ICPSet: Um Conjunto de Dados Estruturados de Itens de Compras Públicas

Gabriel P. Oliveira[1], Mariana O. Silva[1], Lucas G. L. Costa[1], Marco Túlio Dutra[1,2], Gisele L. Pappa[1]

[1]Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

[2]Universidade Federal de Ouro Preto (UFOP) – Ouro Preto, MG – Brasil

{gabrielpoliveira,mariana.santos}@dcc.ufmg.br, lucas-lage@ufmg.br
marco.dutra@aluno.ufop.edu.br, glpappa@dcc.ufmg.br

Abstract. *Transparency and efficiency in public procurement management are essential to ensure the proper use of public resources. However, the complexity and diversity of procured items pose a significant challenge for analyzing and monitoring these purchases. This paper presents the ICPSet, a structured dataset designed to facilitate the analysis of public procurement data. Containing over 30 million standardized and structured items, the ICPSet provides a robust basis for various analyses and tool development.*

Resumo. *A transparência e a eficiência na gestão de compras públicas são essenciais para assegurar a utilização adequada dos recursos públicos. No entanto, a complexidade e a diversidade dos itens licitados representam um desafio significativo para a análise e monitoramento desses dados. Nesse contexto, este trabalho apresenta o ICPSet, um conjunto de dados estruturado que visa facilitar a análise de dados de compras públicas. Contendo mais de 30 milhões de itens padronizados e estruturados, o ICPSet oferece uma base robusta para diversas análises e o desenvolvimento de ferramentas.*

1. Introdução

A transparência e a eficiência na gestão de compras públicas são essenciais para assegurar a utilização adequada dos recursos públicos. No entanto, a complexidade e a diversidade dos itens licitados representam um desafio significativo para a análise e monitoramento dessas compras. Além disso, muitas vezes a alimentação de tais dados ocorre de forma descentralizada pelos municípios, resultando em uma falta de padronização nos registros, especialmente em campos de texto livre. Essa falta de uniformidade dificulta a comparação e a consolidação de informações, prejudicando a eficácia da fiscalização e da tomada de decisões baseadas em dados [Brum et al. 2024].

Apesar de milhares de itens serem licitados e processados diariamente no Brasil, poucos esforços foram feitos para desenvolver uma base de dados padronizada e estruturada. A predominância de registros em formato não estruturado, a presença de erros e inconsistências nos dados, e a falta de padronização no armazenamento das informações contribuem significativamente para essa lacuna [Oliveira et al. 2023]. Além disso, o grande volume de dados gerados diariamente torna ainda mais desafiador o processo de coleta, organização e extração de informações relacionadas às compras públicas [Silva et al. 2023, Silva et al. 2024].

Portanto, a padronização e a estruturação de itens de compras públicas são essenciais para facilitar a análise e a categorização dos mesmos. A padronização dos dados torna as informações consistentes em um formato uniforme, facilitando comparações e análises automatizadas. Já a estruturação envolve a identificação e organização de atributos específicos dos itens, como quantidade, unidade de medida, cor, material, entre outros, convertendo dados inicialmente não estruturados em formatos que podem ser processados de maneira automática e eficiente.

Nesse contexto, este trabalho propõe o ICPSet, um conjunto de dados estruturados de itens de compras públicas. O ICPSet oferece uma base de dados organizada e padronizada, focada na estruturação de informações detalhadas sobre os itens licitados no Estado de Minas Gerais. Contendo mais de 30 milhões de itens, este conjunto de dados não apenas facilita a análise automatizada e a comparação de compras públicas, mas também fortalece a transparência e a eficiência na gestão de recursos públicos.

2. Trabalhos Relacionados

No âmbito governamental, a disponibilidade de conjuntos de dados estruturados é crucial para promover a transparência e a eficiência na gestão de recursos públicos. Nesse sentido, iniciativas como o Portal de Dados Abertos do Governo Federal[1] e os Portais da Transparência são importantes ferramentas para uma governança mais eficiente. Tais ferramentas em conjunto com dados de outras fontes dão origem a outros conjuntos de dados mais específicos, que abrangem desde informações jurídicas [Sousa and Del Fabro 2019, da Mata et al. 2019, Silva Junior et al. 2022] até administrativas [Silva et al. 2022, Davis 2022]. Estes conjuntos não apenas facilitam o monitoramento das atividades governamentais, mas também permitem a análise detalhada de políticas públicas, alocação de recursos e impacto das decisões governamentais sobre a sociedade.

No entanto, nem sempre tais dados são disponibilizados de maneira padronizada e estruturada. Especificamente em relação aos dados de compras públicas, a falta de uniformidade nos registros é um problema recorrente [Oliveira et al. 2023]. A diversidade nos formatos de registro utilizados por diferentes órgãos governamentais e a falta de uma metodologia uniforme para a categorização e descrição dos itens licitados são obstáculos significativos. Além disso, a complexidade e diversidade dos produtos e serviços adquiridos ampliam a dificuldade na normalização e integração dos dados, limitando a eficácia das análises comparativas e da detecção de padrões de gastos públicos [Silva et al. 2023].

Para enfrentar tais desafios, é fundamental desenvolver abordagens que permitam a estruturação e padronização dos dados de compras públicas [Silva et al. 2024, Brum et al. 2024]. Abordagens baseadas em mineração de dados e processamento de linguagem natural (PLN) têm se mostrado promissoras para lidas com diversidade e complexidade dos dados de texto livre [Ghani et al. 2006, Yang et al. 2022]. Essas técnicas permitem a automação da identificação de características como quantidade, qualidade, especificações técnicas e outros detalhes relevantes a partir das descrições textuais dos itens licitados [Silva et al. 2021, Lucena et al. 2022].

Apesar dos avanços na pesquisa sobre extração de atributos de itens, ainda são escassos os esforços voltados especificamente para o desenvolvimento de uma base de

[1] https://dados.gov.br/home

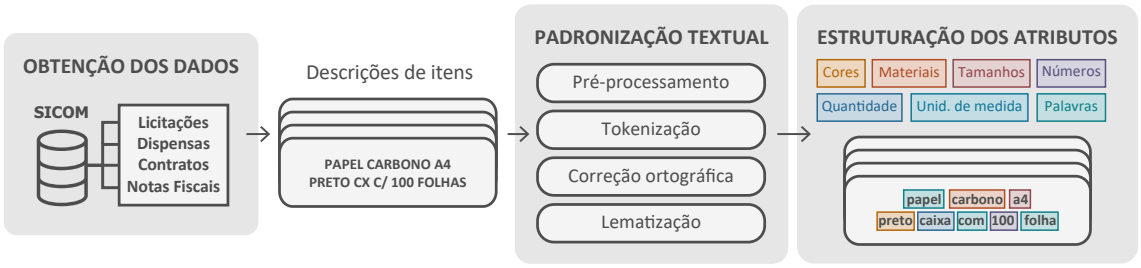

Figura 1. Metodologia de construção do ICPSet.

dados padronizada e estruturada sobre itens de compras públicas. Portanto, propor a criação de um conjunto de dados padronizado e estruturado é uma iniciativa necessária. O ICPSet surge como uma solução notável, visando não apenas organizar e padronizar as informações detalhadas sobre os itens licitados, mas também facilitar a análise automatizada e a comparação de compras públicas.

3. ICPSet

Esta seção apresenta o ICPSet, um conjunto de dados com informações de itens de compras públicas realizadas no Estado de Minas Gerais. A metodologia de construção do conjunto de dados é ilustrada pela Figura 1 e é composta por três passos principais: obtenção (Seção 3.1), padronização (Seção 3.2) e estruturação dos dados (Seção 3.3). Além disso, são detalhadas a organização (Seção 3.4) e a usabilidade do ICPSet (Seção 3.5).

3.1. Obtenção dos Dados

Os dados utilizados para a construção do ICPSet são provenientes do Sistema Informatizado de Contas dos Municípios (SICOM),[2] desenvolvido pelo Tribunal de Contas do Estado de Minas Gerais (TCE-MG) com informações dos portais da transparência dos 853 municípios mineiros. Apesar de estarem disponíveis publicamente, tais dados foram obtidos a partir de consultas em um armazém de dados estruturados presente na infraestrutura do Ministério Público do Estado de Minas Gerais (MPMG) por meio do Programa de Capacidades Analíticas. Especificamente, são considerados itens de compras públicas realizadas pelos municípios entre 2009 e 2023.

Os dados disponibilizados pelo SICOM são exclusivamente textuais e se referem a vários tipos de compras públicas, regidas pela Lei Federal nº 14.133, de 1º de abril de 2021.[3] Para construir o ICPSet, são considerados quatro tipos em específico: (i) **licitações**, processos por meio dos quais a Administração Pública contrata obras, serviços, compras e alienações; (ii) **dispensas**, contendo itens que não necessitam de licitação; (iii) **contratos**, que são ajustes entre a Administração Pública e particulares para a formação de vínculo e a estipulação de obrigações recíprocas; e (iv) **notas fiscais**.

No total, são considerados 30.073.960 itens para serem processados. Cada item é composto por um identificador numérico, uma descrição textual, a descrição da unidade de medida, o ano em que foi comprado, o valor unitário homologado, além de um campo informando o tipo de compra pública (i.e., licitação, dispensa, contrato ou nota fiscal).

[2]SICOM: https://portalsicom1.tce.mg.gov.br/

[3]Lei nº 14.133/21: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm

3.2. Padronização Textual

Após a obtenção dos dados do SICOM, a próxima etapa tem como objetivo padronizar as descrições textuais e corrigir inconsistências para facilitar a comparação e análise automatizada. Em resumo, o processo de padronização textual envolve quatro macro-operações básicas: (i) pré-processamento, (ii) tokenização, (iii) correção ortográfica e (iv) lematização. Cada macro-operação é detalhada a seguir.

Pré-processamento. No pré-processamento, são realizadas seis micro-operações para limpar e preparar as descrições dos itens: (i) *lowercasing*, que transforma todas as palavras em minúsculas; (ii) remoção de *stopwords*, eliminando palavras sem significado essencial para a descrição; (iii) remoção de acentos, substituindo palavras acentuadas por suas formas não acentuadas; (iv) remoção de sinais de pontuação, eliminando todos os sinais de pontuação; (v) remoção de caracteres não-alfanuméricos, excluindo caracteres especiais como, como "#", "", "&", "$"; e (vi) remoção de termos muito longos, retirando palavras com mais de 20 caracteres.

Tokenização. A tokenização transforma uma descrição pré-processada em um vetor de *tokens*, que geralmente correspondem a palavras ou partes delas. Utilizando a biblioteca `NLTK`,[4] especializada para a língua portuguesa, esse processo segmenta o texto em *tokens* que podem ser posteriormente utilizados para análise e processamento textual.

Correção ortográfica. Este processo corrige os erros nos *tokens* das descrições, usando a distância de Levenshtein. Tal distância entre dois *tokens* A e B corresponde ao número mínimo de operações (inserção, remoção e substituição de caracteres) necessárias para transformar um *token* no outro ($A \rightarrow B$ ou $B \rightarrow A$). O corretor ortográfico identifica *tokens* com distância menor ou igual a um determinado valor x e aplica correções apenas aos *tokens* que não possuem correspondência em um dicionário de palavras.[5] Aqui, *tokens* com até duas operações de distância são considerados similares, permitindo a correção de erros ortográficos e garantindo maior consistência e precisão no conjunto de dados.

Lematização. A lematização é um processo linguístico que transforma palavras em sua forma base, chamada de lema. Por exemplo, as palavras "encontrei", "encontraram", "encontrarão" e "encontrariam" seriam reduzidas ao lema "encontrar". Esse processo aumenta a consistência e a padronização das descrições textuais no conjunto de dados. Para realizar a lematização, foi utilizado o DELAF_PB,[6] um dicionário de palavras flexionadas para o português brasileiro. O dicionário possui 880.000 palavras flexionadas e 9.072.338 flexões. Além de fornecer o lema das palavras, o dicionário também apresenta a classe verbal à qual cada palavra pertence. Quando uma forma canônica possui diferentes variações, foi dada prioridade às variações que correspondem à categoria *substantivo*.

Para ilustrar o processo de padronização textual, considere a seguinte descrição de item: *"PAPEL CARBONO A4 PRETO CX C/ 100 FOLHAS"*. Após aplicar as quatro operações de padronização, a saída resultante é uma lista de *tokens* pré-processados: *"papel"*, *"carbono"*, *"a4"*, *"preto"*, *com*, *"caixa"*, *"100"*, *"folha"*. Além de todos os *tokens* estarem em minúsculas e lematizados, duas transformações principais ocorrem nesse processo: os *tokens "c/"* e *"cx"* presentes na descrição original são substituídos por *"com"*

[4]https://www.nltk.org/

[5]https://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS

[6]http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/files/Formato_DELAF_PB.pdf

Tabela 1. Categorias de atributos.

Categoria	Descrição	Exemplos
Cores	Termos que descrevem cores ou tonalidades	*azul, vermelho, claro, escuro*
Materiais	Termos que identificam materiais utilizados na fabricação ou composição dos itens	*metal, inox, madeira, prata*
Números	Todos os termos numéricos	*14, 2015, 500, um, duzentos, iii*
Tamanhos	Termos que indicam tamanho ou escala	*pequeno, grande, p, m, g, unico*
Tipo/Quantidade	Termos que especificam a forma de apresentação dos itens ou sua quantidade	*pacote, cartela, unidade*
Unidade de medida	Termos relacionados a unidades de medida padrão	*litro, grama, ml, kg, cm*
Palavras	Termos que não se encaixam em nenhuma das categorias anteriores	*papel, servico, carne, banana*

e *"caixa"* durante a correção ortográfica.

3.3. Estruturação dos Atributos

A estruturação dos dados visa identificar e extrair atributos específicos dos itens licitados. Esse processo transforma dados não estruturados em formatos que podem ser facilmente processados de forma automática, melhorando a eficiência e precisão das análises. A versão atual do ICPSet considera sete diferentes categorias de atributo, incluindo cores, materiais, números, tamanhos, tipo/quantidade, unidade de medida e palavras. Tais atributos são essenciais para caracterizar detalhadamente cada item. A Tabela 1 descreve cada categoria, apresentando exemplos de termos que se enquadram em cada uma delas.

O processo de estruturação é implementado através do casamento exato de termos. Para isso, conjuntos de termos de cada categoria foram pré-compilados manualmente, exceto para a categoria *Palavras*. Durante a estruturação, cada *token* da descrição já padronizada é comparado com os termos das categorias e direcionado à categoria correspondente. Se o *token* não corresponder a nenhum dos termos pré-compilados das categorias, ele é então incluído na categoria *Palavras*. Isso garante que cada atributo específico dos itens de compras públicas seja devidamente identificado e organizado.

Além disso, um conjunto adicional é criado para armazenar termos ambíguos que se encaixam em mais de uma das categorias pré-definidas. Por exemplo, o termo "laranja" pode referir-se tanto à cor quanto a um produto alimentício, o que o torna ambíguo em termos de classificação. Portanto, os *tokens* que corresponderem a algum termo ambíguo são incluídos não só nas categorias principais (e.g., *Cores*), como também na categoria *Palavras*. Isso garante que a complexidade semântica de certos termos seja adequadamente considerada durante o processo de estruturação, aumentando a precisão na análise dos atributos dos itens de compras públicas.

Para ilustrar o processo de estruturação de atributos, considere a descrição de item padronizada do exemplo da seção anterior: *"papel carbono a4 preto caixa com 100 folha"*. Após aplicar o processo de estruturação nos dados, seis categorias de atributos são identificadas. Os *tokens "papel"*, *"carbono"*, *"com"* e *"folha"* são categorizados como *Palavras*, enquanto *"100"*, *"a4"* e *"preto"* são categorizados como *Números*, *Tamanho* e *Cores*. O *token "caixa"*, por ser um termo ambíguo, é categorizado tanto como *Tipo/Quantidade* quanto como *Palavras*.

Tabela 2. Dicionário de dados do ICPSet.

Campo			Tipo	Exemplo
id_item			*int*	2
id_sicom			*int*	1000001
id_sicom_item			*int*	313215694
id_sicom_fornecedor			*int*	100809
tipo			*string*	Contrato
ano			*int*	2016
preco			*float*	8,0
unidade_medida	original		*string*	UNIDADE
	prep		*string*	unidade
descricao	original		*string*	AGUA MINERAL GALAO DE 20 LITROS
	prep		*string*	agua mineral galao de 20 litro
atributos < *array* >	categoria		*string*	Palavras
	entidades	valor	*string*	agua
		inicio	*int*	0
		final	*int*	4

3.4. Armazenamento e Organização dos Dados

No ICPSet, os 30 milhões de itens são organizados por tipo de registro, ou seja, licitações, dispensas, contratos e notas fiscais. Os itens de cada tipo de registro são armazenados em formato JSON Lines (JSONL), que é um formato leve e eficiente para armazenar grandes volumes de dados estruturados. Entre as vantagens da escolha desse formato estão a facilidade de leitura e escrita sequencial, a compatibilidade com diversas ferramentas de processamento de dados e a eficiência na manipulação de dados em lote.

Cada linha em um arquivo JSONL representa um item individual como um objeto JSON, permitindo uma fácil iteração e o processamento dos dados. A Tabela 2 apresenta os campos presentes em cada arquivo JSONL, o tipo de dado correspondente e um exemplo de entrada para ilustrar a estrutura dos dados. O campo `id_item` representa um identificador único para cada item, enquanto os campos `id_sicom`, `id_sicom_item` e `id_sicom_fornecedor` identificam, respectivamente, o registro no sistema SICOM, o item específico dentro desse registro e o fornecedor correspondente. Já os campos `tipo`, `ano` e `preco` indicam o tipo de registro, o ano da licitação e o preço do item. Além de tais campos básicos, o ICPSet também inclui os campos `unidade_medida` e `descricao`, armazenados tanto na forma original quanto na forma pré-processada (`prep`).

Por fim, o campo `atributos` armazena as categorias e entidades identificadas durante o processo de estruturação dos atributos. Para cada *token*, são registradas informações sobre a categoria à qual ele pertence (e.g., *Palavras*) e a posição inicial e final do *token* na descrição padronizada (`descricao_prep`). Isso permite uma análise detalhada e facilita a recuperação de informações específicas sobre os itens, além de contribuir para a precisão na identificação e categorização dos atributos.

3.5. Usabilidade

Seguindo os princípios da ciência aberta, o ICPSet está publicamente disponível em um repositório no Zenodo.[7] Para cada tipo de registro (i.e., licitações, dispensas, contratos e

[7]https://doi.org/10.5281/zenodo.12691019

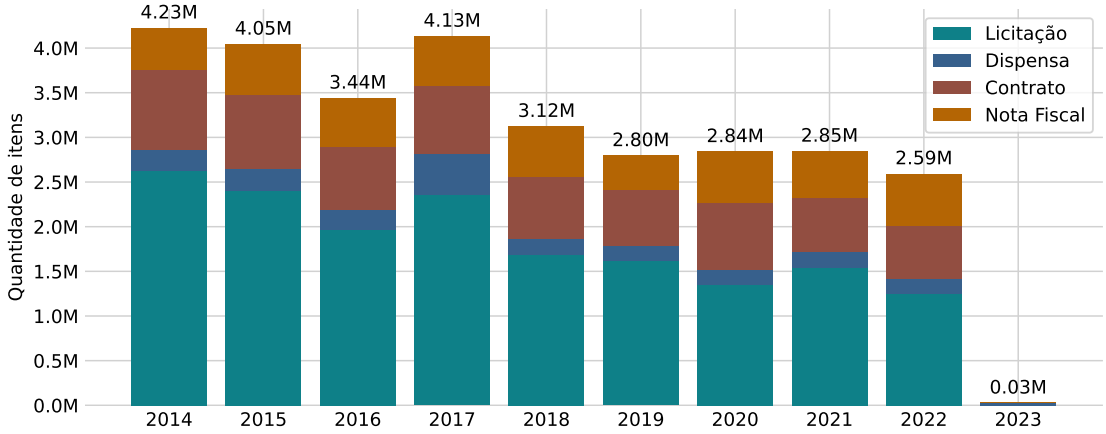

Figura 2. Quantidade de itens de compras públicas por tipo e por ano.

Tabela 3. Distribuição em percentis dos preços dos itens (em reais) por tipo.

	Média	Mín.	25%	50%	75%	90%	99%	Máx.
Licitação	11.468,88	0,00	2,98	13,89	70,00	370,00	22.000,00	$6{,}83 \times 10^{10}$
Dispensa	18.091,19	0,00	5,80	45,85	500,00	5.350,00	140.829,67	$6{,}71 \times 10^{8}$
Contrato	9.203,88	0,00	4,00	19,00	117,00	800,00	76.515,98	$9{,}92 \times 10^{8}$
Nota Fiscal	661,41	0,00	3,50	10,00	55,00	249,52	6.500,00	$5{,}00 \times 10^{6}$

notas fiscais), os dados são organizados em arquivos separados, facilitando o acesso e a análise específica de cada categoria. Além dos arquivos JSONL, os conjuntos de termos pré-compilados para cada categoria de atributo estão disponíveis no mesmo repositório.

4. Análise Exploratória

Esta seção apresenta uma análise exploratória dos dados presentes no ICPSet, a fim de obter informações relevantes sobre suas características principais, além de evidenciar seu potencial de aplicação. De fato, ICPSet pode revelar importantes *insights* sobre a gestão e utilização dos recursos públicos em Minas Gerais. No total, o conjunto de dados possui 30.073.960 itens de compras públicas realizadas pelos 853 municípios de Minas Gerais, sendo tais itens divididos entre licitações (55,83%), dispensas (6,83%), contratos (21,52%) e notas fiscais (15,82%). Tal distribuição informa de antemão as principais formas de aquisição de bens e serviços pelos municípios mineiros.

A Figura 2 apresenta a quantidade de itens presentes no conjunto de dados por ano. Em todo este período, os itens de licitações predominam, refletindo o fato de que as licitações são a principal forma de aquisição de bens e serviços pela Administração Pública. Além disso, não estão disponíveis dados de contratos e notas fiscais para o ano de 2023. Neste caso, é possível que tais dados ainda não estejam presentes no SICOM ou não tenham sido carregados nas tabelas que alimentam a infraestrutura do MPMG.

Em relação ao preço dos itens, a Tabela 3 apresenta a distribuição desses valores por tipo. Além dos valores mínimo, médio e máximo dos itens, são mostrados os percentis 25%, 50% (i.e., mediana), 75%, 90% e 99%. Para todos os tipos, o percentil 90% informa que a grande maioria dos itens foi adquirida pela Administração Pública por valores até

Tabela 4. Top 10 termos mais frequentes nas categorias selecionadas.

Cores		Materiais		Tipo/Quantidade		Palavras	
Termo	**Qtd.**	**Termo**	**Qtd.**	**Termo**	**Qtd.**	**Termo**	**Qtd.**
branco	737.746	plastico	837.798	unidade	1.331.280	de/com/para/sem	21.114.614
preto	431.831	aco	634.555	caixa	1.041.978	servico	1.354.487
transparente	289.939	pvc	631.358	embalagem	868.018	caixa	915.978
verde	277.978	madeira	333.125	pacote	623.164	minimo	724.160
azul	264.270	ferro	233.984	comprimido	543.293	papel	696.113
amarelo	194.750	aluminio	221.554	po	332.081	embalagem	690.251
vermelho	179.048	vidro	210.726	frasco	323.453	produto	661.467
cinza	62.753	inox	201.900	kit	206.018	agua	542.041
rosa	53.947	tecido	179.458	rolo	205.801	dever	535.884
marrom	52.937	metal	144.566	bloco	179.000	cabo	523.156

R$ 5.350. No entanto, existe uma discrepância grande entre tais valores e os valores máximos existentes no conjunto de dados, com itens chegando à casa dos bilhões de reais. Apesar de tais valores poderem ser erros de carga nos dados, é importante ressaltar que existem muitos itens referentes a serviços contratados pelos municípios. Tais itens incluem contratação de empresas e aluguel de estruturas, e naturalmente possuem valores maiores quando comparado com produtos e bens adquiridos.

Por fim, uma das maiores contribuições deste conjunto de dados é a estruturação dos atributos em categorias (conforme Seção 3.3). A Tabela 4 apresenta os dez termos mais frequentes em quatro categorias selecionadas, i.e., cores, materiais, tipo/quantidade e palavras. Tais termos de fato informam as principais composições dos itens, podendo ser úteis em aplicações futuras, incluindo o cálculo de sobrepreço de itens similares. Além disso, conforme mencionado na Seção 3.3, na etapa de estruturação dos atributos, termos ambíguos são incluídos tanto nas categorias principais quanto na categoria *Palavras*. Isso explica a presença dos termos "caixa" e "embalagem" em mais de uma categoria.

5. Aplicações

Esta seção descreve as principais aplicações dos dados do ICPSet, destacando como o conjunto de dados tem sido utilizado em diversas áreas de pesquisa e desenvolvimento.

Inconsistências em Licitações. A metodologia de padronização e estruturação das descrições de itens no ICPSet também podem ser utilizadas para detectar e analisar inconsistências e possíveis irregularidades em processos licitatórios. [Oliveira et al. 2022] propõem uma abordagem para avaliar a compatibilidade entre os itens licitados e os códigos CNAE (Classificação Nacional de Atividades Econômicas) dos licitantes, classificando-os como Válidos, Duvidosos ou Inválidos. Nesse contexto, o primeiro termo identificado entre os *tokens* classificados como *Palavras* é utilizado para verificar a aderência do licitante ao segmento de mercado em questão.

Análise de Sobrepreço. As descrições de itens padronizadas e estruturadas do ICPSet podem ser usadas para realizar o agrupamento e a comparação de preços de itens semelhantes. Isso permite a identificação de sobrepreços e discrepâncias nos valores contratados, possibilitando uma análise detalhada sobre a eficiência e transparência das compras públicas. [Silva et al. 2023] e [Silva et al. 2024] utilizam o ICPSet para analisar indícios

de fraude na aquisição de bens e serviços públicos. Utilizando o intervalo interquartil, ambos trabalhos comparam três estratégias distintas de agrupamento, cada uma enfatizando diferentes facetas do processo de padronização da descrição dos itens, para detectar sobrepreços e desvios significativos nos valores das licitações.

Agrupamento de Itens. A partir das descrições padronizadas dos itens presentes no ICPSet, técnicas de agrupamento de texto podem ser utilizadas para agrupar itens semelhantes com base em atributos como materiais, cores e tamanhos. Tal abordagem permite não só identificar padrões de compra, como também definir preços de referência para produtos ou serviços específicos. [Brum et al. 2024] propõe um *framework* composto pelas etapas de padronização e estruturação do ICPSet, seguidas por uma etapa de representação de texto que captura os componentes mais importantes das descrições dos itens.

Ferramenta Web. O ICPSet também foi utilizado no desenvolvimento da ferramenta Web *Quanto Custa*, projetada para facilitar a consulta e análise de itens licitados [Costa et al. 2024]. Essa ferramenta permite aos usuários buscar por descrições padronizadas, visualizar agrupamentos de itens semelhantes e comparar preços entre diferentes licitações. A integração com os dados estruturados do ICPSet torna o *Quanto Custa* um recurso valioso para gestores públicos e auditores que visam aumentar a transparência e eficiência nas compras. Desenvolvido como parte do Projeto de Capacidades Analíticas, o *Quanto Custa* está disponível exclusivamente para usuários internos do MPMG.

6. Limitações e Oportunidades de Pesquisa

Esta seção aborda os principais desafios e limitações associados ao uso do ICPSet, bem como destaca oportunidades de pesquisa que surgem a partir dessas limitações. Embora o ICPSet forneça uma base sólida e estruturada para análises de compras públicas, certos aspectos podem influenciar a eficácia e a precisão das análises realizadas.

Padronização Textual. A variabilidade e inconsistência nas descrições originais dos itens representam um desafio significativo. Muitas descrições são vagas, incompletas ou utilizam terminologias não padronizadas, dificultando algumas operações de padronização como a tokenização, correção ortográfica e a lematização. Essas inconsistências podem resultar em dados padronizados que ainda contêm variações e erros, afetando a qualidade e precisão das análises subsequentes.
Oportunidades de pesquisa: Desenvolvimento de técnicas avançadas de processamento de linguagem natural (PLN) capazes de lidar com descrições ruidosas e inconsistentes, bem como criação de métodos automatizados para atualização contínua de dicionários de termos e padrões linguísticos.

Estruturação de Atributos. O processo de estruturação de atributos proposto depende de conjuntos de termos pré-compilados para identificar e categorizar atributos específicos dos itens. No entanto, a criação e manutenção desses conjuntos de termos podem ser desafiadoras, especialmente devido à constante evolução das terminologias e à necessidade de adaptação a novos contextos ou tipos de itens. Além disso, a precisão da estruturação depende da abrangência e qualidade dos termos pré-compilados.
Oportunidades de pesquisa: Exploração de abordagens baseadas em aprendizado de máquina para identificação e categorização dinâmica de atributos, bem como a integração de métodos de aprendizado contínuo para adaptação a novas terminologias.

Ambiguidade de Termos. A ambiguidade de termos é uma limitação inerente ao processo de categorização. Alguns termos podem pertencer a mais de uma categoria, como "caixa" podendo referir-se tanto a um tipo de embalagem quanto a uma unidade de quantidade. A abordagem atual tenta mitigar esse problema categorizando termos ambíguos em todas as categorias possíveis, mas isso pode introduzir ruído adicional nas análises.
Oportunidades de pesquisa: Desenvolvimento de métodos de desambiguação de termos que utilizem o contexto em que os termos aparecem, assim como a aplicação de redes neurais para melhorar a precisão na categorização de termos ambíguos.

Cobertura e Generalização. O ICPSet é baseado em dados de compras públicas do Estado de Minas Gerais, o que pode limitar a generalização dos métodos e resultados para outros contextos ou estados brasileiros. Diferenças nas legislações, práticas de compras e terminologias podem afetar a aplicabilidade direta dos métodos desenvolvidos com base neste ICPSet para outras regiões ou sistemas de compras públicas.
Oportunidades de pesquisa: Análise comparativa de práticas de compras públicas entre diferentes estados ou regiões, e adaptação dos métodos e técnicas desenvolvidos para outros contextos, explorando a generalização e transferência de conhecimento.

7. Conclusão

Este trabalho apresentou o ICPSet, um conjunto de dados estruturados que visa facilitar a análise de dados de compras públicas. Contendo mais de 30 milhões de itens padronizados e estruturados, o ICPSet representa uma contribuição significativa para a análise de dados de compras públicas, fornecendo uma base robusta para diversas análises e o desenvolvimento de ferramentas. Entre as possíveis aplicações destacam-se o agrupamento de itens semelhantes, a análise de sobrepreço e a detecção de inconsistências em licitações. A padronização das descrições de itens e a estruturação dos atributos específicos, como materiais, cores e tamanhos, permitem identificar padrões de compra, estabelecer preços de referência e detectar irregularidades com maior precisão.

As limitações e desafios enfrentados, como a variabilidade nas descrições dos itens e a ambiguidade de termos, apontam para áreas que necessitam de atenção contínua. A criação de conjuntos de termos pré-compilados e o desenvolvimento de técnicas avançadas de processamento de linguagem natural são oportunidades de pesquisa que podem contribuir para a superação dessas limitações. Além disso, a expansão do ICPSet para incluir dados de outras regiões e contextos pode aumentar sua relevância e aplicabilidade. De modo geral, as futuras pesquisas e melhorias baseadas neste conjunto de dados têm o potencial de aprimorar ainda mais a transparência e a eficiência das compras públicas, beneficiando gestores, auditores e a sociedade como um todo.

Agradecimentos. A equipe de autoria agradece a Pedro P. V. Brum pela contribuição na primeira fase deste trabalho. Ao Ministério Público do Estado de Minas Gerais (MPMG) pelo apoio através do Projeto Capacidades Analíticas. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Amazon Web Services (AWS) pelo financiamento recebido através do projeto da parceria entre ambos. Por fim, à CAPES e FAPEMIG pelo apoio às pessoas pesquisadoras envolvidas.

Referências

Brum, P. P. V. et al. (2024). Unsupervised grouping of public procurement similar items: Which text representation should I use? In *LREC-COLING*, pages 17176–17185. ELRA and ICCL.

Costa, L. G. L. et al. (2024). Quanto Custa: Banco de Preços de Compras Públicas do Estado de Minas Gerais. In *DS-CoPS*. SBC.

da Mata, W. R. R. et al. (2019). JusBD: Um banco de dados para obtenção de informações do poder judiciário. In *DSW*, pages 398–407. SBC.

Davis, P. (2022). Indicadores e dados municipais: Um banco de dados para avaliar a eficiência das despesas públicas. In *DSW*, pages 79–90. SBC.

Ghani, R., Probst, K., Liu, Y., Krema, M., and Fano, A. E. (2006). Text mining for product attribute extraction. *SIGKDD Explorations*, 8(1):41–48.

Lucena, L. F. et al. (2022). Automatic recognition of units of measurement in product descriptions from tax invoices using neural networks. In *PROPOR*, volume 13208, pages 156–165. Springer.

Oliveira, G. P. et al. (2022). Detecting inconsistencies in public bids: An automated and data-based approach. In *WebMedia*, pages 182–190. ACM.

Oliveira, G. P. et al. (2023). Assessing data quality inconsistencies in brazilian governmental data. *Journal of Information and Data Management (JIDM)*, 14(1).

Silva, F. et al. (2021). Named entity recognition for brazilian portuguese product titles. In *BRACIS*, volume 13074, pages 526–541. Springer.

Silva, M. O. et al. (2022). LiPSet: Um conjunto de dados com documentos rotulados de licitações públicas. In *DSW*, pages 13–24. SBC.

Silva, M. O. et al. (2023). Análise de sobrepreço em itens de licitações públicas. In *WCGE*, pages 118–129. SBC.

Silva, M. O. et al. (2024). Overpricing analysis in brazilian public bidding items. *Journal on Interactive Systems (JIS)*, 15(1):130–142.

Silva Junior, D. et al. (2022). Criação de conjuntos de dados textuais jurídicos em português a partir de processo de extração e heurística. In *DSW*, pages 91–100. SBC.

Sousa, A. W. and Del Fabro, M. D. (2019). Iudicium textum dataset uma base de textos jurıdicos para NLP. In *DSW*, pages 1–11. SBC.

Yang, L. et al. (2022). MAVE: A product dataset for multi-source attribute value extraction. In *WSDM*, pages 1256–1265. ACM.