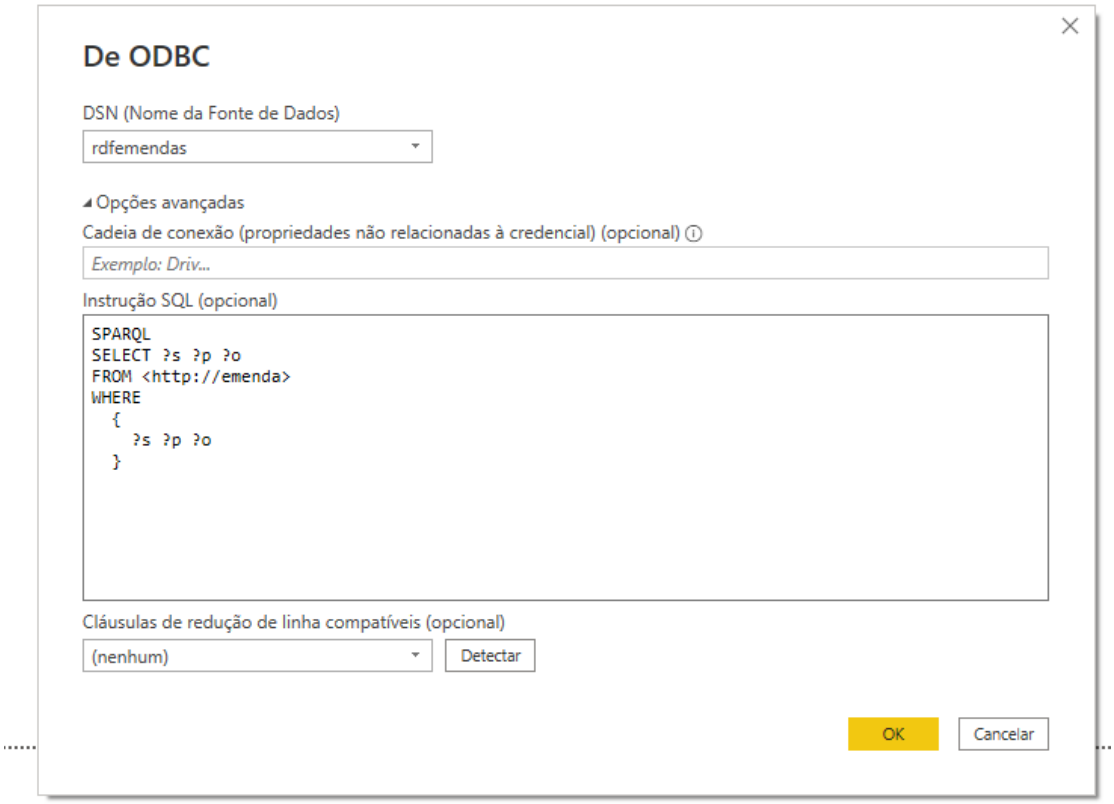
```
INSERT IN GRAPH <http://emenda>
{
<#emenda:52688> <#rdf:type> <#freya:NumeroIndicacao> .
<#emenda:52688> <#rdf:type> <#hasco:originalId> .
<#emenda:52688> <#sio:isAttributeOf> <#freya:indicacao> .
<#emenda:52688> <#rdf:type> <#xsd:integer> .
<#emenda:52688> <#sio:hasValue> 52688 .
<#emenda:52688> <#freya:indicacao> 52688 .
<#emenda:52688> <#rdf:type> <#freya:ResponsavelNome> .
<#emenda:52688> <#rdf:type> <#freya:Indicacao> .
<#emenda:52688> <#sio:isAttributeOf> <#freya:responsavel> .
<#emenda:52688> <#rdf:type> <#xsd:string> .
<#emenda:52688> <#sio:hasValue> "AGOSTINHO PATRUS FILHO" .
<#emenda:52688> <#freya:responsavel> "AGOSTINHO PATRUS FILHO" .
<#emenda:52688> <#rdf:type> <#freya:TipoIndicacao> .
<#emenda:52688> <#rdf:type> <#freya:Indicacao> .
<#emenda:52688> <#sio:isAttributeOf> <#freya:execucaodaindicacao> .
<#emenda:52688> <#rdf:type> <#freya:Resolucao> .
<#emenda:52688> <#rdf:type> <#xsd:string> .
<#emenda:52688> <#freya:execucaodaindicacao> "RESOLUCAO" .
<#emenda:52688> <#rdf:type> <#freya:NomeGrupoDespesa> .
<#emenda:52688> <#rdf:type> <#freya:Indicacao> .
<#emenda:52688> <#sio:isAttributeOf> <#freya:tipodedespesa> .
<#emenda:52688> <#rdf:type> <#freya:Investimento> .
<#emenda:52688> <#rdf:type> <#xsd:string> .
<#emenda:52688> <#sio:hasValue> "INVESTIMENTOS" .
<#emenda:52688> <#freya:tipodedespesa> "INVESTIMENTOS" .
<#emenda:52688> <#rdf:type> <#freya:DescricaoMunicipio> .
<#emenda:52688> <#rdf:type> <#freya:Municipio> .
<#emenda:52688> <#sio:isAttributeOf> <#freya:municipiobeneficiario> .
<#emenda:52688> <#rdf:type> <#xsd:string> .
<#emenda:52688> <#sio:hasValue> "SANTA MARIA DO SUACUI" .
<#emenda:52688> <#freya:municipiobeneficiario> "SANTA MARIA DO SUACUI" .
}
```

Fonte: elaborada pela autora. Versão final disponível no GitHub da pesquisa *insert_freya_sparql v2.txt*.

5.3 Análise do Conhecimento

Após ingerir as informações SPARQL, deve ser acessado o Power BI para criar uma fonte de dados a partir de um ODBC, inserir a fonte de dados criada na Figura 26 e inserir uma *query* para recuperar todas as triplas daquele fragmento de grafo de conhecimento, assim como demonstrado na Figura 31.

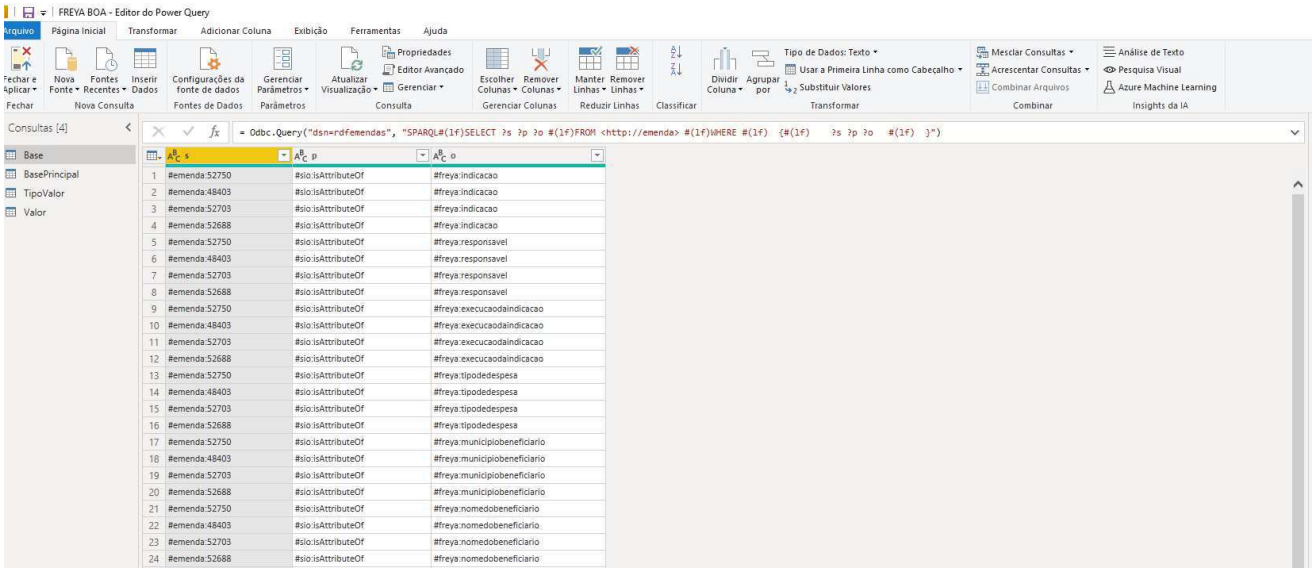
Figura 31 - Configuração do ODBC no Power BI



Fonte: elaborada pela autora.

Depois, será necessário transformar os dados para separar em tabelas fato as classes que foram ingeridas no Power BI. Esse processo deve ser acompanhado pelos especialistas de domínio que poderão validar a construção

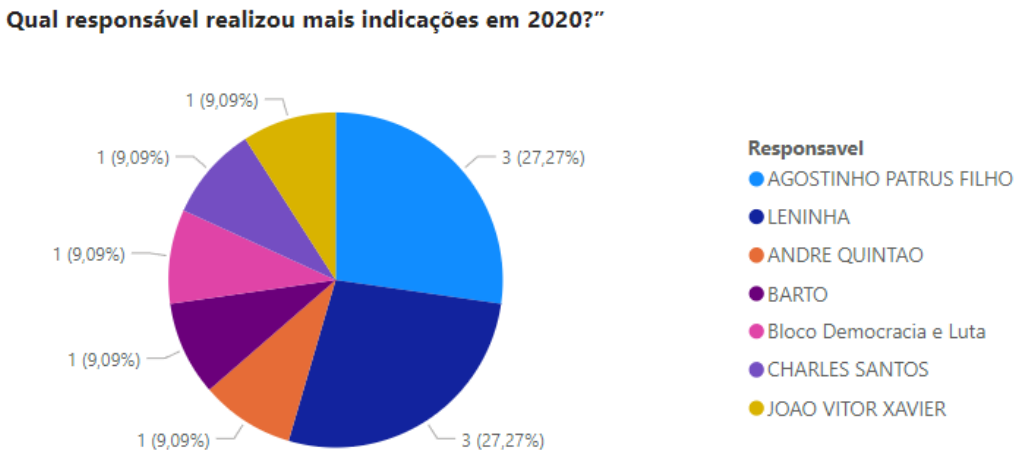
Figura 32 – Visão do Power BI depois da conexão com o ODBC



Fonte: elaborada pela autora.

Após preparar as dimensões no programa, a pergunta de competência que norteia a execução do ciclo da metodologia deve ser representada em formato de gráficos e painéis de acompanhamento. Dessa forma, a resposta é a representada em um formato visual que pode ser validado por diferentes agentes públicos e cidadãos. A Figura 33 apresenta um gráfico pizza em que são contadas quantas indicações cada responsável fez em 2020, de acordo com o conjunto de dados ingeridos.

Figura 33 – Painel criado que responde à pergunta "Qual responsável realizou mais indicações em 2020?"



Responsavel	ValorIndicado
CHARLES SANTOS	R$2.000.002
AGOSTINHO PATRUS FILHO	R$1.000.000
JOAO VITOR XAVIER	R$200.000
LENINHA	R$180.000
BARTO	R$150.000
Bloco Democracia e Luta	R$100.000
ANDRE QUINTAO	R$59.990
Total	**R$3.689.992**

Fonte: elaborado pela autora

Ao montar os recursos, é possível ampliar as informações representadas e avaliar a necessidade de, no próximo ciclo, aprofundar em questões a respeito dos valores indicados e das possíveis relações que podem existir entre outras classes da anotação realizada. Isso é demonstrado na Figura 33 uma vez que, apesar de ter um maior número de indicações, o responsável Agostinho Patrus Filho, utilizou menos recursos financeiros. Isso pode sinalizar uma maior distribuição entre diferentes municípios que poderão receber esse orçamento.

Apesar do conjunto de dados utilizado nessa prova de conceito não ser a totalidade dos registros de 2020, uma possível análise é a relação entre as indicações, valor e tipo de indicação. Isso é relevante para compreender qual o maior volume de recursos em relação ao volume de indicações. Além disso, ao seccionar a pesquisa por responsável, tipo de indicação e valor, é possível encontrar quais os tipos são mais utilizados a depender de um município. Esse painel representado na Figura 34, ainda que não seja uma pergunta de competência sugerida pelo grupo focal, pode ser relevante para alguma tomada de decisão e pode ser levada ao próximo encontro para que seja debatida e anotada. Dessa forma, é possível explorar um conjunto limitado de dados e, aos poucos, agregar informações e conhecimento ao grafo e elaborar painéis mais confiáveis e publicáveis na Internet.

Figura 34 -Painel que relaciona responsável pela indicação e município.

Fonte: elaborado pela autora

A seção seguinte apresenta as conclusões deste trabalho, indica algumas limitações e sugere algumas previsões de trabalhos futuros.

6 CONCLUSÃO

Dados podem ser codificados em diferentes formatos e o conhecimento pode estar implícito neles. Uma tabela que contenha dados sobre emendas parlamentares de um ano específico pode gerar dúvidas a respeito dos conceitos e o contexto em que essas informações devem ser interpretadas. Os Dados Abertos Governamentais são uma iniciativa que busca trazer mais transparência para a gestão e políticas públicas. O acesso à informação é direito de todo cidadão e, assim como determinado pela lei federal 12.527/2011, a publicação de dados deve ser em formatos acessíveis e automatizáveis. A abertura de dados pode ser guiada por processos que facilitem as publicações na Internet. Dentre as soluções tecnológicas possíveis, enfatiza-se as que utilizam as tecnologias da *Web* Semântica como base para esse processo.

Então, a anotação de dados pode ser uma forma de indicar as padronizações, características e as informações descritivas sobre um campo do conhecimento específico. Os metadados que formam esse domínio contribui para a Gestão do Conhecimento. As ontologias são ferramentas que podem auxiliar na representação de recursos na *Web* utilizando linguagens como RDF e OWL uma vez que, a partir de condições lógicas e inferências, definem as hierarquias de conceitos e restrições sobre um conjunto de classes e suas propriedades. Baseada em ontologias, a técnica sugerida do Dicionário Semântico de Dados (*Semantic Data Dictionary* ou SDD) proposta por Rashid *et al.* (2017) permite manipular, anotar e integrar dados de diferentes fontes e formatos.

Essa abordagem utiliza um conjunto de documentos de metadados pré-definidos fundamentados em ontologias para enriquecer os dados presentes em um conjunto. Uma das recomendações para que a técnica seja aplicada com sucesso é a participação de diferentes atores no processo de anotação, que deve ser orientada por especialistas de domínio e engenheiros do conhecimento. Isso se deve ao fato de que é necessário compreender o domínio a ser representado pelo modelo conceitual. Dessa maneira a execução técnica realizada em conjunto permite que conhecimento possa ser representado corretamente.

Este trabalho planejou adaptar metodologias e aplicar o produto para contribuir com a publicação de dados semânticos em contexto governamental. A representação sistemática é baseada nos preceitos das metodologias ágeis e se fundamenta nos conceitos da *Design Science Research*. Essa perspectiva traz ao trabalho apresentado rigor científico que permite unir problemas práticos a soluções teóricas e a partir de questões conceituais buscar soluções em experiências. No estudo de caso das emendas parlamentares de 2020 foi possível condensar, em formato de prova de conceito, o conhecimento do grupo focal aplicado em um a ontologia,

que foi anotada e transformada em um grafo de conhecimento. Esse resultado foi aplicado e processado em uma ferramenta de visualização de painéis de acompanhamento, sempre considerando a pergunta norteadora para aquele ciclo de desenvolvimento. Uma possibilidade para trabalhos futuros é apresentar e aplicar a metodologia em outras áreas correlatas para alinhar conceitos semelhantes e, assim, integrar os dados. Dessa forma, serão alinhadas as dificuldades e visão de outras partes envolvidas no processo.

Dessa forma, é esperado que este trabalho possa contribuir com o movimento de Dados Abertos Governamentais e acrescentar ferramentas que permitam para a anotação e formalização de domínios específicos como forma de aprimorar os processos da gestão pública. Então, para trabalhar com um grande volume de dados, é necessário avaliar como garantir que o conhecimento daqueles conceitos esteja representado em formatos consumíveis por máquinas. A construção de grafos de conhecimento necessita de um esforço contínuo dos atores envolvidos e deve ser incentivado por gestores. Futuramente, poderia ser gerado um *Knowledge Graph* que reúna dados de outros conjuntos anotados com a mesma metodologia e apresentar esses dados em um portal governamental unificado de Dados Abertos Conectados.

A utilização de Dados Abertos Conectados e as tecnologias da *Web* Semântica para inferir informações podem gerar uma melhor compreensão de dados heterogêneos. A forma semiestruturada de análise proposta neste trabalho pode trazer diversos benefícios para a sociedade.

Ir para além da simples elaboração de dicionário de dados auxilia na construção de ferramentas e acelera as discussões em grupos focais. A execução deste trabalho necessitou consultar fontes documentais para delimitar previamente o escopo da discussão com os agentes públicos. Portanto, a continuidade deste trabalho, aplicado ao contexto de Emendas Parlamentares, é factível uma vez que ainda existem classes a serem mapeadas e por se tratar de um domínio de interesse público e político. O controle a ser executado a partir dessa metodologia e das perguntas de competência futuras pode trazer benefícios para o governo. Portanto é necessária a expansão da ontologia, levando em consideração outras perguntas de competência e os dados disponíveis. Além disso, é necessário ampliar a integração e reutilização de termos de outras ontologias para tornar os conceitos mais abrangentes e reutilizáveis em outros contextos.

As formas de executar o processamento e armazenamento do conhecimento não necessitam se limitar à utilização dos mesmos sistemas gerenciadores apresentados neste trabalho. O processo de ingestão dos dados se deu de forma manual para os fins de comprovação

da prova de conceito delimitada nos objetivos do trabalho. As dificuldades encontradas podem ser solucionadas através da automatização de processos utilizando linguagem de programação.

Uma das limitações deste trabalho é partir do pressuposto que existe uma base de dados relacional para extração dos dados. Isso impõe uma necessidade específica para a execução da metodologia. Dessa forma, é necessário avaliar, em trabalhos futuros, outras formas de iniciar a execução desse processo de enriquecimento semântico e publicação de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

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.

ALLEMANG, Dean; HENDLER, James A. **Semantic *web* for the working ontologist modeling in RDF, RDFS and OWL**. 1. ed. Burlington, MA: Morgan Kaufmann, 2008.

ALMEIDA, Mauricio Barcellos. **ONTOLOGIA EM CIÊNCIA DA INFORMAÇÃO: Tecnologia e Aplicações Coleção Representação do conhecimento em Ciência da Informação Volume 2**. 1. ed. Curitiba: EDITORA CRV, 2021. DOI: 10.24824/978652511477.4.

AUER, Soren. Introduction to LOD2. *Em*: AUER, Sören; BRYL, Volha; TRAMP, Sebastian (org.). Lecture Notes in Computer Science1. ed. Cham: Springer International Publishing, 2014. v. 8661. DOI: 10.1007/978-3-319-09846-3. Disponível em: http://link.springer.com/10.1007/978-3-319-09846-3.

AUER, Sören; BIZER, Christian; KOBILAROV, Georgi; LEHMANN, Jens; CYGANIAK, Richard; IVES, Zachary. DBpedia: A Nucleus for a *Web* of Open Data. **The Semantic *Web* - Lecture Notes in Computer Science**, *[S. l.]*, v. 4825, p. 722–735, 2007. DOI: https://doi.org/10.1007/978-3-540-76298-0_52. Disponível em: https://doi.org/10.1007/978-3-540-76298-0_52. Acesso em: 25 jul. 2022.

ÁVILA, Thiago José Tavares. **Uma proposta de modelo de processo para publicação de Dados Abertos Conectados Governamentais**. 2015. Dissertação (Mestrado) - Instituto de Computação da Universidade Federal de Alagoas, Maceió, 2015.

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.

BECK, Kent et al. **Manifesto para Desenvolvimento Ágil de *Software***. 2001. Disponível em: https://agilemanifesto.org/iso/ptbr/manifesto.html. Acesso em: 6 fev. 2022.

BELISÁRIO, Adriano; GEHRKE, Marília; CUBAS, Marina Gama; MENEGAT, Rodrigo. **Fluxo de trabalho com dados-Do zero à prática Escola de Dados**. São Paulo: Open Knowledge Brasil, 2020. v. ePUB Disponível em: https://escoladedados.org/wp-content/uploads/2021/03/livrov2.pdf. Acesso em: 28 jun. 2021.

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. *Em*: (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.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Orla. The Semantic *Web*. **Scientific American**, *[S. l.]*, v. 284, n. 5, p. 34–43, 2001.

BRASIL. **LEI Nº 14.129, DE 29 DE MARÇO DE 2021**. 2011a. Disponível em: https://www.in.gov.br/en/*web*/dou/-/lei-n-14.129-de-29-de-marco-de-2021-311282132. Acesso em: 31 maio. 2021.

BRASIL. **Declaração de Governo Aberto**. 2011b. Disponível em: https://www.gov.br/cgu/pt-br/governo-aberto/central-de-conteudo/documentos/arquivos/declaracao-governo-aberto.pdf. Acesso em: 19 jan. 2021.

BRASIL. **LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011 - Lei de Acesso à Informação, LAI**. 2011c. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm. Acesso em: 19 jan. 2021.

BRASIL. LEI Nº 13.460, DE 26 DE JUNHO DE 2017. **Dispõe sobre participação, proteção e defesa dos direitos do usuário dos serviços públicos da administração pública**, *[S. l.]*, 2017. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/l13460.htm.

BRASIL. **Lei nº 13.709, de agosto de 2018**. 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 19 jan. 2021.

BRITTO, Gustavo C.; RUY, Fabiano B.; AZEVEDO, Carlos L. B. Um Ambiente para Integração de Dados Abertos relativos à Despesa Pública. **CEUR Workshop Proceedings**, *[S. l.]*, v. 2728, n. 2018, p. 176–189, 2020.

BUCHMANN, Robert A.; KARAGIANNIS, Dimitris. Enriching linked data with semantics from domain-specific diagrammatic models. **Business and Information Systems Engineering**, *[S. l.]*, v. 58, n. 5, p. 341–353, 2016. DOI: 10.1007/s12599-016-0445-1.

CARBONARO, Antonella. Linked data and semantic *web* technologies to model context information for policy-making. **Journal of Ambient Intelligence and Humanized Computing**, *[S. l.]*, v. 12, n. 4, p. 4395–4406, 2021. DOI: 10.1007/s12652-019-01341-y. Disponível em: https://doi.org/10.1007/s12652-019-01341-y.

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.

ESTEVANOVIC, Marcela Pires. **Análise da prestação de serviços digitais no estado de Minas Gerais no período de 2009 a 2019: uma contribuição a partir do neoinstitucionalismo**. 2019. Belo Horizonte, Brasil, 2019.

FANTINI, *Web*ber de Souza. **Publicação de dados conectados sobre despesas orçamentárias do governo federal brasileiro**. 2015. Universidade Federal de Pernambuco, Recife, 2015. Disponível em: https://repositorio.ufpe.br/handle/123456789/16779. Acesso em: 19 fev. 2022.

FEKETTIA, Alexandre Loriggio; FARIAS, Victor Mitsunaga; MUSTARO, Pollyana Notargiacomo. Aplicações de gamificação e técnicas de motivação à aprendizagem da metodologia ágil scrum. **VIII International Conference on Engineering and Computer Education**, Luanda, Angola, 2013. DOI: 10.14684/ICECE.8.2013.328-332. Acesso em: 6 fev. 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.

GRUBER, Thomas R. **Toward Principles for the Design of Ontologies**. **International Journal of Human-Computer Studies**, 1995. Disponível em: http://linkinghub.elsevier.com/retrieve/pii/S1071581985710816.

HESSEN, Johannes. **Teoria do Conhecimento**. São Paulo: Martins Fontes, 1999.

HEVNER, Alan R. A Three Cycle View of Design Science Research. **Scandinavian Journal of Information Systems**, *[S. l.]*, v. 19, n. 2, 2007. Disponível em: https://www.uio.no/studier/emner/jus/afin/FINF4002/v13/hefner-design.pdf. Acesso em: 29 jun. 2022.

HITZLER, Pascal; KRÖTZSCH, Markus; RUDOLPH, Sebastian. **Foundations of Semantic *Web* Technologies**. Boca Raton: CRC Press, 2010.

HOXHA, Julia; BRAHAJ, Armand. Open government data on the *web*: A semantic approach. **Proceedings - 2011 International Conference on Emerging Intelligent Data and *Web* Technologies, EIDWT 2011**, *[S. l.]*, p. 107–113, 2011. DOI: 10.1109/EIDWT.2011.24. Acesso em: 24 out. 2021.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados Abertos Conectados**. São Paulo, SP: Novatec Editora, 2015. DOI: 10.13140/RG.2.1.4355.6329. Disponível em: http://ce*web*.br/livros/dados-abertos-conectados/. Acesso em: 14 out. 2019.

JEAN-BAPTISTE, Lamy. **Ontologies with Python**. [s.l: s.n.]. DOI: 10.1007/978-1-4842-6552-9.

JOVANOVIK, Milos; TRAJANOV, Dimitar; KOSTOVSKI, Martin. Open Data Portal based on Semantic *Web* Technologies. *[S. l.]*, 2012. DOI: 10.13140/RG.2.2.23588.88969. Disponível em: https://www.researchgate.net/publication/230691194. Acesso em: 13 set. 2020.

LANGE, Christoph. Ontologies and languages for representing mathematical knowledge on the semantic *web*. **Semantic *Web***, *[S. l.]*, v. 4, n. 2, p. 119–158, 2013. DOI: 10.3233/SW-2012-0059.

LIRA, Márcio Angelo Bezerra De. **Uma Abordagem Para Enriquecimento Semântico De Metadados Para Publicação De Dados Abertos**. 2014. Dissertação (Mestrado) - Universidade Federal de Pernambuco, Recife, PE, 2014. Disponível em: https://repositorio.ufpe.br/handle/123456789/11570. Acesso em: 4 maio. 2020.

LOD PROJECT. **The Linked Open Data Cloud**. 2022. Disponível em: https://lod-cloud.net/. Acesso em: 11 out. 2022.

MARTINS, Livio Cravo; CRAVEIRO, Gisele Silva; ALCÁZAR, José de Jesús. **Definição e Validação de uma Ontologia para o Orçamento Público Federal Brasileiro (v.1.0)Relatório Técnico PPgSI-002/2013**. São Paulo.

MARTINS, Luiz Carlos Barbosa. **Proposta de Arquitetura de Publicação Automatizada de Dados Abertos Conectados Utilizando Meta-Dados e Ontologias**. 2018. Dissertação (Mestrado) - Universidade de Brasília, Brasília, DF, 2018. Disponível em: https://repositorio.unb.br/handle/10482/34816. Acesso em: 4 maio. 2020.

MINAS GERAIS. **Constituição do Estado de Minas Gerais**. 28 ed ed. Belo Horizonte: Assembleia Legislativa do Estado de Minas Gerais, 1989. v. 2021 Disponível em: https://www.almg.gov.br/export/sites/default/consulte/legislacao/Downloads/pdfs/ConstituicaoEstadual.pdf. Acesso em: 8 fev. 2022.

MINAS GERAIS. **Anexo V da Lei Orçamentária Anual**. Volume VI ed. Belo Horizonte: Governo do Estado de Minas Gerais, 2020. Disponível em: https://planejamento.mg.gov.br/pagina/planejamento-e-orcamento/lei-orcamentaria-anual-loa/lei-orcamentaria-anual-loa. Acesso em: 9 fev. 2021.

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

MINAS GERAIS. Decreto 48138, de 17/02/2021 - Assembleia de Minas. . 17 fev. 2021 b.

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.

NOY, Natasha; MCGUINNESS, Deborah. Ontology 101. **Medical Informatics**, *[S. l.]*, p. 1–5, 2011.

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

OCDE. **Revisão do Governo Digital do Brasil Rumo à Transformação Digital do Setor Público - Principais Conclusões**. Paris. Disponível em: http://editor.planejamento.gov.br/seminariodigital/seminario/digital-gov-review-brazil-portugues.pdf. Acesso em: 8 jul. 2020.

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

OPEN DATA INSTITUTE. **The Data Spectrum**. 2021. Disponível em: https://theodi.org/about-the-odi/the-data-spectrum/. Acesso em: 31 jan. 2022.

OPEN KNOWLEDGE BRASIL. **Por que "open"?** 2021. Disponível em: https://www.ok.org.br/dados-abertos/. Acesso em: 20 jan. 2021.

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.

PEREIRA, Larissa Mariany Freiberger. **OGDPub: Uma Ontologia Para Publicação de Dados Abertos Governamentais**. 2017. Dissertação (Mestrado) - Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, 2017. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/179907. Acesso em: 4 jun. 2020.

POSSAMAI, Ana Júlia. Instituições e desempenho do governo digital : Argentina, Brasil, Chile, Colômbia e Uruguai em perspectiva comparada. *[S. l.]*, 2010. Disponível em: https://www.lume.ufrgs.br/handle/10183/28382. Acesso em: 24 mar. 2019.

RAMIREZ, João P. Righi. Vocabulário Controlado do Governo Eletrônico (VCGE): Uma análise com base em em critérios aplicáveis a taxonomias e tesauros. *[S. l.]*, p. 189, 2015.

RASHID, Sabbir M.; MCCUSKER, James P.; PINHEIRO, Paulo; BAX, Marcello P.; SANTOS, Henrique; STINGONE, Jeanette A.; DAS, Amar K.; MCGUINNESS, Deborah L. The Semantic Data Dictionary – An Approach for Describing and Annotating Data. **Data Intelligence**, *[S. l.]*, p. 443–486, 2020. DOI: 10.1162/dint_a_00058. Disponível em: https://www.researchgate.net/publication/341000060_The_Semantic_Data_Dictionary_-An_Approach_for_Describing_and_Annotating_Data. Acesso em: 27 jul. 2020.

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.

REIS JÚNIOR, Cleyton Peixoto Dos. **UnBGOLDProv: Arquitetura de proveniência de dados para um workflow de publicação de dados abertos governamentais**. 2020. Dissertação - Universidade de Brasília, Brasília, 2020. Disponível em: https://repositorio.unb.br/bitstream/10482/38556/1/2020_CleytonPeixotodosReisJ%C3%BAnior.pdf. Acesso em: 10 set. 2022.

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.

SAGI, Tomer; LISSANDRINI, Matteo; PEDERSEN, Torben Bach; HOSE, Katja. A design space for RDF data representations. **The VLDB Journal 2022**, *[S. l.]*, p. 1–27, 2022. DOI: 10.1007/S00778-021-00725-X. Disponível em: https://link.springer.com/article/10.1007/s00778-021-00725-x. Acesso em: 24 jan. 2022.

SANTAREM SEGUNDO, Jose Eduardo. *Web* semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação,** *[S. l.]*, v. 8, n. 2, p. 219–238, 2015. Disponível em: https://revistas.ancib.org/index.php/tpbci/article/view/359. Acesso em: 5 jul. 2020.

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. *Em*: **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.

SCHEIDER, Simon; OSTERMANN, Frank O.; ADAMS, Benjamin. Why good data analysts need to be critical synthesists. Determining the role of semantics in data analysis. **Future Generation Computer Systems**, *[S. l.]*, v. 72, p. 11–22, 2017. DOI: 10.1016/j.future.2017.02.046. Disponível em: http://dx.doi.org/10.1016/j.future.2017.02.046.

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.

SEGOV. **Execução de Emendas 2020**. 2021a. Disponível em: https://www.emendas.mg.gov.br/execucao-de-emendas-2020/. Acesso em: 3 mar. 2022.

SEGOV. **Resolução Segov 001/2021, de fevereiro de 2021**. 2021b. Disponível em: http://www.sigconsaida.mg.gov.br/wp-content/uploads/arquivos/resolucoes/resolucao_segov_001_01_02_2021_SEI.pdf. Acesso em: 2 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.

SIEMENS, George. Connectivism: A Learning Theory for the Digital Age. **International Journal of Instructional Technology and Distance Learning**, *[S. l.]*, v. 2, 2005. Disponível em: http://www.itdl.org/Journal/Jan_05/article01.htm. Acesso em: 6 fev. 2022.

SILVA, Evaldo de Oliveira Da. **Integração de padrões de análise e ontologias de domínio: um estudo de caso no domínio de gestão urbana**. 2008. *[S. l.]*, 2008.

SILVA, Narjara Bárbara Xavier; SALES, Luana Farias; DOS SANTOS, Jhonathan Divino Ferreira. Estudo de categorias para sistematização de conceitos em Gestão do Conhecimento. **AtoZ: novas práticas em informação e conhecimento**, *[S. l.]*, v. 9, n. 1, p. 32, 2020. DOI: 10.5380/atoz.v9i1.74363.

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.

TETHERLESS WORLD. **Semantic Data Dictionary**. 2022. Disponível em: Semantic Data Dictionary. Acesso em: 19 fev. 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.

W3C. **OWL - Semantic *Web* Standards**. 2012. Disponível em: https://www.w3.org/2001/sw/wiki/OWL. Acesso em: 27 jan. 2022.

W3C. **RDF 1.1 Concepts and Abstract Syntax**. 2014. Disponível em: https://www.w3.org/TR/rdf11-concepts/. Acesso em: 1 mar. 2022.

W3C. **Vocabularies**. 2015. Disponível em: https://www.w3.org/standards/semantic*web*/ontology. Acesso em: 5 jun. 2022.

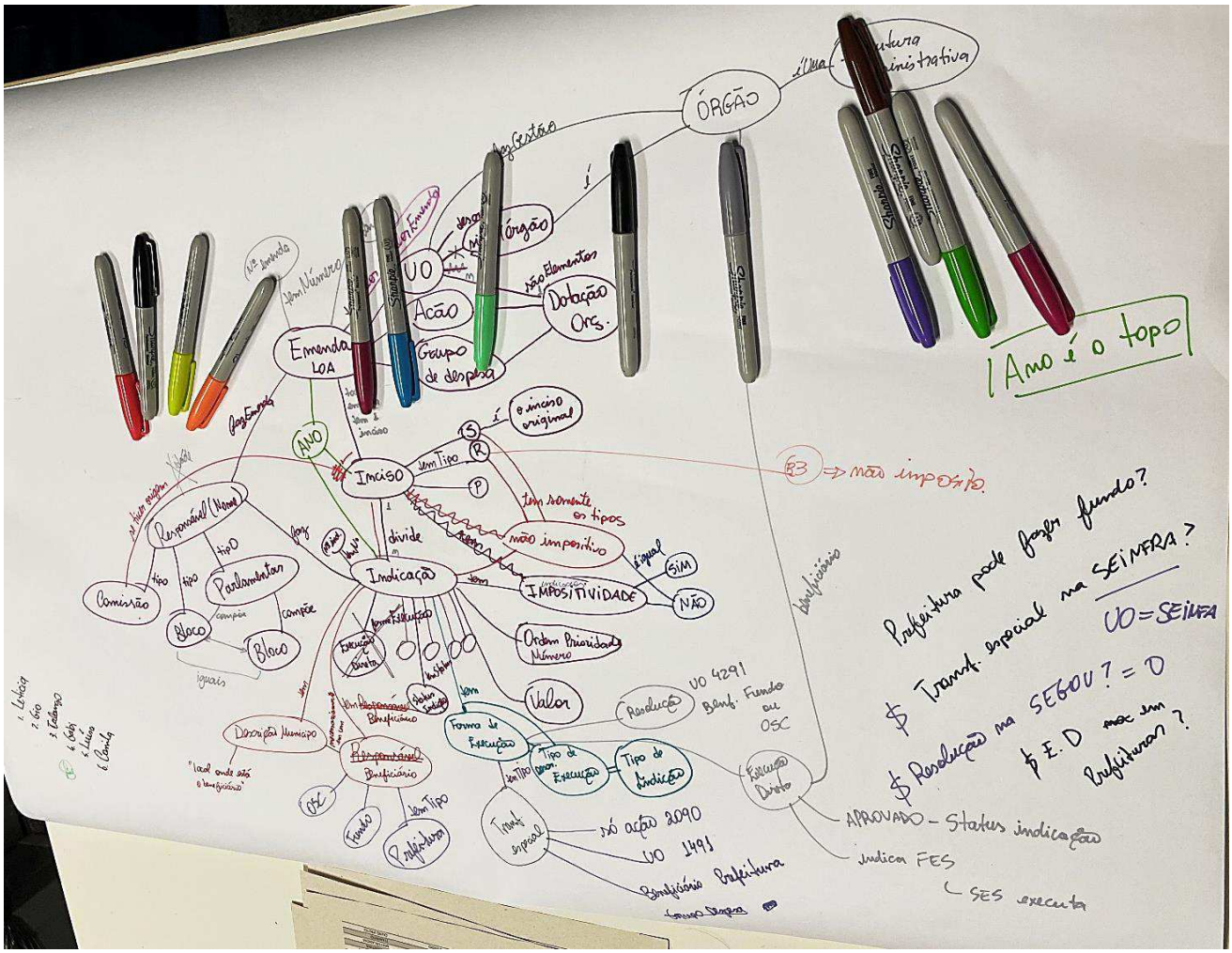
WIERINGA, Roel. Design science as nested problem solving. *Em*: PROCEEDINGS OF THE 4TH INTERNATIONAL CONFERENCE ON DESIGN SCIENCE RESEARCH IN INFORMATION SYSTEMS AND TECHNOLOGY - DESRIST '09 2009, New York, New York, USA. **Anais** [...]. New York, New York, USA: ACM Press, 2009. p. 1. DOI: 10.1145/1555619.1555630.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

ZOU, Xiaohan. A Survey on Application of Knowledge Graph. **Journal of Physics: Conference Series**, Cingapura, v. 1487, n. 1, 2020. DOI: 10.1088/1742-6596/1487/1/012016.

APÊNDICE I - ROTEIRO E IMAGENS DO GRUPO FOCAL

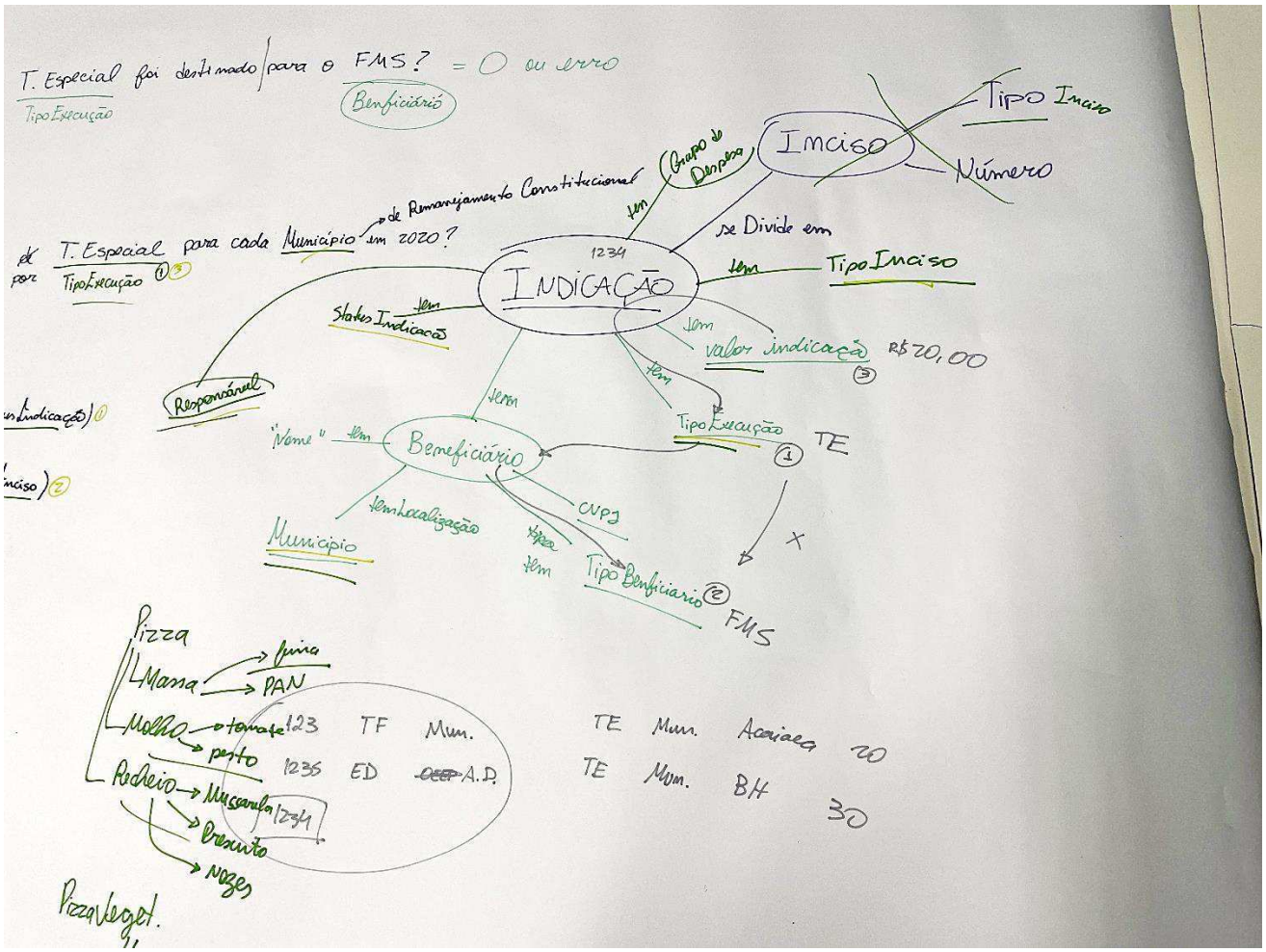
- Qual o objetivo do grupo focal? Validar a ontologia construída e avaliar o dataset construído em ciclos de reuniões
- Roteiro (3 ou 4 ciclos)
    - Perguntas que podem direcionar as demandas
    - Apresentação dos conceitos básicos: dados estruturados, ontologia, dicionário de dados, dicionário semântico de dados
    - Apresentação gráfica da ontologia
    - Discussão dos relacionamentos entre as classes
    - Possíveis perguntas para serem respondidas que vão além do que já foi mapeado
        - Na sua percepção, qual é o problema dos dados das Emendas Parlamentares Impositivas?
            - Elencar problemas
        - Quais são as ações a serem realizadas para sanar os problemas? Quais propostas para a solução dos problemas?

Figura 35 – Mapa conceitual desenvolvido no primeiro encontro do grupo focal, no dia 10/11/2021.



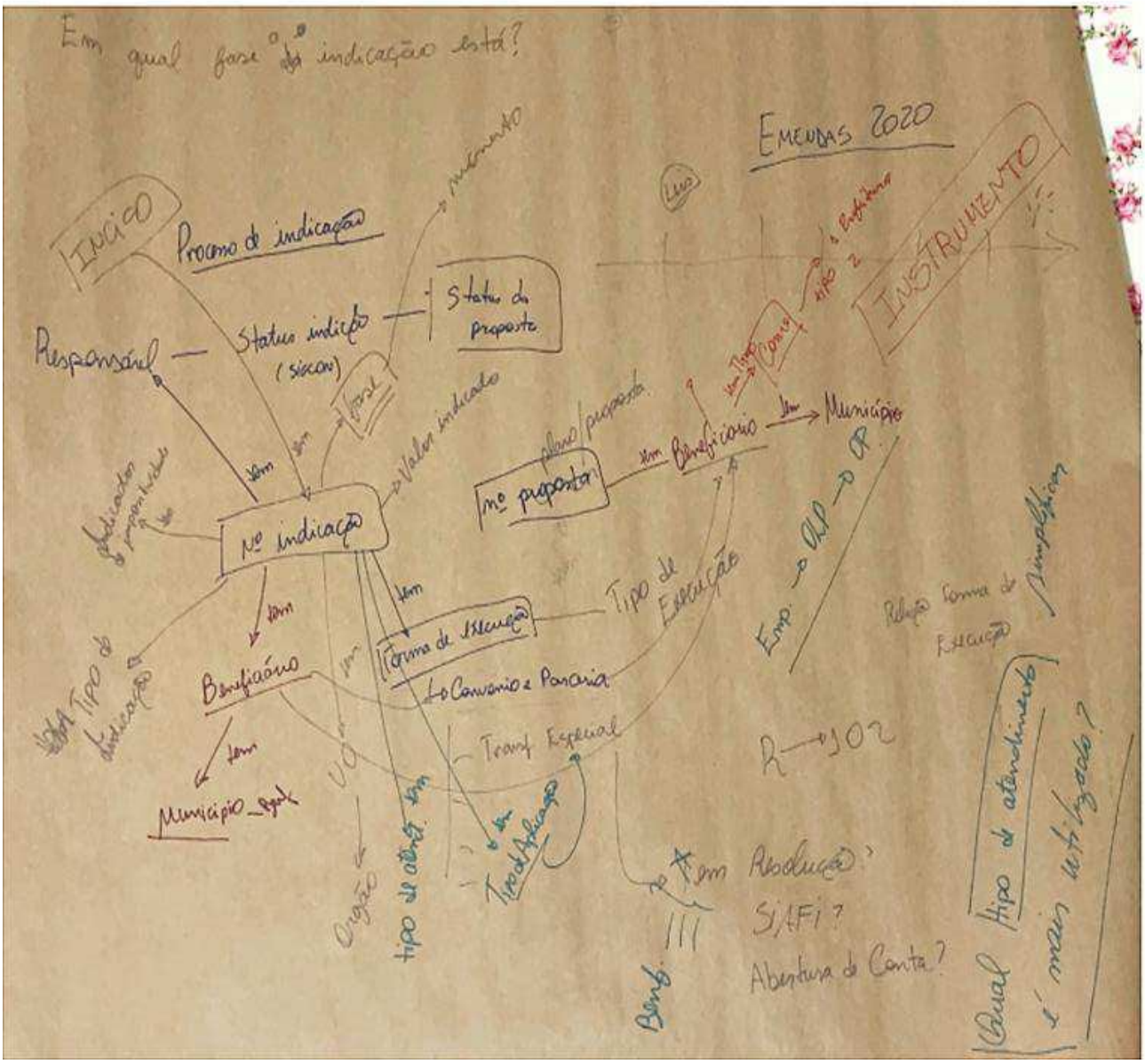
Fonte: elaborada pela autora.

Figura 36 - Mapa conceitual desenvolvido no segundo encontro do grupo focal, no dia 17/11/2021.



Fonte: elaborada pela autora.

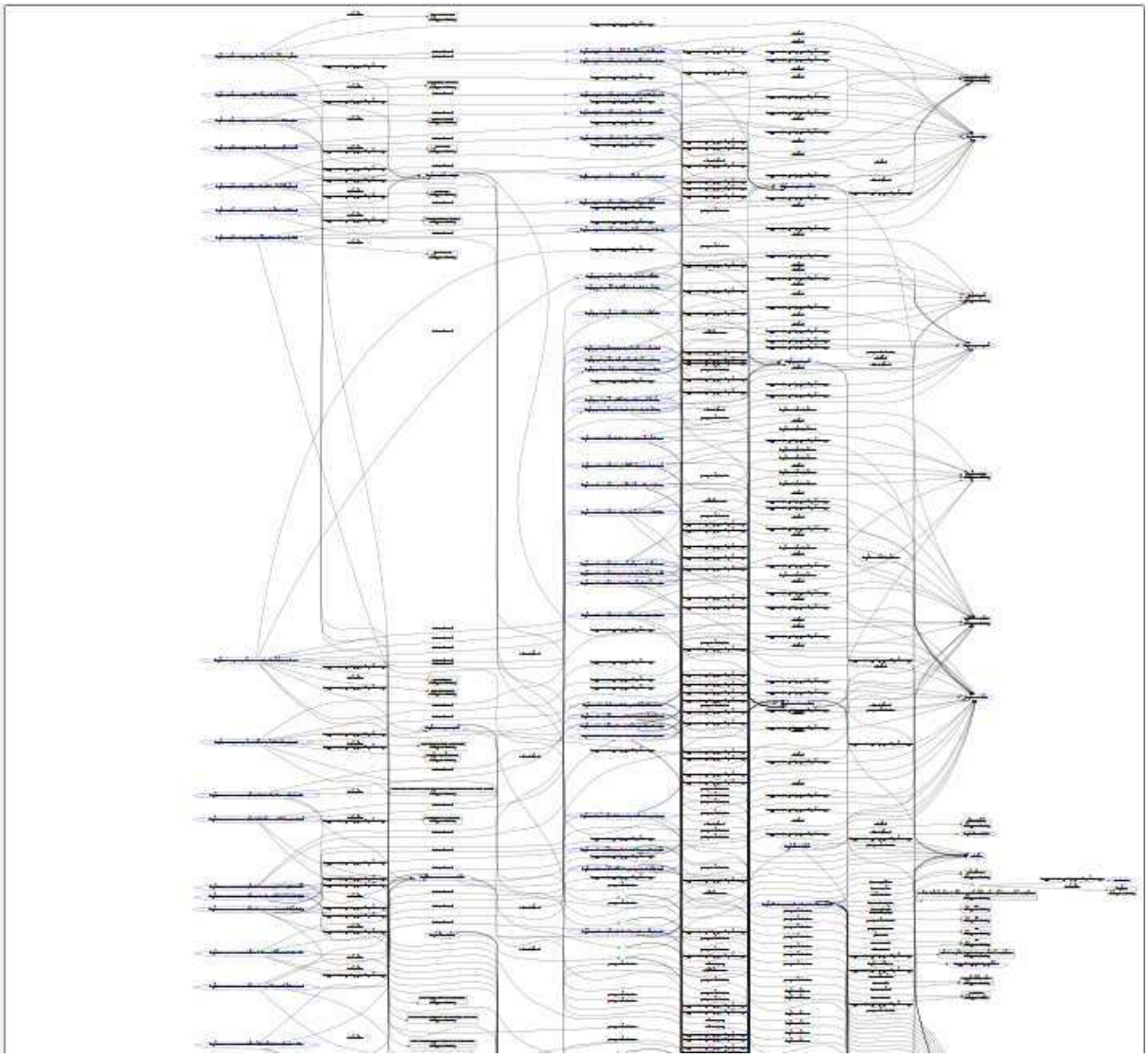
Figura 37 - Mapa conceitual desenvolvido no terceiro encontro do grupo focal, no dia 24/11/2021.



Fonte: elaborada pela autora.

APÊNDICE II – REPRESENTAÇÃO DO GRAFO DE CONHECIMENTO GERADO

O arquivo “freya-kg.trigg” está disponível no GitHub da pesquisa[43].



[43] Disponível em: https://github.com/marci-pires/EmendasParlamentaresMG/tree/main/FinalFiles . Acesso em 19 out 2022.