Programa de Pós-Graduação em
Ciência da Informação - PPGCINF

UnB

Faculdade de Ciência da Informação

MINERAÇÃO DE TEXTO USANDO ARQUITETURA DA INFORMAÇÃO E ONTOLOGIA COMO MÉTODO PARA AUXÍLIO DE AUDITORIA EM NOTA FISCAL DE CONSUMIDOR ELETRÔNICA DO ESTADO DO AMAZONAS

Área de concentração: Gestão da Informação.
Linha de pesquisa: Arquitetura da Informação, Tecnologias e Organização da Informação e do Conhecimento.

Discente: Diana Maria da Camara Gorayeb
Orientador: Prof. Dr. Claudio Gottschalg Duque

INTRODUÇÃO

- A **Era da Informação** traz impactos para a sociedade e altera os mecanismos e estruturas da aquisição e comunicação da informação e incorpora a capacidade de codificação, decodificação, armazenamento e recuperação das unidades de informacionais (dados) que geram conhecimento;

- O **desafio** é reconhecer os contornos de um contexto específico, entender seus processos informacionais e torná-los acessíveis e producentes para atingir objetivos de usuários fazendo uso da revolução tecnológica.

INTRODUÇÃO

- A **Ciência da Informação (CI)**, contribui para alcançar soluções de problemas considerando **unidade informacional e conhecimento** no contexto social, institutional e individual. “Esta contribuição é proveniente de um ramo da pesquisa que, pode-se dizer, é parte do núcleo duro da Ciência da Informação: **os vocabulários controlados**” (ALMEIDA, 2020).

- A **CI** estuda a construção de instrumentos informacionais que hoje estão associados às **ciências técnicas computacionais** utilizando propriedades automáticas que incorporam o tratamento de grandes volumes de dados.

INTRODUÇÃO

- O papel da **Arquitetura da Informação** é afirmar a capacidade do “design” a ser desenvolvido para criar entendimento nas diversas formas de **apresentação da informação: verbalmente, visualmenbte e numericamente**, baseado na **habilidade de organizar e habilidade de achar algo** (WURMAN, 1997).

INTRODUÇÃO

- **Ontologia é um artefato tecnológico** acompanhada de abordagens conceitual, lógica, terminológica e filosófica que elimina as contradições na especificação do domínio, promove vocabulário de consenso, com mecanismo de inferência e que geram **conhecimento**.

- O modelo de **Ontologia** na **CI** colabora:
  - Representação do domínio;
  - Permite a compreensão da realidade e organização dos fatos;
  - Permite a descoberta, recuperação, análise e filtragem de informações como ferramenta de pesquisa ao usuário final visando o aprimoramento do trabalho especializado.

INTRODUÇÃO

- Esta pesquisa tem como **objeto de estudo** os **dados** de produtos descritos em campo livre nos documentos eletrônicos, **Notas Fiscais de Consumidor Eletrônicas (NFC-e)**, e armazenados em banco de dados da Secretaria da Fazenda do Estado do Amazonas;

- Está direcionada à construção de um modelo de **Ontologia** para reconhecer "[...] padrões de intercâmbio, controle de linguagem e modelos de representação do objeto por meio de metadados" (CAMPOS et al., 2006).

INTRODUÇÃO

- Na **Ciência da Informação** abrange os contextos de:
  - Organização e Representação da Informação;
  - Ontologia; e
  - Arquitetura da Informação.

- Utiliza técnicas aplicadas da **Ciência da Computação**:
  - Processamento da Linguagem Natural;
  - Aprendizado de Máquina;
  - Mineração de textos; e
  - Metadados.

PROBLEMA DE PESQUISA

- **Nota Fiscal de Consumidor Eletrônica (NFC-e),** documento emitido e armazenado eletronicamente que registra:
  - Operações relativas aos produtos industrializados;
  - Circulação de mercadorias;
  - Transporte interestadual, intermunicipal e de comunicação.

PROBLEMA DE PESQUISA

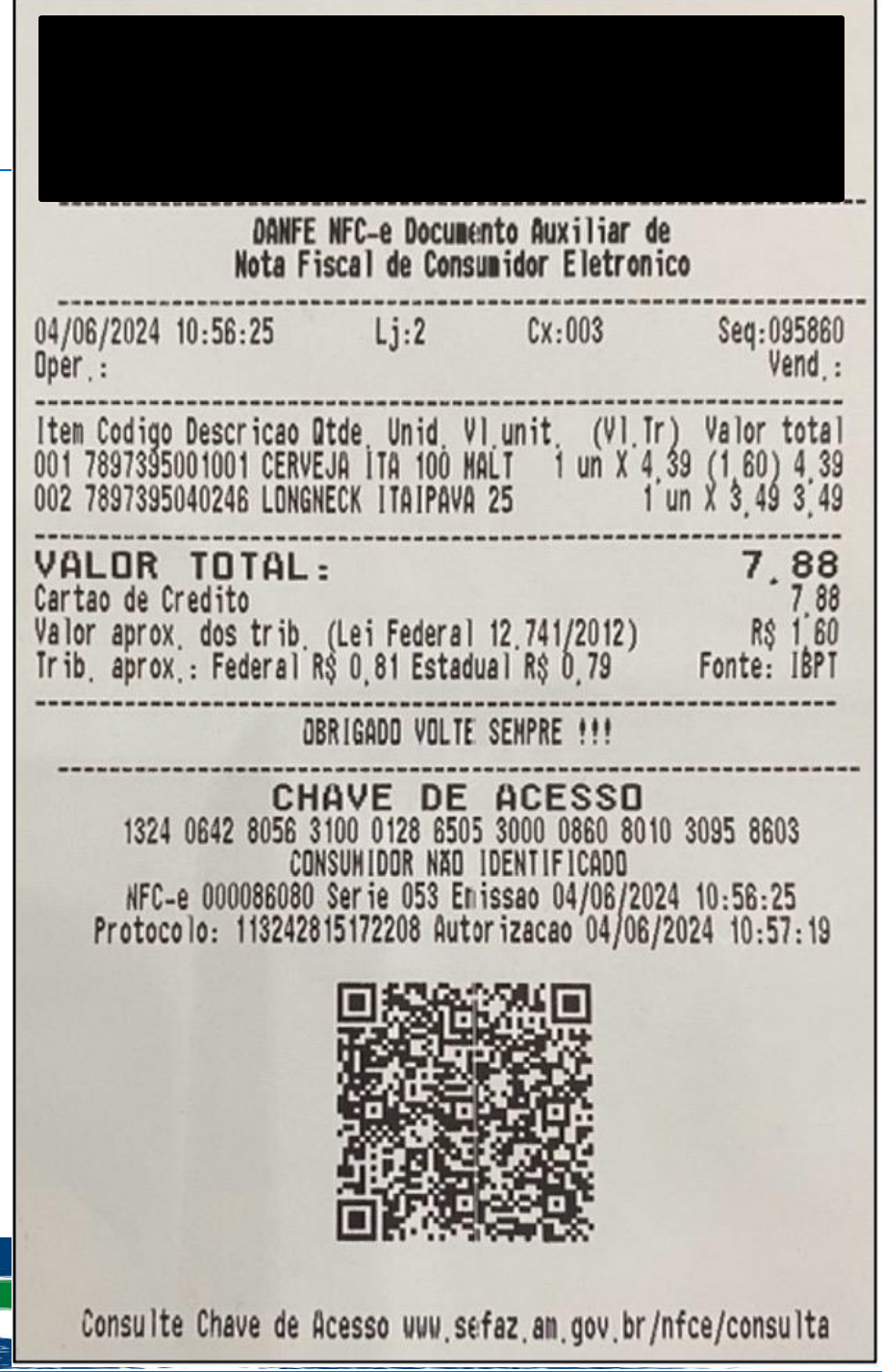
- Arquivos digitais no padrão *Extended Markup Language* (**XML);**
- **Chave de acesso** com 44 dígitos para identificação (metadados lógicos computacionais);
- Campos de informação:
  - Denominação;
  - Número de ordem, série, subsérie e número da via;
  - Data de emissão;
  - Hora de emissão;
  - Nome, endereço, número de inscrição estadual e do CNPJ do emitente;
  - CPF ou CNPJ do consumidor (quando identificado);
  - Destaque dos tributos;
  - **Discriminação das mercadorias.**

PROBLEMA DE PESQUISA

- Metadados da identificação da mercadoria :
  - Nomenclatura Comum do Mercosul (NCM);

- Metadados da identificação da venda:
  - Número do item;
  - Código do item;
  - **Descrição do item;**
  - Quantidade do item;
  - Unidade;
  - Valor unitário do item;
  - Valor total da operação.

PROBLEMA DE PESQUISA



[redacted]

DANFE NFC-e Documento Auxiliar de
Nota Fiscal de Consumidor Eletronico

04/06/2024 10:56:25 Lj:2 Cx:003 Seq:095860
Oper.: Vend.:

Item	Codigo	Descricao	Qtde.	Unid.	Vl unit.	(Vl Tr)	Valor total
001	7897395001001	CERVEJA ITA 100 MALT	1	un	X 4,39	(1,60)	4,39
002	7897395040246	LONGNECK ITAIPAVA 25	1	un	X 3,49		3,49

VALOR TOTAL: 7,88
Cartao de Credito 7,88
Valor aprox. dos trib. (Lei Federal 12.741/2012) R$ 1,60
Trib. aprox.: Federal R$ 0,81 Estadual R$ 0,79 Fonte: IBPT

OBRIGADO VOLTE SEMPRE !!!

CHAVE DE ACESSO
1324 0642 8056 3100 0128 6505 3000 0860 8010 3095 8603
CONSUMIDOR NÃO IDENTIFICADO
NFC-e 000086080 Serie 053 Emissao 04/06/2024 10:56:25
Protocolo: 113242815172208 Autorizacao 04/06/2024 10:57:19

Consulte Chave de Acesso www.sefaz.am.gov.br/nfce/consulta

PROBLEMA DE PESQUISA

DOCUMENTO AUXILIAR
DA NOTA FISCAL DE CONSUMIDOR ELETRONICA

ITEM	COD	DESC	QTDE	UN	VL. UNIT	VL. TOTAL R$
001	7891149010509	CERV BRAHMA 350ML				
	1,000	La x			2,69	2,69
002	78936683	CERV HEINEKEN 330ML				
	1,000	Un x			5,99	5,99
003	7896045506590	CERVEJ HEINEKEN 269M				
	1,000	Un x			3,29	3,29

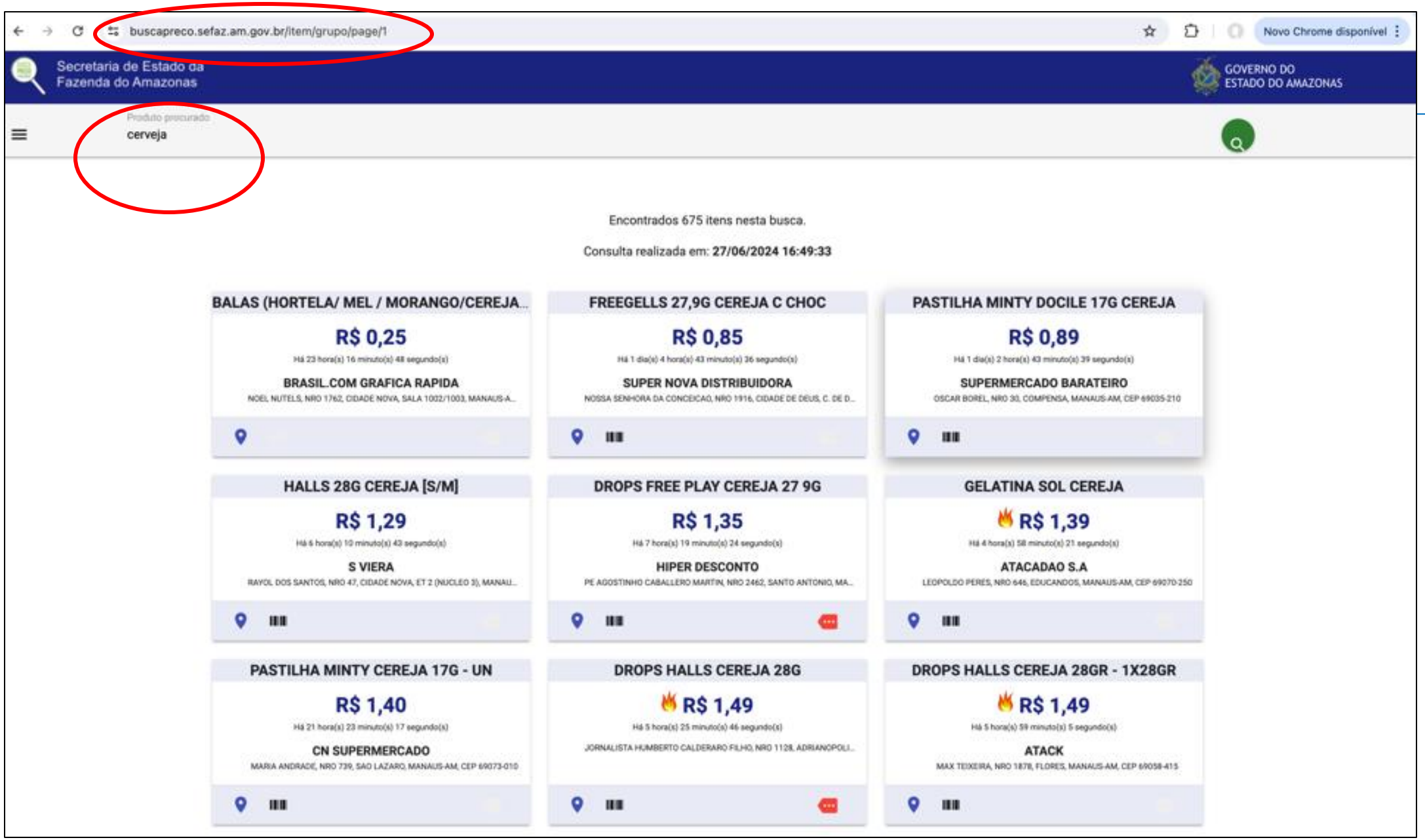
QTD. TOTAL DE ITENS	3
VALOR TOTAL R$	11,97
VALOR A PAGAR R$	**11,97**
FORMA DE PAGAMENTO	VALOR PAGO
Cart Credito	11,97

Consulte pela Chave de Acesso em:
www.sefaz.am.gov.br/nfce/consulta
1324 0606 0572 2304 3571 6502 2000 0778 4312 2112 3920

CONSUMIDOR NAO IDENTIFICADO
NFC-e n. 77843 Serie 22 04/06/2024 10:30:35
Protocolo de Autorizacao:113242815135158
Data de Autorizacao:04/06/2024 10:30:35

PROBLEMA DE PESQUISA

- **Divergências** apresentadas entre a descrição do produto e os controles da SEFAZ/AM para tributar:
  - Nomenclatura Comum do Mercosul (NCM);
  - cEAN (GTIN);
- **Dificuldades para** acompanhar o trânsito do produto com divergências na descrição no início e final da cadeia de venda;
- **Inconsistência** e ausência de metadados na descrição dos produtos;
- **Polissemia** e **falta de padrões** da terminologia para categorização dos produtos;
- O **grande volume** de documentos eletrônicos e **de dados** armazenados.


buscapreco.sefaz.am.gov.br/item/grupo/page/1
Novo Chrome disponível
Secretaria de Estado da Fazenda do Amazonas
GOVERNO DO ESTADO DO AMAZONAS
Produto procurado
cerveja
Encontrados 675 itens nesta busca.
Consulta realizada em: 27/06/2024 16:49:33
BALAS (HORTELA/ MEL / MORANGO/CEREJA...
R$ 0,25
Há 23 hora(s) 16 minuto(s) 48 segundo(s)
BRASIL.COM GRAFICA RAPIDA
NOEL NUTELS, NRO 1762, CIDADE NOVA, SALA 1002/1003, MANAUS-A...
FREEGELLS 27,9G CEREJA C CHOC
R$ 0,85
Há 1 dia(s) 4 hora(s) 43 minuto(s) 36 segundo(s)
SUPER NOVA DISTRIBUIDORA
NOSSA SENHORA DA CONCEICAO, NRO 1916, CIDADE DE DEUS, C. DE D...
PASTILHA MINTY DOCILE 17G CEREJA
R$ 0,89
Há 1 dia(s) 2 hora(s) 43 minuto(s) 39 segundo(s)
SUPERMERCADO BARATEIRO
OSCAR BOREL, NRO 30, COMPENSA, MANAUS-AM, CEP 69035-210
HALLS 28G CEREJA [S/M]
R$ 1,29
Há 6 hora(s) 10 minuto(s) 43 segundo(s)
S VIERA
RAYOL DOS SANTOS, NRO 47, CIDADE NOVA, ET 2 (NUCLEO 3), MANAU...
DROPS FREE PLAY CEREJA 27 9G
R$ 1,35
Há 7 hora(s) 19 minuto(s) 24 segundo(s)
HIPER DESCONTO
PE AGOSTINHO CABALLERO MARTIN, NRO 2462, SANTO ANTONIO, MA...
GELATINA SOL CEREJA
R$ 1,39
Há 4 hora(s) 58 minuto(s) 21 segundo(s)
ATACADAO S.A
LEOPOLDO PERES, NRO 646, EDUCANDOS, MANAUS-AM, CEP 69070-250
PASTILHA MINTY CEREJA 17G - UN
R$ 1,40
Há 21 hora(s) 23 minuto(s) 17 segundo(s)
CN SUPERMERCADO
MARIA ANDRADE, NRO 739, SAO LAZARO, MANAUS-AM, CEP 69073-010
DROPS HALLS CEREJA 28G
R$ 1,49
Há 5 hora(s) 25 minuto(s) 46 segundo(s)
JORNALISTA HUMBERTO CALDERARO FILHO, NRO 1128, ADRIANOPOLI...
DROPS HALLS CEREJA 28GR - 1X28GR
R$ 1,49
Há 5 hora(s) 59 minuto(s) 5 segundo(s)
ATACK
MAX TEIXEIRA, NRO 1878, FLORES, MANAUS-AM, CEP 69058-415

OBJETIVOS

- Objetivo geral:
  - **Auxiliar o processo de auditoria das Notas Fiscais da SEFAZ/AM, por meio da elaboração de um modelo de ontologia de descrição do produto.**

OBJETIVOS

- **Objetivos Específicos (OE):**
  - **Descrever a relevância dos SOCs**, especificamente, da **Ontologia** para os processos de organização e recuperação da informação **(OE01)**;
  - Definir as principais informações extraídas quando aplicada a **Mineração de Texto** nas Notas Fiscais da SEFAZ/AM **(OE02)**;
  - Identificar possíveis requisitos de uma **Arquitetura da Informação** em Notas Fiscais da SEFAZ/AM **(OE03)**;
  - Apresentar um modelo de **Ontologia** para orientar um padrão de descrição do produto para a **auditoria** da Nota Fiscal da SEFAZ/AM **(OE04)**.

JUSTIFICATIVA

- Colaborar com **meios confiáveis de recuperação da informação** na SEFAZ/AM permitindo a percepção do produto e do seu significado em determinado contexto;
- Auxiliar o **processo de auditoria**:
  - Pesquisa de preço (PMPF) para substituição tributária;
  - Controle de Estoque (Fiscalização);
  - Tributação (MVA);
  - Desembaraço fiscal (Fiscalização);
  - Compras por licitação.
- Promover melhoria na **transparência da informação**.

METODOLOGIA

- MÉTODO DE PESQUISA:
  - **Estruturalismo,** (MARCONNI, LAKAKTOS, 2003), investiga um fenômeno concreto de “descrição do produto” e o transporta para o nível abstrato identificando termos e variáveis, estudando seus relacionamentos relevantes, compreendendo e indicando seu uso na construção de um modelo de ontologia que represente o objeto e dê significado, por fim a aplicação em uma realidade estruturada e considerando a experiência com o sujeito social.

METODOLOGIA

- TIPO DA PESQUISA (TRIPODI et al., 1975)
  - ABORDAGEM:
    - **Quantitativa** ao analisar os elementos de descrição dos produtos e propor uma classificação relevante e com significado ao que está escrito baseado em técnicas de Mineração de Texto orientadas pela Arquitetura da Informação;
  - FINS:
    - **Descritiva**, pois pretende construir um modelo de Ontologia voltado para a representação do produto da NFC-e;

METODOLOGIA

- NATUREZA (APPOLINARIO, 2006):
  - **Aplicada** objetiva resolver um problema concreto e imediato do usuário, nesta pesquisa propõe o aprimoramento do processo de auditoria das Notas Fiscais de Consumidor Eletrônicas (NFC-e);
- PROCEDIMENTOS TÉCNICOS (GIL, 2002):
  - **Bibliográfico e Documental** de arquivos .csv tipo texto, no período de **01/02/2023 e 31/05/2023** disponibilizados pela SEFAZ/AM.
- ANÁLISE DOS DADOS:
  - **Corpus** por meio da sistematização das informações fornecidas pela SEFAZ/AM propondo categorias para as transações da NFC-E.

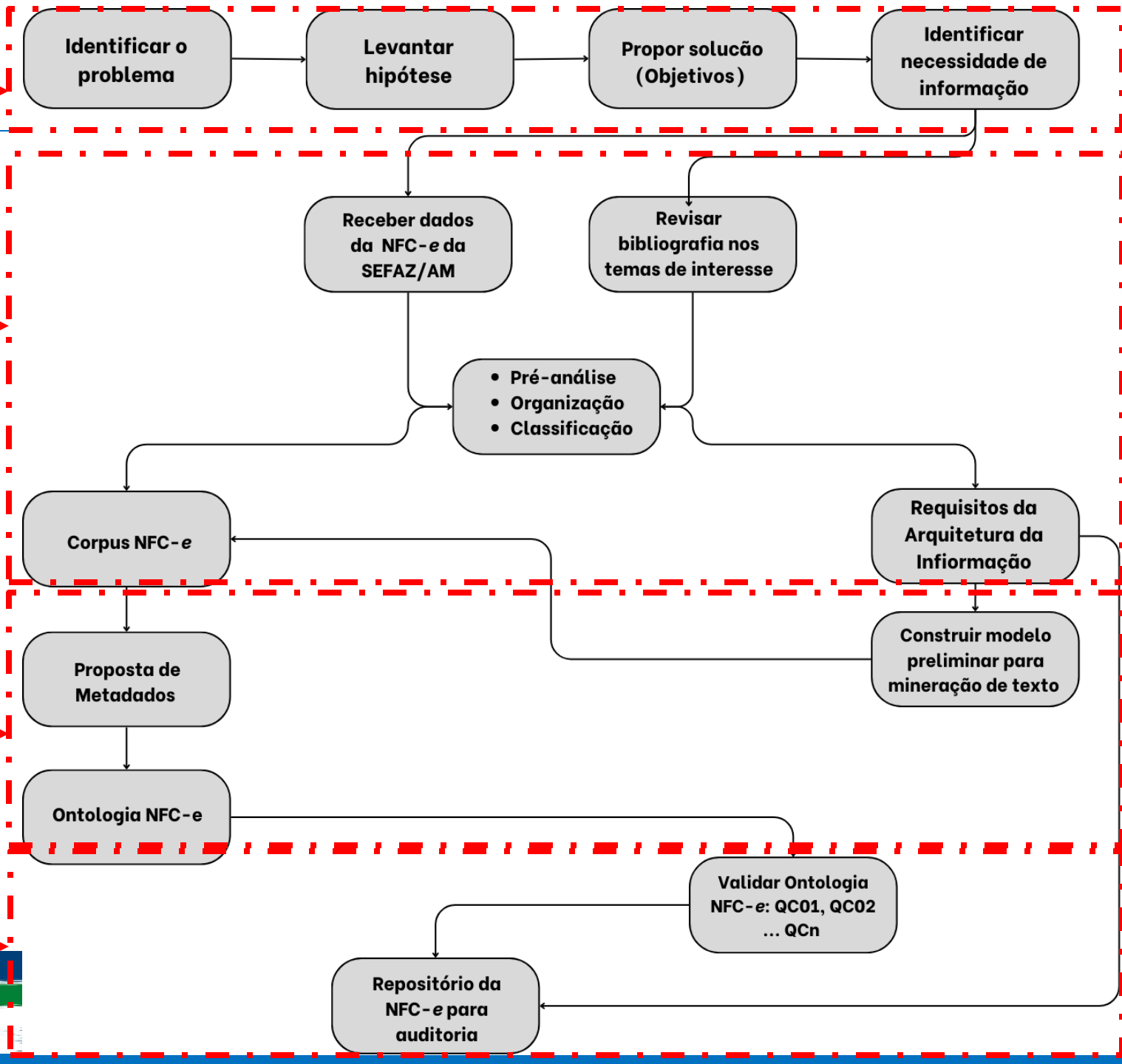
PERCURSO METODOLÓGICO

Descoberta e apresentação do problema à luz dos conhecimentos.

Identificação dos dados, teorias, instrumentos, técnicas e medidas relevantes.

Desenho e obtenção da solução.

Investigação dos resultados e comprovação da solução aplicada à realidade concreta.



REFERENCIAL TEÓRICO DA PESQUISA

- **CIÊNCIA DA INFORMAÇÃO**
  - **Base conceitual da informação e do conhecimento**: Borko (1968); Buckland (1991); Saracevic (1996); Pinheiro (2003); Capurro; Hjorland (2007); Lima-Marques (2011);
  - **Interdisciplinaridade da CI:** Araújo (2003); Zins (2007);
  - **Processo de comunicação:** Shannon; Weaver (1949); Belkin (1978); Brookes (1980); Wersig (1993).

REFERENCIAL TEÓRICO DA PESQUISA

- **ORGANIZAÇÃO E REPRESENTAÇÃO DA INFORMAÇÃO E CONHECIMENTO**
  - **Conceitos da Organização da Informação:** Hjørland (2012); Woledge (1983);
  - **Sistemas de Organização do Conhecimento (SOC):** Bräscher e Café (2008); Hodge, (2000); Duque (2005); Carlan (2010); Gorayeb; Gottschalg-Duque (2022).

REFERENCIAL TEÓRICO DA PESQUISA

- **ARQUITETURA DA INFORMAÇÃO**
  - **Conceito, elementos e objetivos da Arquitetura da Informação:** Wurman (1997); Zachman (1997); Bailey (2002); Hinton (2009); Toms (2002); Downey e Banerjee (2010); Ding e Lin (2010); Siqueira (2012); Rosenfeld; Morville e Arango (2015); Cartaxo e Duque (2016); Cartaxo; Basílio e Duque (2017); Luz (2020); Kuroki Junior e Duque (2023);
  - **Propostas de desenvolvimento de Arquitetura da Informação:** Lima-Marques e Macedo (2006); Orlandi (2019); Costa (2009); Ding e Lin (2010); Downey e Banerjee (2010).

REFERENCIAL TEÓRICO DA PESQUISA

- **ONTOLOGIAS**
  - **Conceito:** Gruber (1993); Choukri (2014); Hamouda; Chourabi e Boughzala (2016); Chang *et al*. (2020); Miranda; Marcelino; Silva (2023);
  - **Ciclo de vida:** Fox et al. (1993); Grüninguer; Fox (1995); Quine (1961 apud Kors, 1997); Vizcaino et al. (2004); Haridy et al. (2023);
  - **Metodologia de desenvolvimento:** Fernández; Gómez Pérez; Juristo (1997); Noy; McGuinness (2001); Sure; Studer (2003); Almeida (2020); Suárez-Figueroa; Gómez-Pérez; Fernandéz-López (2015); Haridy *et al*., (2023); Ghozi *et al*., (2023);
  - **Aplicação:** Standaert; Yaroslaski; Castro (2021); Schulze, *et al*. (2021). Warren (2024);

REFERENCIAL TEÓRICO DA PESQUISA

- **LINGUÍSTICA - BASE PARA CIÊNCIA DA COMPUTAÇÃO (LINGUÍSTICA COMPUTACIONAL):**
  - **Linguagem:** Chomsky (1996); Coseriu (1996); Mendonça (2000); Maculan; Lima (2017);
  - **Termo, Conceito e Significado:** Wittgenstein (1958); Gonzalez; Lima (2003); Novaes (2011); Maimone; Silveira; Tálamo (2011); Maculan; Lima (2017);
  - **Linguística computacional:** Duque (2005); Farias (1998); Othero (2006). Baptista (2015); Oliveira (2020).

REFERENCIAL TEÓRICO DA PESQUISA

- **PROCESSAMENTO DA LINGUAGEM NATURAL (PLN) e APRENDIZADO DE MÁQUINA (AM):**
  - **Técnicas de PLN e Algoritmos de AM Não Supervisionado:** Schuneider (2002); Gonzalez; Lima (2003); Duque (2005); Nadeau; Satoshi (2007); Cambria; White (2014); Sorato, *et al.* (2016); Faceli, *et al.* (2021); Mboli, *et al.* (2021); Chandra, *et al.* (2022); Albuquerque, *et al.* (2023); Haralambous (2024);
- **MINERAÇAO DE TEXTO:**
  - **Descoberta de conhecimento em texto:** Fayyad; Piatetsky-Shapiro; Smyth (1996); Morais; Ambrósio (2007);
- **METADADOS**
  - **Informação e Documentação:** ISO 23081.

PROCESSOS E RESULTADOS DO MODELO

Processos	Resultados
Levantamento e compreensão das normas e documentos da gestão do fisco (Resolução n.º 0028/2023 SEFAZ/AM);	Escopo do projeto para fiscalização: • Filtros para segmentos mais importantes para arrecadação; • Lista de produtos de substituição tributária com Preço Médio Ponderado ao Consumidor Final (PMPF); • Nome dos produtos que poderão enriquecer a ontologia.
Entrada dos filtros que serão aplicados na base NF-e e NFC-e;	Corpus a partir de: • Número Comum do Mercosul (NCM); • Indicação de termo principal; • Períodos pré-determinados de arrecadação etc.
Definição dos metadados descrição dos produtos;	Sentença completa útil à auditoria.
Levantamento e definição de características sintáticas e semânticas dos termos, atributos e associações por meio de algoritmos de IA;	Ontologia: lista de candidatos às classes, subclasses, qualificadores (atributos) e relações existentes.
Definição, validação e enriquecimento dos termos.	Reuso de ontologias referenciadas e utilização de termos adicionais que estão nas Resoluções da SEFAZ/AM.
Incorporar propriedades de dados	Colocar os dados das NFC-e - Data Property Assertions de instâncias das classes para validação dos dados.
Incorporar requisitos da Arquitetura da Informação	Modelo de repositório para auxiliar a auditoria.

ABORDAGEM E CLASSIFICAÇÃO DO MODELO

CRONOGRAMA

ATIVIDADES	2024											
	JAN	FEV	MAR	ABR	MAIO	JUN	JUL	AGO	SET	OUT	NOV	DEZ
Qualificação do Projeto de Tese											X	
Ajustes na redação da Tese											X	X
Tabulação de dados e desenvolvimento da Ontologia											X	X

CRONOGRAMA

ATIVIDADES	2025											
	JAN	FEV	MAR	ABR	MAIO	JUN	JUL	AGO	SET	OUT	NOV	DEZ
Redação da Tese	X	X										
Validação e verificação da ontologia	X	X	X									
Elaboração do modelo de repositório para auditoria			X	X								
Ajustes na redação da Tese					X							
Defesa da Tese						X						