Um Ambiente para Integração de Dados Abertos relativos à Despesa Pública

Gustavo C. Britto, Fabiano B. Ruy, Carlos L. B. Azevedo

Coordenadoria de Informática, Instituto Federal do Espírito Santo – Campus Serra
ES-010 Km-6,5 – Manguinhos, Serra – ES - Brasil

gustavocbritto@gmail.com, {fabianoruy, carlos.azevedo}@ifes.edu.br

Abstract. *In recent years Brazil has assumed a set of commitments for improving its transparency, especially on the Public Budget. The resulting actions led to the availability of a large amount of public data from the federation, states and municipalities. However, there are still many challenges to deal with all this data in a unified way, seeking for more valuable information. This paper proposes an environment able to capture data from different government entities and store them in a common repository, based on a Public Expense Ontology, for providing uniform queries, regardless of the original data sources.*

Resumo. *O Brasil assumiu nos últimos anos uma série de compromissos visando melhorar sua transparência, especialmente em relação ao Orçamento Público. As medidas decorrentes levaram à disponibilização de um grande volume de dados públicos pela federação, estados e municípios. Entretanto, ainda há muitos desafios para lidar com esses dados de maneira conjunta, buscando obter informações mais relevantes. Este artigo propõe um ambiente capaz de capturar dados de diferentes entidades governamentais e armazená-los em um repositório comum, baseado em uma Ontologia de Despesa Pública, provendo consultas uniformes, independentemente das fontes de dados originais.*

CEUR-WS.org/Vol-2728/paper13.pdf

1. Introdução

O Brasil é membro da parceria internacional *Open Government Partnership* (OGP), em que, por meio de um plano de ação, firmam-se compromissos frente aos desafios que envolvem um governo mais transparente, participativo e colaborativo. No país, a Lei de Acesso à Informação (Lei 12.527/2011) regulamenta a gestão de dados pelas entidades governamentais, bem como sua disponibilização para o público em geral. Atualmente, o Governo Brasileiro está em seu quarto plano de ação [Brasil 2020a], em que define um conjunto de princípios e compromissos de transparência. Dentre os quais, vale destacar a melhoria na disponibilização de dados abertos, o incentivo a estados e municípios a implementar ações de governo aberto e a geração de um ecossistema de dados mais suscetível a participação social e que possa ser bem utilizado pela comunidade científica.

Desde a entrada da lei em vigor, em 2011, um grande volume de dados de interesse público tem sido provido pelas entidades governamentais, sobretudo em seus Portais de Transparência. Entretanto, são muitos os desafios até que essas entidades sejam capazes de disponibilizar seus dados publicamente em conformidade com os compromissos estabelecidos legalmente e em seus planos de ação. Neste contexto, em uma revisão sistemática sobre dados abertos governamentais (OGD – *Open Government Data*), De Oliveira *et al.* (2018) destacam: (i) a heterogeneidade na organização dos dados

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

disponibilizados; (ii) a dificuldade em obtê-los, principalmente atualizados e, ainda; (iii) a viabilidade de processamento dos dados nos formatos em que são liberados.

Parte desse problema decorre do fato de cada entidade do Governo, nas diferentes esferas Federal, Estadual e Municipal, ser responsável por publicar seus próprios dados, como, por exemplo, determina a Lei Complementar 131/2009, relativa ao Orçamento Público. Isso leva a ambientes geralmente independentes, com variações na organização, formato, meio de acesso, além do escopo dos dados publicados em cada portal. Quando há necessidade de trabalhar com os dados de diferentes entidades conjuntamente, seja para rastreamento, busca por inconsistências, indícios de fraude, ou mesmo comparações, é preciso lidar com um problema de interoperabilidade. Dessa maneira, uma importante questão a ser tratada é possibilitar a utilização conjunta desses dados, favorecendo uma interpretação mais coerente, o que perpassa desde uma representação consistente dos dados até meios de consulta uniformes, independentemente da fonte.

Uma abordagem que tem obtido sucesso em relação à representação, é o uso de Ontologias para criar um modelo comum para os dados [Santos *et al.* 2018] [De Giacomo *et al.* 2018] [Cruz e Xiao 2005]. Ontologias podem representar consistentemente o conhecimento de um domínio, por meio de seus conceitos e relações, indicando a interpretação desejada ao domínio, independentemente das aplicações específicas pretendidas. Neste contexto, uma importante aplicação é seu uso como referência semântica para a integração de dados [Cruz e Xiao 2005] [Gruber 1991], promovendo a interoperabilidade de dados a partir de uma base comum para interpretação e redução de inconsistências conceituais [Campos, 2019] [Guarino, 1998]. De fato, uma ontologia no domínio de Despesa Pública pode servir como base conceitual, dando suporte à criação de um ambiente integrado com um repositório comum abrigando, de forma homogênea, dados obtidos de diferentes fontes. Isso facilita a realização de consultas, provendo uniformidade para a organização e recuperação dos dados ao indicar sua semântica.

Além disso, ainda é preciso obter os dados dos diversos portais de transparência para popular este ambiente. De Oliveira *et al.* (2018) apontam a escassez de estudos abordando a obtenção automática dos dados diretamente da fonte, geralmente exigindo o *download* para que sejam, então, trabalhados. Uma alternativa é a utilização de mecanismos de captura, como os baseados em *web scraping*, para navegar automaticamente pelos portais obtendo os dados de interesse. Outro desafio é, após capturar os dados, classificá-los e armazená-los corretamente em um repositório comum.

Este trabalho propõe um ambiente capaz de representar uniformemente as informações relativas à Despesa Pública obtidas de diferentes portais de transparência disponibilizados pelas entidades governamentais. O ambiente captura os dados de um conjunto selecionado de entidades e, baseado em uma Ontologia de Despesa Pública (nas versões de referência e operacional), os armazena em um repositório comum, provendo serviços de consulta homogêneos, a partir de um mesmo local, facilitando análises e comparações sobre os dados das fontes selecionadas.

O artigo está assim organizado: a seção 2 descreve as principais noções do domínio de Despesa Pública; a seção 3 apresenta o desenvolvimento do ambiente proposto; a seção 4 discute os trabalhos correlatos; e a seção 5 as considerações finais.

2. Execução de Despesa no Orçamento Público

No contexto de dados abertos governamentais (OGD), um aspecto relevante para a promoção da transparência é o entendimento acerca dos gastos efetuados pelo governo, envolvendo o Orçamento Público e, mais especificamente, execução da Despesa Pública.

O **Orçamento Público** é o instrumento de planejamento feito pelo governo para estimar a arrecadação no decorrer do ano subsequente e autoriza um limite a ser gasto com esses recursos [Brasil 2020a]. O processo orçamentário é constituído por quatro etapas: (i) elaboração, (ii) aprovação, (iii) execução e (iv) controle. Este trabalho tem interesse nos dados disponibilizados referentes a etapa (iii), de **execução orçamentária**. Neste domínio há uma variedade de conceitos descritos no Manual Técnico Orçamentário - MTO [Brasil 2020b], que são apresentados resumidamente no Quadro 1.

Quadro 1. Dicionário de Orçamento Público [Brasil 2020b]

Conceito	Descrição
Unidade Orçamentária	Entidades públicas com o papel de coordenar o processo orçamentário, integrando e proporcionando o trabalho de suas unidades administrativas.
Órgão	Unidades responsáveis por uma função de governo.
Função	Maior nível de agregação das diversas áreas de atuação do setor público.
Subfunção	Representa um nível mais detalhado da função.
Elemento Despesa	Responsável por identificar os objetos de gasto. Por exemplo, diárias, material de consumo, serviços de terceiros prestados sob qualquer forma, etc.
Ação	Nível operacional - Operação (ação) da qual resultam produtos (bens ou serviços) que contribuem para atingir o objetivo de um programa.
Programa	Nível Estratégico - Composto por um conjunto de ações (orçamentárias e não-orçamentárias) para enfrentar um problema.
Subtítulo	Usados especialmente para identificar a localização física da ação orçamentária.
Contrato	Acordo entre as partes que definem direitos e obrigações.
Despesa	Gastos realizados pelo governo para funcionamento e manutenção dos serviços públicos prestados à sociedade.

Despesa Pública é definida por meio de um sistema de classificação estruturado [Brasil 2020b], que permite sua identificação quanto a importância e o monitoramento do destino do dinheiro público. Ela pode ser dividida em blocos de classificação, respondendo perguntas a respeito de uma determinada despesa, como mostra o Quadro 2.

Quadro 2. Classificações - adaptado do MTO [Brasil 2020b]

Bloco da Estrutura	Item da Estrutura	Pergunta a ser Respondida
Classificação por Esfera	Esfera Orçamentária	Em qual Orçamento?
Classificação Institucional	Órgão	Quem é o responsável por fazer?
	Unidade Orçamentária	
Classificação Funcional	Função	Em que áreas de despesa a ação governamental será realizada?
	Subfunção	
Estrutura Programática	Programa	O que se pretende alcançar com a implementação de Política Pública?
Informações Principais da Ação	Ação	O que será desenvolvido para alcançar o objetivo do programa?
	Descrição	O que é feito? Para que é feito?
	Forma de Implementação	Como é feito?
	Produto	O que será produzido ou prestado?
	Unidade de Medida	Como é mensurado?
	Subtítulo	Onde está o beneficiário do gasto?
Programação Financeira	Categ. Econômica Despesa	Qual o efeito econômico da realização da despesa?
	Grupo Natureza de Despesa	Em qual classe de gasto será realizada a despesa?
	Modalidade de Aplicação	De que forma serão aplicados os recursos?
	Elemento de Despesa	Quais os insumos a utilizar ou adquirir?
	Fonte de Recursos	De onde virão os recursos para a despesa?
	Dotação	Qual o montante alocado?

A Figura 1 apresenta um exemplo de uma despesa classificada, incluindo sua codificação.

CÓDIGO COMPLETO*		10.	39.	252.	26.	782.	2075.	7M64.	0043.
Esfera: Orçamento Fiscal		10							
CLASSIFICAÇÃO INSTITUCIONAL	Órgão: Ministério da Infraestrutura		39						
	Unidade Orçamentária: Departamento Nacional de Infraestrutura de Transportes - DNIT			252					
CLASSIFICAÇÃO FUNCIONAL	Função: Transporte				26				
	Subfunção: Transporte Rodoviário					782			
CLASSIFICAÇÃO PROGRAMÁTICA	PROGRAMA: Transporte Terrestre						2075		
	AÇÃO: Construção de Trecho Rodoviário							7M32	
	SUBTÍTULO: Paraíba								0043

Figura 1. Exemplo estrutura (parcial) - Manual Técnico do Orçamento (2020)

De acordo com a Lei 4.320/64, a execução da despesa pública deve passar por três estágios: empenho, liquidação e pagamento. No **empenho**, o governo assume o compromisso de contratar e realizar o gasto reservando o dinheiro para tal. Nesta etapa é criada uma nota de empenho, documento que indica o credor, a representação e a importância da despesa. Na **liquidação**, será feita a verificação se o produto, serviço ou obra cumpriu toda a obrigação descrita no empenho. No último estágio, o governo realiza, enfim, o **pagamento** ao credor pelo serviço ou produto fornecido.

Licitação é a forma com que a Administração Pública pode realizar compras e vendas de produtos e serviços, conforme define a Lei 8.666/63. É lançado um edital referente a uma licitação com o objeto bem definido, prazos e condições gerais de participação. Os interessados enviam propostas e, ao final do processo, é feito o julgamento e classificação, sendo declarado vencedor da licitação aquele que apresentar proposta mais vantajosa; e, assim, é chamado para a assinatura do contrato.

3. Ambiente de Integração

O ambiente proposto visa lidar com os dados de despesa pública, desde sua captura e armazenamento até a disponibilização para consultas. A Figura 2 apresenta os principais componentes de sua arquitetura: (A) um Modelo Comum capaz de representar consistentemente os dados pretendidos, e dar origem ao repositório a ser populado; (B) Mecanismos de Captura para obter os dados conforme são disponibilizados pelas entidades governamentais e armazená-los no repositório comum; e (C) Meios de Consulta para permitir acesso uniforme aos dados pelos usuários.

O **modelo comum** foi criado com base em uma Ontologia de Referência no domínio de Despesa Pública. Tal ontologia tem o papel de representar consistentemente os conceitos e relações do domínio, de forma a suportar os dados de despesa pública de diferentes fontes, servindo como referência semântica do ambiente. Esta referência é fundamental tanto para uma compreensão clara do domínio abordado como para orientar e fornecer os termos e sentidos adequados para a captura, armazenamento dos dados e realização das consultas. A partir da ontologia de referência, é derivada uma ontologia operacional para guiar o armazenamento dos dados obtidos nos portais de transparência. Além de viabilizar o armazenamento padronizado, independentemente da fonte, a ontologia operacional dá origem ao repositório que pode ser computacionalmente processado, permitindo consultas e inferências sobre os dados. Ver bloco A na Figura 2.

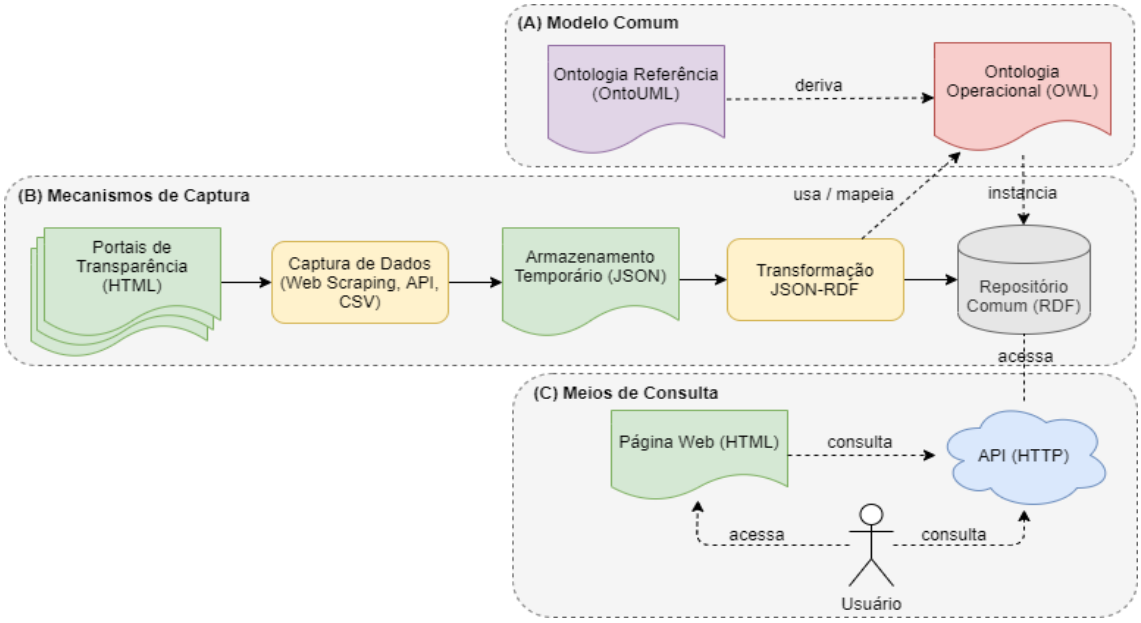

Figura 2. Arquitetura do Ambiente

Os **mecanismos de captura** são construídos para obter os dados dos portais de transparência, fontes oficiais de divulgação, e armazená-los no repositório do ambiente. Para cada portal pode ser elaborado um *plugin* que captura os dados usando técnicas de *web scraping*. Esta atividade também pode ser feita por outros meios, tais como por APIs ou *download* em formatos específicos (e.g. CSV), conforme provido pelos portais. Um mapeamento dos tipos de dados disponíveis em cada portal para os conceitos da ontologia favorece a compreensão dos dados e indica quais porções devem ser capturadas. Os dados são armazenados temporariamente em formato JSON em um esquema similar ao de cada fonte de dados. Em seguida, são tratados para limpeza e ajustes necessários, e transformados em triplas RDF conforme o mapeamento com a ontologia. Assim, os dados disponíveis em cada portal específico ganham uma nova representação, agora padronizada, em um modelo comum, conforme a ontologia. As triplas geradas formam o repositório comum do ambiente com os dados obtidos. Ver bloco B na Figura 2.

Os **meios de consulta** são acessados por uma página web. Um usuário ou agente externo pode acessar o repositório criado com a ontologia e o conjunto completo de triplas (OWL+RDF) e realizar consultas específicas em SPARQL, ou acessar consultas pré-estabelecidas navegando na página. Vale destacar que no ambiente as consultas possuem o mesmo formato, independentemente da entidade governamental a ser pesquisada, sendo possível, inclusive, obter e comparar dados de entidades distintas na mesma consulta. Ver Bloco C na Figura 2. Detalhes dos componentes do ambiente são descritos a seguir.

3.1. Ontologia de Despesa Pública

A Ontologia de Despesa Pública constitui a base semântica do ambiente, atuando em duas versões: **de referência**, buscando representar fielmente a conceituação do domínio de despesa pública e prover a fundamentação ao ambiente como um todo; e **operacional**, para permitir a população dos dados em um formato padronizado, uniformizando o armazenamento, processamento e consultas.

A Ontologia de Referência foi desenvolvida com base no método SABiO [Falbo 2014], tomando como fonte as Leis 4.320/64 (controle de orçamento) e 8.666/93 (licitações e contratos) e seus materiais derivados; ontologias relacionadas, tais como [Bassetti *et al.* 2014] e [Araújo, *et al.* 2015]; e análises dos dados disponíveis nos portais de transparência. O seu escopo priorizou os conceitos e relações mais frequentemente usados nos portais pesquisados, para viabilizar consultas mais úteis e consistentes. Assim, a versão da ontologia aqui apresentada não tem intenção de ser extensiva / completa, mas sim de representar o universo de discurso necessário a este trabalho.

A ontologia é fundamentada em UFO e modelada em OntoUML, visando maior expressividade na representação do domínio e facilidades na transformação para a ontologia operacional. UFO é uma ontologia de fundamentação que define distinções úteis para compreender e representar um domínio. São providas distinções básicas tais como sortais rígidos (*kind*) e antirrígidos (*role*, *phase*); mediadores em relações materiais (*relator*), além de não sortais para generalizações (*category*, *rolemixin*). A linguagem OntoUML captura essas distinções em uma extensão da UML, e tem sido usada na construção de modelos conceituais em diversos domínios [DoD 2011] [Pergl *et al.* 2013]. UFO e OntoUML são descritas em detalhes em [Guizzardi 2005] [Guizzardi *et al.* 2018].

O modelo está organizado nas subontologias de ***Entidade Governamental***, ***Licitação e Contrato*** e ***Execução de Despesa***. A Figura 3 mostra a subontologia de Entidade Governamental, representando as entidades públicas simplificadamente. A categoria **Governo** generaliza as três esferas da federação: **Governo Federal**, **Governos Estaduais** e **Governos Municipais**. Estes são compostos por **Órgãos** que, por sua vez, podem assumir o papel de **Unidades Gestoras** quando responsáveis por gerir os recursos financeiros públicos por meio de **Processos**.

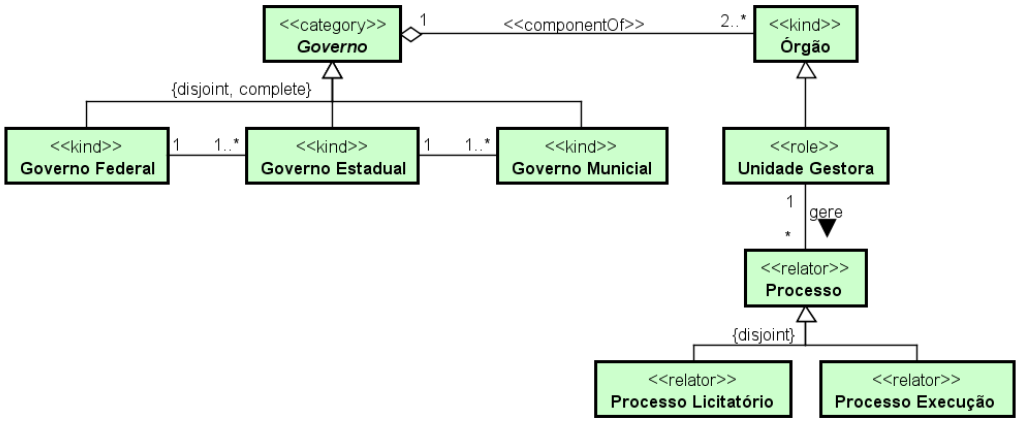

Figura 3. Subontologia de Entidade Governamental

Processos administrativos são abertos iniciar o procedimento de realização de gastos públicos. Tipicamente, um **Processo Licitatório**, dá origem a uma licitação, ou sua dispensa, para se estabelecer um contrato que contemple os gastos pretendidos. Pode ocorrer ainda um **Processo de Execução**, que permite a realização de despesas diretamente, sem contrato ou licitação, em casos de emergência ou calamidade pública, serviços de baixo custo, e outros itens descritos nos Art. 24 e 62 da Lei 8.666/93.

A Figura 4 descreve Licitações e Contratos. Uma **Ação Licitatória** ocorre com o objetivo de se estabelecer um **Contrato** da Unidade Gestora com um **Credor**. Tal ação pode ser uma **Licitação**, ou uma **Dispensa de Licitação**, conforme os casos previstos em

lei, dando origem aos respectivos **Contratos**. Uma vez estabelecido o **Contrato** com um **Credor** (seja uma Pessoa ou Empresa), as despesas podem ser executadas. Como exemplo, a Licitação *20/2018*, estabelecida pelo Processo *15525/2018* gerido pela *Prefeitura da Serra*, dá origem ao Contrato *111/2018*, relacionando os Empenhos *73/2020* e *183/2019*, cujo Credor é a empresa *Cac Comercial Ltda*. As etapas do processo licitatório foram omitidas no diagrama, pois o foco não é o fluxo do processo, mas o seu resultado, capaz de gerar despesa, tal qual está disponível nos portais de transparência.

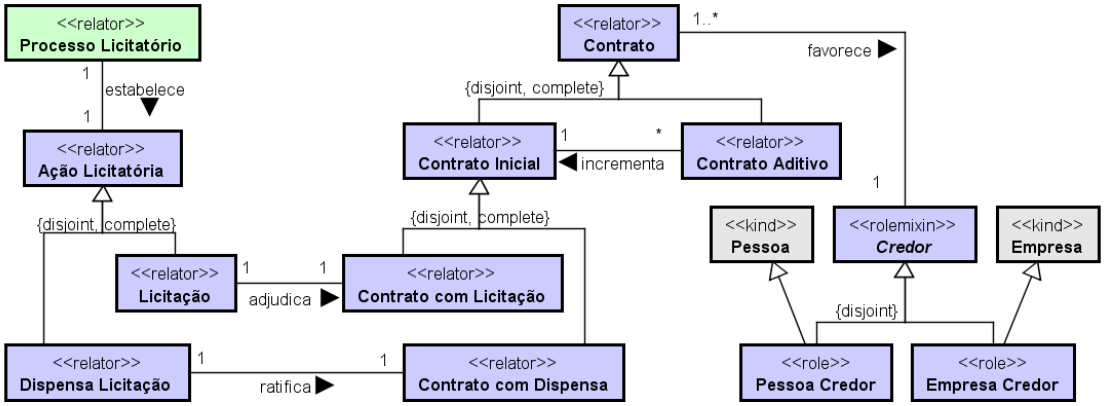

Figura 4. Subontologia de Licitação e Contrato

Os atos de **Empenho**, **Liquidação** e **Pagamento** de uma **Despesa**, generalizados como **Execução Despesa**, são registrados mediante um **Contrato** (ou **Processo de Execução**, nas exceções mencionadas) da Unidade Gestora, beneficiando um **Credor**, conforme mostra a Figura 5.

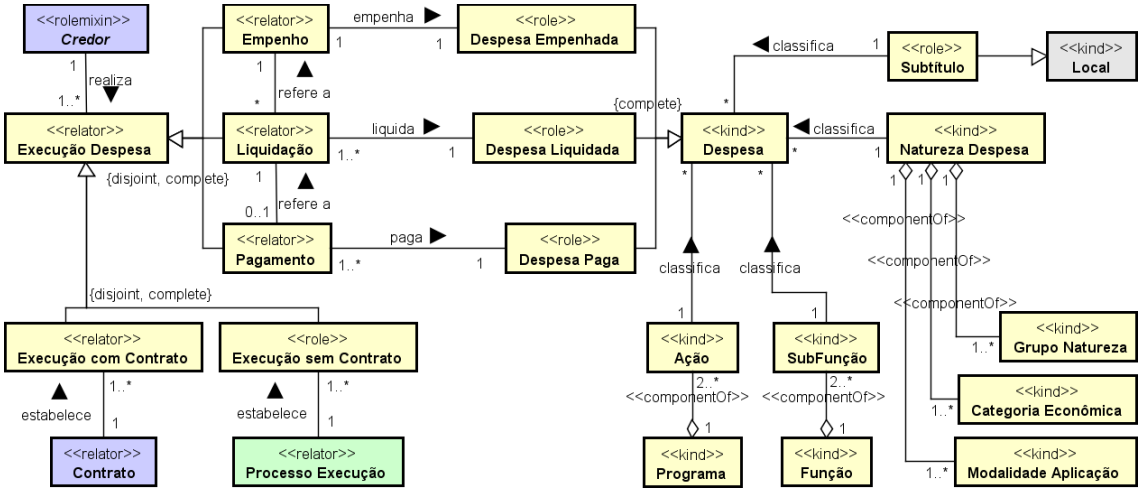

Figura 5. Subontologia de Execução de Despesa

Cada registro dos atos de **Empenho**, **Liquidação** ou **Pagamento** eventualmente altera a classificação da **Despesa** (para **Despesa Empenhada**, **Despesa Liquidada** e/ou **Despesa Paga**), que pode estar desempenhando um, dois ou os três papeis ao mesmo tempo. Por exemplo, quando uma dada despesa está empenhada, mas já foi parcialmente liquidada e paga (no exemplo anterior, no Empenho *73/2020*, dos *R$ 4.482,00 empenhados*, apenas *R$ 1.892,40* foram *liquidados* e *pagos* até maio/2020). Assim, toda **Despesa** é criada como uma **Despesa Empenhada**, mas pode assumir também as demais classificações conforme os atos de **Liquidação** e de **Pagamento** são registrados. Além

disso, a **Despesa** pode ser classificada com diversas informações: quanto à **Ação**, que faz parte de um **Programa**, **SubFunção**, que faz parte de uma **Função**, **Subtítulo**, e **Natureza de Despesa**, que é composta por **Categoria Econômica**, **Grupo de Natureza** e **Modalidade de Aplicação**, conforme descritos na Seção 2. Por exemplo, o Empenho *73/2020*, está associado à Ação *1284 - Serra Mais Você*, Programa *70 - Serra Mais Participativa*, SubFunção *122 - Administração Geral*, Função *4 - Administração*, Categoria Econômica *3.0.00.00 - Despesas Correntes*, Grupo de Natureza *3.3.00.00 - Outras Despesas Correntes*, Modalidade de Aplicação *3.3.90.00 - Aplicações Diretas*.

A Ontologia Operacional é derivada a partir da ontologia de referência, realizando-se algumas tarefas de projeto (*design*) para adequar o modelo para a implementação e uso pretendidos. A linguagem de destino definida é OWL, por permitir maior diversidade de ferramentas e facilidades de consulta. Assim, o modelo, em OntoUML, passa por simplificações, como a redução de algumas classes e relacionamentos, para possibilitar melhor correspondência com os dados obtidos, além de otimizar as consultas. Por exemplo, foram criadas relações derivadas (e.g. **Processo** *estabelece* **Execução Despesa**) e as hierarquias referentes a **Processo**, **Licitação**, **Contrato** e **Credor** foram reduzidas a uma classe cada, com atributos para determinar seus tipos. Tal simplificação, no entanto, não foi aplicada às hierarquias referentes a **Execução de Despesa** e **Despesa**, pois implicaria na redução da capacidade de representar informações relevantes. Esse processo tende a causar perdas semânticas em relação à ontologia de referência, que visa representar fielmente o domínio do problema. Entretanto, a ontologia operacional resultante está mais focada em garantir que os dados tipicamente disponíveis nos portais sejam adequadamente representados e processados.

A partir do modelo adaptado (*design*), a Ontologia Operacional foi gerada como um arquivo OWL utilizando a ferramenta Menthor (https://github.com/MenthorTools/menthor-editor). Assim, a operacionalização ocorre, permitindo a criação das instâncias segundo a ontologia desenvolvida, mantendo as relações definidas entre os conceitos, e podendo ser interpretada por agentes computacionais e sofrer consultas.

3.2.Obtenção de Dados

Antes de capturar os dados providos pelas entidades governamentais, é preciso compreender as informações disponíveis e como serão organizadas no repositório comum. Inicialmente é necessária uma análise do site ou arquivo disponível para compreender sua estrutura. Em seguida, técnicas de mapeamento vertical [Calhau e Falbo 2010] [Ruy 2017] são aplicadas para associar os tipos de dados da fonte aos conceitos da ontologia. Dessa forma, o mapeamento é feito observando-se a semântica dos dados (e não apenas os termos utilizados) para identificar a quais classes as instâncias serão atribuídas posteriormente no repositório. Assim, além de atribuir ao dado a semântica definida pela ontologia de domínio, o mapeamento também facilita a criação dos algoritmos de busca e transformação dos dados.

O Quadro 3 demonstra o mapeamento realizado para quatro entidades governamentais. Foram escolhidas entidades de diferentes esferas para experimentar uma maior variação dos tipos de dados. Além do Brasil, na esfera federal, foi escolhido o estado do Espírito Santo por ser avaliado com bom grau de transparência[1], e dois

[1] Open Knowledge Brasil: https://transparenciacovid19.ok.org.br/

municípios de sua região metropolitana: Serra e a capital, Vitória. As entidades selecionadas possuem portais de transparência em efetivo funcionamento, provendo os dados necessários para lidar com despesa pública neste trabalho.

Quadro 3. Mapeamento dos Tipos de Dados para a Ontologia de Despesa Pública

Conceito da Ontologia	Termo Esfera Federal (Brasil)	Termo Esfera Estadual (ES)	Termo Esfera Municipal (Vitória/Serra)
Órgão	Órgão/Entidade Vinculada	Órgão	Órgão
Unidade Orçamentária	Unidade Orçamentária	Unidade Gestora (disponível apenas no site)	Unidade Orçamentária
Processo	Processo	Processo	Processo
Licitação	Licitação	Licitação	Licitação
Contrato	Número do Contrato	Número do Contrato	Número do Contrato
Credor	Dados do Favorecido	Favorecido	Beneficiário
Empenho	Nº do documento (em Documento de Empenho)	Documento	Empenho
Liquidação	Nº do documento (em Documento de Liquidação)	Documento	Liquidação
Pagamento	Nº do documento (em Documento de Pagamento)	Documento	Número do Pagamento
Despesa	Detalhamento do Gasto	Descrição da Despesa	Descrição
Ação	Ação	Ação	Ação de Governo
Programa	Programa	Programa	Programa de Governo
SubFunção	Subfunção	Subfunção	Sub Função
Função	Área de Atuação (função)	Função	Função
SubTítulo	Subtítulo (localizador)	Subtítulo	Subtítulo/Localização
Categoria Econômica	Categoria da Despesa	Categoria Econômica	Categoria Econômica
Grupo Natureza	Grupo de Despesa	Grupo de Despesa	Grupo de Natureza da Despesa
Modalidade Aplicação	Modalidade de Aplicação	Modalidade de Aplicação	Modalidade de Aplicação

A partir do mapeamento e da compreensão da estrutura de cada portal almejado, foram construídos *plugins* de captura de dados aplicando *web scraping*. Esta técnica usa agentes computacionais para navegar em páginas web e coletar dados, mantendo sua estrutura [Reitz e Schlusser 2016], e é comumente utilizada quando os dados são disponibilizados em formatos pouco ou não estruturados. A cada adição de nova fonte de dados um *plugin* é criado, desenvolvendo uma parte específica dos *scripts* de captura e reutilizando uma parte genérica parametrizável, com base em princípios de navegação e na semântica e mapeamentos com a ontologia. Quando há portais de entidades distintas utilizando o mesmo sistema de disponibilização de dados, *plugins* podem ser reutilizados. Vale mencionar que nos portais utilizados como exemplo neste trabalho, as outras opções disponíveis de obtenção dos dados (por *download* ou API) não forneciam os mesmos dados que estão acessíveis na página do portal, deixando de fora informações relevantes (como números de notas de empenho, liquidação e pagamento das despesas). Assim, é importante selecionar o modo de extração que forneça o conjunto de dados mais apropriado para popular o repositório.

O mecanismo de captura armazena temporariamente os dados em formato JSON seguindo uma estrutura similar à do site de origem, para, então, convertê-los em triplas RDF com um *script* que usa a biblioteca RDFLIB (https://rdflib.readthedocs.io/). O *script* implementa o mapeamento para garantir que os dados a serem instanciados nas triplas respeitem o esquema estabelecido em OWL, sendo, assim, reflexo da ontologia. Os dados em formato de triplas RDF podem, então, ser consumidos por agentes computacionais.

A Figura 6 mostra, como exemplo, o mesmo dado da classe **Execução de Despesa** representado, primeiro, em JSON e, então, transformado em triplas RDF.

Figura 6. Exemplo de Transformação de Dado JSON para RDF

Ao final de cada atividade de captura, o repositório é populado com os dados do portal selecionado. À medida que mais portais são selecionados, o repositório cresce em volume de dados e possibilidades de consultas, formando o repositório comum almejado. Uma vez criado o *plugin* de captura para determinado portal, este pode ser executado sempre que for conveniente ou de forma agendada, atualizando os dados no repositório.

3.3. Consultas

As consultas ao repositório ficam acessíveis por meio de uma página web, de duas formas principais. A primeira, voltada ao cidadão comum, oferece consultas já preparadas, respondendo a questões típicas, pré-estabelecidas. O repositório também pode ser acessado por um *endpoint*, para que especialistas, com conhecimento em SPARQL, possam explorar o conjunto de dados de forma livre. Os dados de diferentes entidades governamentais estão representados da mesma forma no repositório, possibilitando aplicar a mesma consulta (*query*) para quaisquer entidades da base. Além disso, uma única consulta pode buscar informações de diferentes entidades públicas, o que facilita comparações, evitando o esforço de buscar nos diversos portais e em variados formatos.

Outra possibilidade é a realização de consultas mais elaboradas buscando nos campos textuais (e.g. descrição da despesa), o que não é frequentemente disponibilizado pelos portais de transparência. Por exemplo, considere o Empenho *1719/2020* da *Prefeitura de Vitória*, referente a *Cobrir despesas com aquisição de álcool gel*. Com a busca padrão oferecida no portal da prefeitura, só é possível filtrar os empenhos por elemento de despesa do tipo *Material de Consumo*. Já no ambiente, realizando a consulta pelo texto dos campos, pode-se filtrar os empenhos cuja descrição contenha, por exemplo, expressões como “álcool gel”, “máscara” ou “respirador”. Além disso, é possível realizar a busca por informações de todas as entidades governamentais presentes no repositório, afim de fazer um comparativo de informações das despesas (como preço, unidades, fornecedores etc.), facilitando uma análise sobre quais entidades estão investindo mais em determinado produto ou serviço, ou pagando menos por esses itens.

Como exemplo, o código SPARQL a seguir realiza uma consulta agrupando despesas, no ano de 2020, que contenham “álcool” em sua descrição. A Figura 7 apresenta o resultado, considerando três entidades governamentais. O ambiente é flexível para consultas, permitindo, por exemplo, comparações entre diferentes entidades, com anos anteriores, ou verificando a relação do gasto com a quantidade de habitantes. De fato, a partir da captura e representação uniforme dos dados no repositório, as possibilidades de busca por meio das consultas são inúmeras.

```
PREFIX ontoDesp: <http://localhost/ontoDesp#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?ano ?govName (sum(xsd:decimal(?valorEmpenho)) AS ?valorEmpenhado) where {
   ?gov rdfs:label ?govName .
   ?orgao ontoDesp:temGoverno ?gov .
   ?orgao ontoDesp:temUnidadeGestora ?ug .
   ?orgao rdfs:label ?orgaoDesc .
   ?ug ontoDesp:temProcesso ?processo .
   ?processo ontoDesp:temExecucao ?empenho .
   ?empenho ontoDesp:empenha ?desp .
   ?desp ontoDesp:temDescricao ?descricao .
   ?desp ontoDesp:temValorEmpenho ?valorEmpenho .
   ?desp ontoDesp:temAno ?ano .
   FILTER ( regex(?descricao, "alcool", "i") || regex(?descricao, "álcool", "i") )
} group by ?govName ?ano limit 1000
```

#	Entidade	Valor Empenhado
1	Espírito Santo	R$ 27.402.118,42
2	Vitória	R$ 129.823,50
3	Serra	R$ 78.242,30

Figura 7. Busca por Despesas específicas em Entidades distintas.

4.Trabalhos Correlatos

De Oliveira *et al.* (2018) relatam em seu mapeamento sistemático a dificuldade de encontrar trabalhos contemplando todos os aspectos presentes no ambiente aqui proposto. Em um trabalho com escopo semelhante, Tosin *et al.* (2016) propõem "uma solução para interligar bases de dados isoladas e disponibilizar estes dados de forma aberta e conectada". O trabalho almeja resolver o desafio de captura de dados abertos, incluindo os governamentais, e sua disponibilização para consulta, em uma base unificada, usando anotação semântica. A captura dos dados, realizada via APIs, embora tenha a vantagem de demandar menos codificação, nem sempre é interessante para dados governamentais, pois algumas entidades suprimem dados via API, deixando-os apenas nas páginas web. O método usado no presente trabalho, adotando também *web scraping*, permite acesso mais amplo aos dados, e resolve este problema de quantidade e qualidade diferente de dados de acordo com o meio de acesso.

Como modelo comum, Tosin *et al.* (2016) adotam uma ontologia e a mapeiam para um repositório de triplas para posterior consulta. Porém, tal ontologia é descrita somente como uma anotação semântica, e o mapeamento é realizado majoritariamente pela proximidade da nomenclatura, ansiando-se automatizar o processo com base na similaridade dos nomes dos elementos. Tal solução poderia causar o problema do Falso Acordo (*False Agreement Problem*), quando dados semanticamente diferentes, mas com nomenclatura similar, são mapeados para o mesmo item semântico [Guarino 1998]. Isso pode gerar interpretações diferentes da originalmente desejada nos dados consolidados. No presente trabalho, foi aplicada uma ontologia de referência, e sua versão operacional, com mapeamento focado na semântica dos dados, como forma de minimizar tal problema.

Outro trabalho que se baseia em dados abertos governamentais, também com o foco em disponibilizar uma consulta integrada é descrito em [Monteiro e Gálvez 2012], abordando dados sobre vulnerabilidade social. Assim como em [Tosin *et al.* 2016],

utilizaram somente uma ontologia operacional para basear a integração de dados, com menor foco em sua semântica. Tal escolha pode ser justificada pelo uso de poucas fontes de dados e oriundas somente do governo federal. Porém, um tratamento semântico é crucial, especialmente quando se busca integrar informações de fontes mais distintas, como entidades governamentais em suas três esferas.

Por fim, Bassetti *et al.* (2014) realizam um estudo de caso sobre a integração de dados governamentais utilizando uma ontologia de organizações. Diferentemente dos anteriores, este também utiliza uma ontologia de referência, realizando um mapeamento para grafos em RDF e posterior possibilidade de consultas. Porém, os autores adotam um escopo mais fechado, sem a intenção de prover um ambiente expansível, o que difere da arquitetura aqui adotada, em que novas entidades governamentais podem ser adicionadas, a depender da construção de novos *plugins* para os mecanismos de captura.

5. Considerações Finais

Dados governamentais sobre despesas públicas são frequentemente publicados de forma não estruturada ou em formatos não padronizados. Como cada entidade do governo é responsável individualmente por publicar seus dados, comumente as diferentes bases de dados utilizam formatos distintos, havendo pouca padronização entre elas.

Para abordar o problema de interoperabilidade gerado, este trabalho propõe um ambiente pautado em três tópicos: captura, representação consistente e consultas aos dados. A captura dos dados nas fontes diversas pode ser realizada de diferentes formas (e.g. via *web scraping*, API ou *download*), conforme a que proporcionar maiores benefícios em termos de acesso e qualidade do conteúdo obtido. Uma representação consistente é obtida a partir da Ontologia de Despesa Pública (desenvolvida nas versões de referência e operacional), para prover semântica ao ambiente, apoiando a captura, armazenamento e consultas, além de orientar esforços na expansão do mesmo. O mapeamento das fontes de dados para a ontologia favorece a compreensão dos dados e guia os esforços de captura. Os dados são então armazenados em um repositório comum, de maneira uniforme, independentemente de suas fontes. Finalmente, as consultas são realizadas em um ambiente homogêneo, permitindo maior flexibilidade e poder de acesso aos dados, incluindo o cruzamento das informações pretendidas.

O ambiente tem arquitetura modular e é passível de extensão, inclusive por esforços de terceiros. Seu núcleo (ontologia, repositório e estruturas de captura e consultas) se mantém estável enquanto *plugins* para novas fontes de dados podem ser adicionados. Como trabalhos futuros, pretende-se evoluir a ontologia; ampliar as entidades governamentais alcançadas; disponibilizar o ambiente e adaptar a arquitetura para facilitar a expansão por interessados; liberar o site com meios de consulta mais intuitivos, como parametrização e seleção de campos de pesquisa; além de prover resultados visuais mais elaborados, usando gráficos. Isso permitiria um uso mais amplo pelo cidadão interessado sem conhecimentos técnicos de computação.

Ademais, a arquitetura do ambiente poderia ser reutilizada em outros contextos que demandem integração semântica de dados abertos disponibilizados por diferentes fontes. Naturalmente, uma nova ontologia e adaptações para o domínio específico seriam necessárias. Além dos dados governamentais, é possível observar demanda crescente por integração de dados, por exemplo, nas áreas de saúde (e.g. durante a pandemia de COVID-19) e financeira (para acompanhamento de índices e cotações de ativos diversos).

Referências

Araújo, L.S.O, Santos, M.T. and Silva, D.A. (2015) The Brazilian federal budget ontology: a semantic web case of public open data. In Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems, pp. 85-89.

Brasil, Ministério da Transparência e Controladoria-Geral da União, Secretaria de Transparência e Prevenção da Corrupção, Diretoria de Transparência e Controle Social, Coordenação Geral de Governo Aberto e Transparência, (2020a), "4º Plano de Ação Nacional em Governo Aberto", Disponível em: https://www.gov.br/cgu/pt-br/governo-aberto/noticias/2018/4o-plano-de-acao-brasileiro-e-lancado-em-reuniao-com-coordenadores/4o-plano-de-acao-nacional_portugues.pdf/view, Junho/2020.

Brasil, Ministério da Economia. Secretaria Especial da Fazenda, Secretaria do Orçamento Federal (2020b). Manual Técnico de Orçamento MTO. Versão 2020.

Calhau, R.F. and Falbo, R.A. (2010) An Ontology-based Approach for Semantic Integration. In: 14th IEEE International Enterprise Distributed Object Computing Conference, Vitória. Los Alamitos: IEEE Computer Society, pp.111-120.

Campos, P.M.C. (2019) Designing a Network of Reference Ontologies for the Integration of Water Quality Data. Master Thesis, Federal University of Espírito Santo.

Cruz, I.F., and Xiao, H. (2005) The Role of Ontologies in Data Integration. *Journal of Engineering Intelligent Systems*, vol. 13, p. 245-252.

De Giacomo G., Lembo D., Lenzerini M., Poggi A., Rosati R. (2018) Using Ontologies for Semantic Data Integration. In: Flesca S., Greco S., Masciari E., Saccà D. (eds) *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*. Studies in Big Data, vol 31. Springer, Cham.

De Oliveira, E.F. and Silveira, M.S. (2018) Open government data in Brazil a systematic review of its uses and issues. Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age. ACM.

Falbo, R.A. (2014) SABiO: Systematic Approach for Building Ontologies. In Proceedings of the 1st Joint Workshop ONTO.COM / ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering, FOIS.

Fonseca, L.B., Azevedo, C.L.B. e Almeida, J.P.A. (2014) Mapeando Dados Governamentais com uma Ontologia de Organizações. LOD Brasil Linked Open Data.

Gruber, T.R. (1991) The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases, in Principles of knowledge representation and reasoning: Proceedings of the Second International Conference.

Guarino, N. (1998) Formal Ontology and Information Systems. In Proceedings of the International Conference on Formal Ontology and Information Systems, pp. 3-15.

Guizzardi G., Fonseca C.M., Benevides A.B., Almeida J.P.A., Porello D., Sales T.P. (2018) Endurant Types in Ontology-Driven Conceptual Modeling: Towards OntoUML 2.0. In: Trujillo J. et al. (eds) Conceptual Modeling. ER 2018. Lecture Notes in Computer Science, vol 11157. Springer, Cham. https://doi.org/10.1007/978-3-030-00847-5_12.

Guizzardi, G. (2005) Ontological Foundations for Structural Conceptual Model, CTIT - Centre for Telematics and Information Technology, University of Twente, Doctoral Thesis.

Monteiro, A.C. and Gálvez L.E.Z. (2012). Prospects and limitations in the context of knowledge discovery in database for manipulation of domains through ontologies to support the modeling of data warehouse - Case study in social databases. XXXVIII Conferencia Latinoamericana En Informatica (CLEI).

Pergl, R., Sales, T.P. and Rybola, Z. (2013) Towards OntoUML for software engineering: from domain ontology to implementation model. In International Conference on Model and Data Engineering, pp. 249-263. Springer, Berlin, Heidelberg.

Reitz, K. and Schlusser T. (2016) The Hitchhiker's Guide to Python: Best Practices for Development, O'Reilly Media, 1st edition.

Ruy, F.B. (2017) Software Engineering Standards Harmonization: An Ontology-based Approach. PhD Thesis, Federal University of Espírito Santo.

Santos, L.A., Miranda, G.M., Campos, S.L., Falbo, R.D., Barcellos, M.P., Souza, V.E., and Almeida, J.P. (2018). Using an Ontology Network for Data Integration: A Case in the Public Security Domain. Ontobras, São Paulo.

Tosin, T., Rigo, S.J., Barbosa, J.L.V., and Rodrigues, C. (2016) A model for data integration in open and linked databases with the use of ontologies. 35th International Conference of the Chilean Computer Science Society (SCCC) Part F1262.

U.S. Department of Defense – DoD (2011), Data Modeling Guide (DMG) For an Enterprise Logical Data Model, V2.3, 15 March 2011.