

DIANA MARIA DA CAMARA GORAYEB

**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**

BRASÍLIA, DF

2024

DIANA MARIA DA CAMARA GORAYEB

**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**

Projeto de Tese para qualificação apresentado ao curso de Doutorado do Programa de Pós-Graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília, como exigência parcial para a obtenção do título de Doutora em Ciência da Informação.

Área de concentração: Gestão da Informação.

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

Orientador: Prof. Dr. Cláudio Gottschalg Duque

BRASÍLIA, DF

2024

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

Projeto de Tese para qualificação apresentado ao curso de Doutorado do Programa de Pós-Graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília, como exigência parcial para a obtenção do título de Doutora em Ciência da Informação.

Área de concentração: Gestão da Informação.

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

Aprovado em: XX/XX/2024

Banca Examinadora

Prof. Dr. Cláudio Gottschalg Duque

Presidente

Universidade de Brasília – PPGCINF/UnB

Dr. xxxxxxxxxxxxxxxxxxxxxxxx

Membro

Universidade xxxxxxxxxxxxxxxx - XXXX

BRASÍLIA, DF

2024

Dr. xxxxxxxxxxxxxxxxxxxx

Membro

Universidade de Brasília – PPGCINF/UnB

Dr. xxxxxxxxxxxxxxxxxxxx

Membro

Universidade de Brasília – PPGCINF/UnB

BRASÍLIA, DF

2024

BRASÍLIA, DF

2024

FICHA CATALOGRÁFICA

Gorayeb, Diana Maria da Camara

Xxxxxx

Tese (Doutorado) – Universidade de Brasília

DEDICATÓRIA

XXXXXXXXXX

AGRADECIMENTOS

XXXXXX

RESUMO

Este trabalho, cujo objetivo geral é auxiliar o processo de auditoria das Notas Fiscais da Secretaria de Fazenda do Estado do Amazonas (SEFAZ/AM), por meio da elaboração de um modelo de ontologia, se configura como uma pesquisa na área da Ciência da Informação. Possui como base teórica o enfoque nos conceitos e categorias de: Sistemas de Organização do Conhecimento; Arquitetura da Informação; e Ontologia acompanhados de técnicas aplicadas da área da Ciência da Computação, como: Processamento de Linguagem Natural; Aprendizado de Máquina; Mineração de Dados e Mineração de Texto; e Metadados. O método de procedimento da pesquisa é o Estruturalismo; quanto à natureza, se assenta como uma pesquisa aplicada; considerando a finalidade da pesquisa, se enquadra na pesquisa descritiva; no tocante à abordagem do problema, a pesquisa se apresenta como pesquisa quantitativa; em relação aos procedimentos técnicos, é uma pesquisa bibliográfica e documental, cujos dados provêm de arquivos disponibilizados pela SEFAZ/AM por meio de uma amostra de dados em arquivo .csv, tipo texto, do período de 01/02/2023 a 31/05/2023, contendo transações de Notas Fiscais de Consumidor Eletrônicas, selecionadas a partir da Nomenclatura Comum do Mercosul para o produto cerveja. Para a elaboração da ontologia, o conhecimento será um padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações e onde a lógica permite a produção de regras lógicas a partir das inferências para criação de modelos, representação da informação e extração de conhecimento. A modelagem da informação, o fluxo informacional, o mapeamento da recuperação e a apresentação dos resultados na perspectiva dos processos de negócio e da necessidade do usuário são alguns temas explorados neste trabalho. Para melhor organizar o trabalho, optou-se por uma divisão que apresenta a teoria que resguarda a base científica da pesquisa; na sequência, disserta sobre a metodologia, para dar a conhecer a pretensão e os caminhos que se trilharão para se chegar aos resultados; por fim, apresentam-se os resultados esperados quando da plena implantação da ontologia.

Palavras chaves: Ciência da Informação; Arquitetura da Informação; Sistema de Organização do Conhecimento; Ontologia; Mineração de Texto.

ABSTRACT

This work, whose general objective is to assist the audit process of the Invoices of the Amazonas State Finance Department (SEFAZ/AM), through the elaboration of an ontology model, is configured as research in the area of Science of Information. Its theoretical basis focuses on the concepts and categories of: Knowledge Organization Systems; Information Architecture; and Ontology accompanied by applied techniques from the area of Computer Science, such as: Natural Language Processing; Machine Learning; Data Mining and Text Mining; Logically structured metadata or not. The research procedure method is Structuralism; as for its nature, it is based on applied research; considering the purpose of the research, it fits into descriptive research; regarding the approach to the problem, the research presents itself as quantitative research; in relation to technical procedures, it is a bibliographic and documentary research, whose data come from files made available by SEFAZ/AM through a sample of data in a .csv file, text type, from the period from 02/01/2023 to 05/31/2023, containing electronic consumer invoice transactions, selected from the Mercosur Common Nomenclature for the beer product. For the elaboration of the ontology, knowledge will be a pattern or set of patterns whose formulation can involve and relate data and information and where logic allows the production of logical rules based on inferences for creating models and representing information and extracting knowledge. Information modeling, information flow, recovery mapping and presentation of results from the perspective of business processes and user needs are some themes explored in this work. To better organize the work, a division was chosen that presents the theory that protects the scientific basis of the research; next, talks about the methodology, to make known the intention and the paths that will be taken to reach the results; finally, the expected results are presented when the ontology is fully implemented.

Keywords: Information Science; Information Architecture; Knowledge Organization System; Ontology; Text Mining.

LISTA DE FIGURAS

Figura 1 – Planejamento expandido para elaboração do SOC.....	38
Figura 2 – MIA – Elementos relacionados à Arquitetura da Informação Multimodal	46
Figura 3 – Proposta de Modelo de Arquitetura da Informação.....	48
Figura 4 – Modelo de Arquitetura de Informação apoiado pela Multimodalidade	49
Figura 5 – Modelo de arquitetura da informação para sistemas automatizados	50
Figura 6 – Representação gráfica do Método de Arquitetura da Informação Aplicada (MAIA)	50
Figura 7 – Visão detalhada do processo de projeto da Arquitetura da Informação.....	52
Figura 8 – NFC-e (exemplo 1).....	84
Figura 9 – NFC-e (exemplo 2).....	85
Figura 10 – Tela sítio eletrônico da SEFAZ/AM - consulta para o nome “cerveja”.	87
Figura 11 – Percurso metodológico da pesquisa.....	98
Figura 12 – Procedimentos metodológico de construção do modelo de ontologia.....	100
Figura 13 – Abordagem e classificação da ontologia	103

LISTA DE QUADROS

Quadro 1 – Resumo conceitual de ontologia	27
Quadro 2 – Ciclo de vida da ontologia	54
Quadro 3 – Metodologias para construção de ontologias	55
Quadro 4 – Endereços eletrônicos de ontologias relacionadas com a pesquisa	59
Quadro 5 – Processos e resultados para efetivação da ontologia.....	101

LISTA DE TABELAS

Nenhuma entrada de índice de ilustrações foi encontrada.

LISTA DE SIGLAS E ABREVIACÕES

AI	Arquitetura da Informação
AM	Aprendizado de Máquina
ArS	Arquitetura de Sistemas
CC	Ciência da Computação
CI	Ciência da Informação
CRF	<i>Conditional Random Fields</i>
HMMs	<i>Hidden Markov Models</i>
ICMS	Imposto sobre a Circulação de Mercadorias e Serviços
ISO	<i>International Organization for Standardization</i>
KDD	<i>Knowledge Discovery in Data Bases</i>
KOS	<i>Knowledge Organization System</i>
LC	Linguística Computacional
MIA	Arquitetura da Informação Multimodal
MAIA	Método de Arquitetura da Informação Aplicada
MD	Mineração de Dados
MT	Mineração de Texto
NCM	Nomenclatura Comum do Mercosul
NED	<i>Named Entity Desambiguation</i>
NER	<i>Named Entity Recognition</i>
NFC-e	Nota Fiscal de Consumidor Eletrônica
NF-e	Nota Fiscal Eletrônica

PNL	Processamento da Linguagem Natural
PMPF	Preço Médio Ponderado ao Consumidor Final
POS	<i>Parts-Of-Speech</i>
QC	Questões de Competência
RI	Recuperação da Informação
RWR	<i>Relevant Words Recognition</i>
SEFAZ/AM	Secretaria de Fazenda do Estado do Amazonas
SH	Sistema Harmonizado de Designação e de Codificação de Mercadorias
SOC	Sistema de Organização do Conhecimento
TIC	Tecnologia da Informação e Comunicação
TBT	<i>Transformation-Based Tagging</i>
UDC	Design Centrado no Usuário
UF	Unidade Federativa
XML	<i>Extended Markup Language</i>

SUMÁRIO

1 INTRODUÇÃO	18
1.1 Elementos da pesquisa	20
1.1.1 Problema.....	21
1.1.2 Objetivo.....	22
1.1.3 Justificativa	22
1.2 Organização do Documento	23
2 CIÊNCIA DA INFORMAÇÃO E CIÊNCIA DA COMPUTAÇÃO: AS BASES PARA SE CONSTRUIR UMA ONTOLOGIA.....	25
2.1 Bases teóricas da Ciência da Informação	28
2.1.1 Ciência da Informação: conceitos e princípios.....	28
2.1.2 Ciência da Informação: conceitos e princípios.....	36
2.1.3 Arquitetura da Informação	41
2.1.4 Ontologia.....	53
2.2 Bases teóricas da Ciência da Computação	60
2.2.1 Linguística: base para a Ciência da Computação.....	61
2.2.2 Processamento de Linguagem Natural	66
2.2.3 Aprendizado de Máquina	70
2.2.4 Mineração de Dados e Mineração de Texto.....	72
2.2.5 Metadados	75
3 METODOLOGIA	80
3.1 Contextualização e relação com os temas e conceitos da CI	81
3.2 Contextualização da pesquisa científica: configurações metodológicas.....	95
3.3 Configurações do procedimento metodológico da Ontologia	98
4 RESULTADOS ESPERADOS	105
5 CRONOGRAMA.....	107
REFERÊNCIAS	108
ANEXO A - SOLICITAÇÃO PARA COLABORAÇÃO NA PESQUISA.	120
ANEXO B - DESPACHO FUNDAMENTADO DA SOLICITAÇÃO PARA COLABORAÇÃO NA PESQUISA.....	122

ANEXO C - DESPACHO DE AUTORIZAÇÃO PARA ACESSO AOS DADOS E INFORMAÇÕES DA BASE DE DADOS SEFAZ/AM	123
--	------------

1 INTRODUÇÃO

A Era da Informação surge a partir dos avanços tecnológicos trazidos pela Terceira Revolução Industrial – caracterizada pela automatização exponencial das fábricas, com implantação de sistemas ciberfísicos, de nanotecnologia, neurotecnologia, inteligência artificial, robôs, impressão 3D, biotecnologia e outras tecnologias – e traz grandes impactos para setores da saúde, do trabalho, da educação e, em especial, altera os mecanismos e estruturas da comunicação e da aquisição de informação.

Nesse contexto histórico, as informações são muitas e o desenvolvimento em seus mais diversos setores depende, de “[...] reconhecer os contornos do nosso novo terreno histórico, ou seja, o mundo em que vivemos” (CASTELLS, 2006, p.19). O que fazer com as informações? Como transformar informação em comunicação? Quais veículos, quais técnicas usar para melhor produzir, reproduzir, divulgar a informação? Como tornar o ser humano mais capaz diante dessa Era que desponta e se consolida em dimensões amplas e tempo reduzido?

Essas são questões de pensadores, estudiosos, grandes intelectuais desse tempo histórico. Não se trata apenas de entender os processos da informação e da comunicação, mas de torná-los acessíveis, produtores, principalmente, que sirvam para melhorar a vida das pessoas. Dessa forma, é preciso entender que a sociedade em rede, a sociedade da informação e da comunicação, não é passageira, nem é um monstro que servirá para a destruição da humanidade. Esse entendimento é necessário para que se possa:

[...] identificar os meios através dos quais, sociedades específicas em contextos específicos, podem atingir os seus objetivos e realizar os seus valores, fazendo uso das novas oportunidades geradas pela mais extraordinária revolução tecnológica da humanidade, que é capaz de transformar as nossas capacidades de comunicação, que permite a alteração dos nossos códigos de vida, que nos fornece as ferramentas para realmente controlarmos as nossas próprias condições, com todo o seu potencial destrutivo e todas as implicações da sua capacidade criativa (Castells, 2006, p. 19)

É seguindo o entendimento de que a sociedade da informação pode ser produtiva e benéfica para a humanidade que esta pesquisa se funda. No contexto da

web semântica, é possível conhecer “[...] padrões de intercâmbio, controle de linguagem e modelos de representação por meio de metadados como as ontologias” (CAMPOS *et al.*, 2006, *apud* CARLAN, 2010, p. 53).

Assim, para Noy e McGuinness (2001 *apud* CARLAN, 2010, p. 53) são cinco os motivos pelos quais desenvolver ontologias é importante e considerável:

a) compartilhamento de conhecimento comum em estruturas de informação entre outros povos ou para os agentes de software; b) permite o reuso do conhecimento; c) realiza inferências em um domínio de conhecimento; d) separa o conhecimento de domínio do conhecimento operacional; e) realiza a análise do conhecimento estruturado tendo como resultado respostas mais relevantes. As ontologias promovem e facilitam a interoperabilidade entre sistemas de informação. Por meio de um processo “inteligente” dos agentes (computadores), é possível compartilhar e reutilizar o conhecimento entre os sistemas. As ontologias, fornecem, ainda, um entendimento comum de um domínio entre pessoas de determinada comunidade, entre computadores e pessoas e entre um ou mais computadores (Carlan, 2010, p. 53).

Neste sentido, a execução desta pesquisa na área da Ciência da Informação (CI), cujo ponto de chegada é a organização de uma ontologia, passa, inicialmente por analisar e particionar dados de notas fiscais de consumidor eletrônica, além de estudo teórico de áreas diversas como aplicação de técnicas de mineração de texto para que os objetivos desta pesquisa possam ser cumpridos.

Para a elaboração da ontologia, o conhecimento será um padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações (GOLDSCHMIDT; PASSOS, 2005) e onde a lógica permite a produção de regras lógicas a partir das inferências para criação de modelos e representação da informação e extração de conhecimento.

A modelagem da informação, o fluxo informacional, o mapeamento da recuperação e a apresentação dos resultados na perspectiva dos processos de negócio e da necessidade do usuário são alguns dos entendimentos do uso da CI conforme demonstram as pesquisas no tema que serão exploradas neste trabalho.

Como finalidade da CI são utilizados argumentos, conteúdos, ferramentas, técnicas, métodos e metodologias que se classificam como conceitos, relacionamentos e propriedades para estruturação de espaços informacionais (Mori, 2009), a ontologia nesta pesquisa seguirá o destacado por Carlan (2010):

Considera-se que as ontologias dizem respeito a vocabulários e seus significados, com semântica expressiva, explícita e bem-definida, possivelmente interpretável por máquina. As ontologias são o elemento da web semântica que possibilita o nível de representação semântico. Para isso, é necessário descrever e representar modelos mentais sobre domínios específicos, de maneira utilizável pelo computador, ou seja, é preciso que parte da interpretação semântica possa ser automatizada. As ontologias permitem que isso seja feito, e, por meio delas os softwares usados na web semântica, como agentes inteligentes e web services, são capazes de utilizar o conhecimento codificado para, ao menos parcialmente, entender, isto é, interpretar semanticamente, os documentos e objetos (CARLAN, 2010, p. 55).

Os estudos de ontologia na CI são explorados em técnicas da Ciência da Computação, tendo em vista o problema da recuperação da informação, o vocabulário controlado e técnicas automatizadas do processamento da informação. O modelo produzido é uma representação do domínio que permite a compreensão e organização dos fatos da realidade de acordo com o pensamento do mundo.

Dentre as diversas aplicações do modelo de ontologia destaca-se o suporte à descoberta, recuperação, análise e filtragem das informações na forma de: ferramenta de pesquisa para usuários finais; ser base para classificação e clustering; ser base para algoritmos; ser suporte para integração de bases diversas de informação (interoperabilidade sintática e semântica) promovendo acesso a itens informacionais.

Neste sentido, a pesquisa direcionada à construção do modelo de ontologia abrange, quando se trata de informações para usuários finais, alguns aspectos importantes para o aprimoramento no campo da auditoria como: credibilidade, abrangência semântica na área de conhecimento especializado, disponibilidade e políticas de atualização da informação (FELIPE; SOUZA. 2020).

1.1 Elementos da pesquisa

O objetivo desta seção é descrever de maneira clara e precisa o problema de pesquisa identificado, os objetivos do trabalho e as justificativas para a sua elaboração.

1.1.1 Problema

A Secretaria de Estado da Fazenda do Amazonas (SEFAZ/AM) é o órgão responsável pela efetivação do controle fiscal dos tributos do Estado do Amazonas, o Imposto sobre Circulação de Mercadorias e Serviços (ICMS), como principal tributo, é avaliado por meio da fiscalização e auditoria das notas fiscais emitidas no Estado. A Nota Fiscal Eletrônica (NF-e) e a Nota Fiscal de Consumidor Eletrônica (NFC-e), são modelos nacionais de documento fiscal eletrônico com finalidade de registrar o fato gerador do ICMS.

Dentre as atividades de fiscalização, a identificação do produto a partir das descrições dos itens relacionados nas notas fiscais é um processo importante, pois diversos outros processos dependem do resultado dessa identificação, como por exemplo a definição da alíquota do ICMS a ser aplicada sobre uma operação comercial e o cálculo do preço médio ponderado ao consumidor final (PMPF), utilizado para o cálculo do imposto sobre mercadorias em regime de substituição tributária.

Entretanto, as descrições de produtos e outros dados como tipos de embalagens, unidades e quantidades presentes nas NF-e/NFC-e não são totalmente confiáveis para uma simples avaliação. A variação das possíveis combinações de textos que indicam um produto e suas características é enorme, por exemplo, para o tipo de embalagem que contém 6 latas de um produto, essa embalagem pode ser descrita como “CX06”, “CX-6”, “CXA 6”, “PCT 6”, “6X350ml” e diversas outras maneiras possíveis, inclusive pode não constar nenhuma informação de que o produto é uma embalagem. Portanto, uma interpretação de descrição de produtos baseada em uma análise de texto necessita de um extenso procedimento empregando inteligência humana para explorar as diferentes formas de descrever o objeto de interesse, e mesmo após essa investigação, caso o contribuinte informe na descrição uma nova maneira de representar o produto, o procedimento manual ou pouco inteligente não irá identificar a mudança rapidamente.

Como problema de pesquisa neste trabalho apresenta: **é possível que as notas fiscais se tornem atributos de qualidade se houver um modelo de ontologia para auxiliar o processo de auditoria das Notas Fiscais da SEFAZ/AM?**

Como **hipótese**, explora a elaboração de uma ontologia, poder-se-á auxiliar o processo de auditoria das Notas Fiscais da SEFAZ/AM, pois servirá como um modelo e haverá mais especificidade a respeito do assunto e as diversas variáveis da problemática das notas fiscais serão evidenciadas.

1.1.2 Objetivo

Os objetivos deste trabalho serão apresentados na forma de Objetivo Geral complementado por Objetivos Específicos (OE). Assim o objetivo geral desta pesquisa é:

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

Os Objetivos Específicos necessários para se chegar ao objetivo são:

- Descrever a relevância dos SOC's, 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 para a mineração de texto em Notas Fiscais de Consumidor Eletrônicas (OE03); e
- Apresentar um modelo de Ontologia para orientar um padrão de descrição de produto para as auditorias da Notas Fiscais da SEFAZ/AM (OE04).

1.1.3 Justificativa

As justificativas estão interligadas entre si e destacam a **informação** como elemento fundamental para o desenvolvimento do trabalho no contexto da auditoria e fiscalização:

- Existe à disposição da SEFAZ/AM grande quantidade de dados sobre produtos nas bases da NFC-e, contudo, falta clareza na descrição e na organização da

informação. Os meios de recuperação não são confiáveis a ponto de permitir a fiscalização e a percepção do produto **e do seu significado em determinado contexto**. O uso dos dados de diferentes formas faz com que a extração, para as diversas finalidades da fiscalização, seja basicamente manual, somente sobre o que já é conhecido, limitando as ações de controle da arrecadação e o combate aos atos fraudulentos;

- A ontologia pode servir como um modelo para auxiliar o processo de auditoria melhorando e ampliando o escopo dos produtos auditados do setor de fiscalização;
- Apresentar as relações existentes entre produto e atributos na ontologia permite melhor definição do objeto e consequente expansão da precisão da pesquisa na base de dados da nota fiscal eletrônica;
- Dispor de informação – por meio digital – com capacidade de auxiliar o trabalho humano e torná-lo mais eficiente, promovendo um mecanismo de acesso à informação, tornando mais acessível e inteligível um vocabulário complexo; e
- Promover melhoria na transparência da informação.

1.2 Organização do Documento

Além da introdução e de elementos da pesquisa, no início do documento e das referências e dos anexos ao final do documento, as demais partes são estruturadas até o momento na seguinte forma:

Capítulo 2 - Referenciais teóricos: a primeira parte do capítulo apresenta os referenciais teóricos da Ciência da Informação com ênfase nos conceitos de **Informação** e processo de **Comunicação** com objetivo de recuperar a informação e apresentá-la quando e como for necessário para gerar **Conhecimento** ao encontro de Sistemas de Organização da Informação e Sistemas da Organização do Conhecimento; a seguir e relaciona os principais elementos da **Arquitetura da Informação** enquanto forma de organizar espaços físicos apresentando modelos propostos para aplicação, classificação e organização da informação. Finaliza,

apresentando conceitos da **Ontologia** sob os diversos aspectos da metodologia de desenvolvimento. A segunda parte do capítulo apresenta as referências teóricas das teóricas que suportam aplicações técnicas da Ciência da Computação como a **Linguística Computacional** tratada pelas técnicas de **Processamento da Linguagem Natural** e algoritmos de **Aprendizado de Máquina**. A seguir, apresenta tarefas de processamento da **Mineração de Dados** e de **Texto** incluindo ao final a definição e aplicação de **Metadados**.

Capítulo 3 – Metodologia aborda os elementos teóricos representativos para o entendimento do processo metodológico e científico adotado e inclui detalhamento justificado sobre o contexto e escopo da pesquisa bem como o procedimento metodológico adotado para chegar ao objetivo deste trabalho com a apresentação do modelo de ontologia para auxiliar no processo de auditoria. Para tanto, conceito, processos e ferramentas utilizadas no campo da Auditoria são apresentados ao final do capítulo e sua correlação com o tema investigado. De maneira objetiva, ocupa-se dos seguintes temas: a) conceito de auditoria utilizado para fiscalização; b) relacionamento entre auditoria e o órgão gestor responsável – SEFAZ/AM; c) processo de auditoria que envolvem notas fiscais eletrônicas; d) auditoria e aspectos teóricos da organização da informação disponível nas bases de dados; e e) as ferramentas que são atualmente utilizadas para alcançar os objetivos do tesouro estadual.

Capítulo 4 - Resultados Esperados: Apresenta os resultados que devem ser obtidos durante a pesquisa.

Capítulo 5 – Cronograma aproximado das etapas de desenvolvimento do modelo da ontologia, redação e ajustes na Tese e proposta de Defesa.

2 CIÊNCIA DA INFORMAÇÃO E CIÊNCIA DA COMPUTAÇÃO: AS BASES PARA SE CONSTRUIR UMA ONTOLOGIA

Realizar uma pesquisa sobre organização de conhecimento requer diversos passos teóricos para que ela se firme em conceitos substanciais para empreender a pesquisa, alcançar o objetivo geral e chegar a um resultado sólido. Parte-se, então, de que a CI não pode, sozinha, ser o caminho desta tese. Em princípio, o sentido do termo “conhecimento” deve ser evidenciado para que seja compreendido seu uso nas Ciências Aplicadas, posto que não é tão rigoroso. Observa-se que denominações como “Gestão do Conhecimento”, “Representação do Conhecimento”, “Organização do Conhecimento”, dentre outras, não estão relacionadas ao conhecimento “como um corpo de crenças incontestavelmente verdadeiras” (Almeida, 2020, p. 22).

Assim sendo, a acepção de “conhecimento”, na acepção aplicada, volta-se para “[...] o conjunto de crenças que as pessoas têm boas razões para aceitar, mas as quais não vão aderir de forma dogmática se receberem boas razões para pensar diferente”. A transmissão do conhecimento é uma característica da espécie humana (FELIPE; SOUZA, 2020, p. 83) e sua capacidade inclui fatores como a representação da informação, sua codificação, decodificação, armazenamento e recuperação.

[A CI] apesar de se ter primazia no estudo e construção de instrumentos como as linguagens documentárias e interfaces de sistemas da informação, estes temas são hoje associados às ciências mais técnicas, que com mais propriedades os têm incorporado aos seus fazeres e construtos (FELIPE; SOUZA, 2020, p. 85).

Assim, iniciativas modernas e automáticas tratam grandes volumes de informacionais, com maior respaldo na Ciência da Computação. Considera-se, desse modo, que o conhecimento pode ser “guardado, processado ou manipulado em sistemas computacionais, como sugerem as denominações mencionadas” (ALMEIDA, 2020, p. 22).

Nessa concepção de conhecimento, há de se entender o termo “ontologia” como “um artefato representacional que, incorporado à sistemas, provê à máquinas algum tipo de estrutura de mundo” (ALMEIDA, 2020, p. 22). Considera-se que esse

termo: “ontologia” teve seu início na CI na década de 1990 e trouxe consigo inúmeras dificuldades, dentre elas:

i) os estudos em ontologia seriam uma impostura, pois diziam respeito à classificação e classificação é algo que não se pode reinventar; ii) o uso do termo “ontologia” em outras áreas para denominar um tipo de estrutura de classificação seria apenas uma questão etimológica; iii) o termo “ontologia” seria resultado da aplicação de conceitos antigos da Ciência da Informação à novas tecnologias (ALMEIDA, 2020, p. 23).

No avançar histórico, as dúvidas iniciais cederam diante dos esclarecimentos e a de uma teoria em constante debate. A partir disso, a abordagem ontológica apresenta os três modos de investigação a se considerar no escopo da ciência:

O ontológico, quando se pergunta: qual é a natureza dos objetos que podem ser conhecidos? Qual é a natureza da realidade?
 O epistemológico, quando se pergunta: qual é a natureza da relação entre o sujeito que conhece e o objeto que é conhecido?
 O metodológico, quando se pretende saber: como o sujeito deve proceder para descobrir o conhecimento? (ALMEIDA, 2020, p. 23-24).

Com isso, fica claro que a CI, por vocação, contribui para se chegar a soluções de problemas considerando dados, informação e conhecimento. Estende-se, ainda, para seus registros, no contexto social, institucional ou individual. “Essa contribuição é proveniente também de um ramo da pesquisa que, pode-se dizer, é parte do núcleo duro da Ciência da Informação: os vocabulários controlados” (ALMEIDA, 2020, p. 27) utilizados na recuperação da informação.

Embora se considere a ontologia um artefato computacional, não se pode considerar “tecnologia pela tecnologia”. Isso porque ela “[...] vem acompanhada de abordagem conceitual, lógica, terminológica e filosófica, sob pena de se tornar mais do mesmo”. De tal forma, pode-se evidenciar que “[...] a ontologia é o único artefato dentre as tecnologias recentes ligadas a Inteligência Artificial que carece de um forte componente humano” (ALMEIDA, 2020, p. 27).

Há de se considerar que “[...] o trabalho em ontologias começa naturalmente quando grupos interdisciplinares se envolvem no compartilhamento e na tradução da informação e conhecimento”. Quando se considera o ponto de vista tecnológico, “[...] a ontologia tem potencial para ser parte da solução para a falta de interoperabilidade, um problema que exige ação em diversas frentes” (ALMEIDA, 2020, p. 31), elimina

contradições na especificação do domínio, promove vocabulário de consenso e com notação formal para mecanismo de inferência que geram novos conhecimentos e um grande potencial de reuso (VICTORINO; PINHEIRO; SANTOS, 2015).

Por isso, um arcabouço conceitual bem elaborado que se proponha à execução de atividades diversas antes de implementação do projeto e do desenvolvimento de ontologias será a definição para se ter ontologias bem fundamentadas. Considera-se, portanto, que o “[...] aspecto conceitual é tanto ou mais importante que o aspecto tecnológico porque sem ele, em grande medida, mantém-se muitos dos problemas” (ALMEIDA, 2020, p. 31). No caso, o arcabouço conceitual envolve noções de representação, terminologias e vocabulários da Metafísica e Filosofia da Linguagem, da Lógica e Semântica, da Ciência da Informação, da Ciência da Computação, da classificação e categorização, da gestão de projeto e da Arquitetura da Informação (ALMEIDA, 2020).

O quadro 1 abaixo considera as bases conceituais, o campo e o propósito de uma ontologia para melhor compreensão:

Quadro 1 – Resumo conceitual de ontologia

Distinção	Campo	O que é?	Propósito	Exemplo
Ontologia como uma disciplina	Filosofia	Ontologia como um sistema de categorias	Entender a realidade, as coisas que existem e suas características	Sistemas de Aristóteles, Kant, Husserl
Ontologia como um artefato	Ciência da Computação	ontologia como uma teoria (baseada em lógica)	Entender um domínio e reduzi-lo à modelos	BFO, DOLCE (genéricas)
		ontologia como um artefato de software	Criar um vocabulário para representação em sistemas e para gerar inferências	OWL (linguagem de RC)
	Ciência da Informação	ontologia como uma teoria (informal)	Entender um domínio e classificar termos	Sistema de classificação de Ranganathan
		ontologia como um sistema conceitual informal	Criar vocabulários controlados para recuperação da informação a partir de documentos	um catálogo, um glossário, um tesouros

Fonte: Almeida (2014, p.252).

A partir do exposto, sabe-se que a CI, por seu caráter interdisciplinar, dialoga com diversas áreas do conhecimento. De tal forma, para que se alcance os objetivos desta pesquisa, há de se enveredar por duas áreas de atuação: a CI e suas

abordagens de SOCs; de Arquitetura da informação; de ontologia; e a Ciência da Computação (PLN; AM; Mineração de Dados e Mineração de Texto; Metadados).

2.1 Bases teóricas da Ciência da Informação

A CI, enquanto base conceitual para a construção de uma ontologia e para os estudos da informação e do conhecimento, precisa ser delineada, visando estabelecer os campos conceituais da pesquisa. Como é uma ciência que “[...] está preocupada com o corpo de conhecimentos relacionados à origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação” (Borko, 1968, p. 3), é pertinente que se destaquem os campos da CI que esta pesquisa tomará como base.

Mesmo porque “[...] a investigação da representação da informação em ambos os sistemas, naturais e artificiais, o uso de códigos para a transmissão eficiente da mensagem e o estudo do processamento de informações e de técnicas aplicadas aos computadores e seus sistemas de programação” (Borko, 1968, p. 3) apresenta um vasto domínio. Assim, segue-se o estudo da CI no tocante a esta pesquisa.

2.1.1 Ciência da Informação: conceitos e princípios

A CI constitui campo interdisciplinar com perspectivas, procedimentos e ferramentas nas ciências humanas, naturais e computacionais, abrangendo áreas de estudo como biblioteconomia, filosofia, psicologia, linguística e matemática, com aplicações que vão desde as práticas de gestão, documentação, educação, política, economia e de forma explosiva, as tecnologias para coleta, organização, recuperação e disseminação da informação.

Assim, os estudos de Saracevic (1996) serão uma das bases no que tange à compreensão da área de CI do presente trabalho. Para o autor, uma das contribuições da CI diz respeito ao “modo como a informação é manipulada na sociedade e pela tecnologia”, de modo a assegurar “[...] melhor compreensão para um rol de problemas,

processos e estruturas associados ao conhecimento, à informação e ao comportamento humano frente à informação” Saracevic (1996, p. 20).

Destaca-se que a CI não apenas é um campo das questões científicas, mas também de prática profissional. Ambos dedicados a resolver “[...] os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, instituição ou individual do uso e das necessidades da informação”. Por certo, que para lidar com tais questões consideram-se as vantagens das modernas tecnologias informacionais (SARACEVIC, 1996, p. 7).

Ao longo do tempo, diversos conceitos sobre Informação buscam atender todo o campo que pertence a essa área e assim definir o que seria do escopo da CI. Tarefa difícil é chegar a um conceito amplo o suficiente que consiga abrigar todas as atividades e todos os processos pertencentes ao campo da Informação. Se considerar o que Buckland (1991, p. 1-2) enfatiza, “[...] as definições podem não ser completamente satisfatórias, os limites entre esses usos podem ser confusos e até uma abordagem pode não satisfazer qualquer dos significados determinados como o correto sentido do termo ‘informação’”. Para melhor entender o conceito de Informação, o autor classifica em três principais usos da palavra “informação”, quais sejam:

- (1) Informação-como-processo: Quando alguém é informado, aquilo que conhece é modificado. Nesse sentido “informação” é “o ato de informar...; comunicação do conhecimento ou “novidade” de algum fato ou ocorrência; a ação de falar ou o fato de ter falado sobre alguma coisa” (Oxford English Dictionary, 1989, v.7, p.944).
- (2) Informação-como conhecimento: “Informação” é também usado para denotar aquilo que é percebido na “informação-como-processo”: o “conhecimento comunicado referente a algum fato particular, assunto ou evento; aquilo que é transmitido, inteligência, notícias” (Oxford English Dictionary, 1989, v.7, p.944). A noção de que informação é aquela que reduz a incerteza poderia ser entendida como um caso especial de “informação-como conhecimento”. Às vezes informação aumenta a incerteza.
- (3) Informação-como-coisa: O termo “informação” é também atribuído para objetos, assim como dados para documentos, que são considerados como “informação”, porque são relacionados como sendo informativos, tendo a qualidade de conhecimento comunicado ou comunicação, informação, algo informativo (Oxford English Dictionary, 1989, v.7, 1946) (BUCKLAND, 1991, p. 1-2).

Definida a informação, ainda se mantém o questionamento: que tipos de coisas podem ser informativas? Buckland (1991, p. 5) define dois tipos: (a) Dados - na forma plural, seguindo o sentido da palavra Latina “*datum*”, cujo significado é “coisas que podem ser dadas”. O autor considera que esse é “um termo aceitável para o significado de informação como coisa que tenha sido processada de alguma forma para uso posterior”; (b) Textos e documentos – presentes em arquivos, bibliotecas, e escritórios como, por exemplo, “artigos, cartas, formulários, livros, periódicos, manuscritos, e registros escritos de vários tipos, em papel, em microfilmes, e no formato eletrônico”. O termo “documento” seria usado para denotar objetos textuais, incluindo.

Ampliando o arcabouço teórico, Borko (1968, p. 1) traz grande contribuição à área da CI ao formular a natureza da CI, a qual é pura e aplicada. Pura quando “[...] investiga os fundamentos sem se preocupar com o campo de aplicação”; torna-se aplicado, à medida que “desenvolve produtos e serviços”. Posto isso, pode-se acrescentar que ela “[...] investiga propriedades e comportamentos da informação, as forças que governam seu fluxo e os meios de processá-la para otimizar sua acessibilidade e uso” (BORKO, 1968, p. 6).

Nesse caso, a CI investiga o processo de representação da informação para transmissão da mensagem natural e artificial; os serviços e técnicas de processamento em sistemas de informação para coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação (BORKO, 1968).

Zins (2007) entende a multidisciplinaridade da CI e considera três elementos chaves: fenômeno, domínio e escopo para sistematizar o vasto campo de abrangência dessa ciência.

Como fenômeno, o autor divide em dado, que corresponde ao “conjunto de símbolos que representam percepções empíricas”; informação, que seria o “conjunto de símbolos que representam o conhecimento empírico”; conhecimento, que se apresenta como um “conjunto de símbolos que representam os pensamentos que o indivíduo acredita justificadamente como os verdadeiros”; mensagem, que incide

sobre o “[...]conjunto de símbolos que representa qualquer conteúdo significativo (ZINS, 2007, p. 336).

Quanto ao domínio, Zins (2007, p. 337) elenca três categorias: (1) - Cultural, social e tecnológico, esfera da “atividade humana geral e criativa, qualquer ferramenta física criada por humanos”; (2) *High-tech*, que se configura como “uma subcategoria do tecnológico baseado nos dispositivos computacionais”; (3) Mundo humano, mundo vivo e físico, engloba “todos os organismos biológicos, humanos e não humanos”.

Referente ao escopo, Zins (2007) apresenta os modelos mediador e inclusivo. O modelo mediador compreende a CI “[...] como um campo próximo de todos os outros campos do conhecimento humano”. Já o modelo inclusivo, depreende a CI “[...] como ciência guarda-chuvas que engloba em si todos os campos do conhecimento humano” (ZINS, 2007, p. 339).

Um fator de grande relevância que deve ser considerado nas pesquisas da CI é a interdisciplinaridade. Para que esse domínio fosse contemplado, Araújo (2003) é um autor que traz clareza a esse tema. Para ele, a interdisciplinaridade de CI destaca-se pela sua natureza de ciência pós-moderna, pois constitui sua identidade a partir da computação e da recuperação automática da informação nos moldes das ciências exatas, porém se alinha às propostas e métodos das ciências sociais.

O autor discute a amplitude da ciência pós-moderna demonstrando as implicações do modelo exato bem como da impossibilidade de uma definição dissociada das ciências sociais:

[...] é por pretender se aproximar da “ciência pós-moderna”, superando os limites do modelo até então dominante, buscando superar seus impasses metodológicos simplificadores e abarcar um pensamento pautado pela complexidade, que a ciência da informação evolui para novas etapas de diálogo e inserção nas ciências sociais (ARAÚJO, 2003, p. 26).

Araújo (2003) também entende que a aceitação da natureza interdisciplinar de CI está pautada pela complexidade, que se dá na perspectiva estatística e quantitativa da realidade social e na quebra do paradigma que separava o sujeito do objeto das relações sociais com a evolução da compreensão da realidade social e com o enfoque nas atitudes interpretativas do sujeito. Continua Araújo (2003):

A questão da intersubjetividade conformada a partir da informação se torna central para a compreensão dos diferentes planos de realidade, da distinção entre as diferentes formas de conhecimento e dos mecanismos de sua configuração e legitimação. **Os sujeitos precisam, necessariamente, ser incluídos nos estudos sobre a informação e, sobretudo, precisam ser incluídos em suas interações cotidianas, formas de expressão e linguagem, ritos e processos sociais** (ARAÚJO, 2003, p. 25, grifo nosso).

Os resultados das mudanças de paradigmas inerentes à formação da CI afetaram em grande parte os processos e a necessidade de comunicação, da representação da realidade ou do registo do pensamento humano, fatores sociais imprescindíveis para o surgimento da linguagem natural, para a circulação da informação, da formação e da transformação do conhecimento.

Outro aspecto importante a ser evidenciado em uma pesquisa na CI é o que se entende por informação. Nesse contexto, Belkin (1978) revela que definir a informação facilita a efetiva comunicação da informação desejada entre emissor e receptor e estabelecer uma visão particular do objeto (problema) da CI, demonstrando amadurecimento para o status científico. Como requisitos para o conceito de informação o autor destaca os elementos definidores, comportamentais e metodológicos.

Os elementos definidores dizem respeito ao processo de comunicação social, à natureza da informação desejada e ao efeito da informação no usuário; como elementos comportamentais, o autor considera a reação dos usuários quando expostos ao conjunto de dados. Por conseguinte, tem-se que: (1) usuários diversos reagem de forma diversa ao mesmo conjunto de dados; (2) o mesmo usuário reage de forma diferente ao mesmo conjunto de dados dependendo do momento; (3) a forma de apresentação do conjunto de dados influencia na resposta do usuário. Para além disso, como elementos metodológicos, Belkin (1978) destaca que, quanto à utilidade, o conceito não pode ser situacional, mas generalizável, e que se ofereça a previsão do efeito de determinada informação para determinado usuário.

Não se pode esquecer que o conceito de informação, enquanto fenômeno central da CI, está em constante evolução, dada a sua capacidade de criar conhecimento, além de apresentar múltiplas atribuições para registro, busca e

tratamento do processo de comunicação. Seguindo essa linha de pensamento, Buckland (1991) apresenta três usos da palavra informação:

A informação como processo: informação é o ato de informar alguém e o conhecimento se altera; a informação como conhecimento: o conhecimento comunicado no ato de informar; (fato, assunto ou evento) redução da incerteza; a informação como coisa: atribuído ao objeto, como documento ou dado, são referidos como informações porque são considerados informativos, como tendo a qualidade de transmitir conhecimentos ou comunicar informações, são manuseados, armazenados e recuperados (BUCKLAND, 1991, p. 1).

O autor pretende alcançar tudo o que é potencialmente informativo, abrindo portas para qualquer recurso ou atividade que possa ser fonte de informação e que possa ficar disponível para ser recuperado e utilizado para novas produções, para novas percepções do uso e status de conhecimento da sociedade.

Além disso, pode-se perspectivar a abordagem proposta por Wersig (1993), que considera a informação como matéria e estrutura física do objeto; como conhecimento e dados de valor para determinado fim; como mensagem em um conjunto de símbolos produzidos pelo comunicador para realizar seu intento comunicativo; como significado atribuído ao objeto dependente da interpretação de um agente cognitivo; como efeito para alteração do conhecimento no receptor; e como processo de comunicação para determinado fim.

Os estudos de Brookes (1980) dimensionam o conceito de informação como uma reunião de conhecimentos sobre determinado assunto, parte de uma matéria intelectual percebida pela ação cognitiva fundamentada na lógica e expressa pela equação fundamental da CI. Tal autor reconhece nos estudos filosóficos dos 3 mundos de Popper (1978) a importância do conhecimento objetivo para a CI como representação do pensamento humano criado essencialmente pelo homem, atribuído em artefatos e a serviço de qualquer necessidade de obtenção de conhecimento. A expressão apresenta a relação entre a informação e o conhecimento como uma transformação incremental nas estruturas de conceitos que o indivíduo possui.

$$K[S] + \Delta I = K[S + \Delta S]$$

A relação apresenta o termo $K[S]$ como um conhecimento existente, mas que em determinado momento apresenta uma deficiência pela falta de conhecimento específico desejado “estado anômalo de conhecimento”. A este termo adiciona-se uma porção de informação desejada ΔI , o resultado provoca uma transformação na estrutura de conceitos que o indivíduo possui suprimindo momentaneamente a deficiência anteriormente existente e expressa por $K[S + \Delta S]$.

Ainda em relação a conceituação da informação, Capurro e Hjørland (2007, p. 187) consideram como aquilo que é de importância para se responder a uma questão, sem desconsiderar que:

Deve ser definida em relação a necessidade dos usuários não de modo universal ou individualista, mas de modo coletivo ou particular. Informação responde questões importantes relacionadas ao interesse do usuário, a geração, coleta, organização, interpretação, armazenamento, recuperação, disseminação e transformação da informação e deve ser baseada em visões, teorias, questões e objetivos que a informação deverá satisfazer.

Consideram, então, os autores o aspecto cognitivo, haja vista acontecer interpretação seletiva sobre a informação para a comunicação do conhecimento, reconhecendo o significado da mensagem informativa a partir da relevância e observação do receptor usuário. Com isso, os autores ampliam o conceito ao entender que:

A informação pode ser identificada, descrita e representada em sistemas da informação para diferentes domínios do conhecimento [...] alguns domínios têm alto grau de consenso e critérios de relevância explícitos. Outros domínios têm paradigmas diferentes, conflitantes, cada um contendo sua própria visão, mais ou menos explícitas, da informatividade dos diferentes tipos de informação (CAPURRO; HJØRLAND, 2007, p. 159).

É viável destacar, também, a informação como conteúdo a ser comunicado com finalidade de educação e de produção, dependente do processo de interação entre indivíduos (transmissor e receptor) e da assimilação crítica pelo indivíduo ou grupo social (Vieira, 1983). É possível afirmar que a informação como objeto de interesse reúne, em um determinado escopo, conhecimentos de um domínio explorado a partir de um fenômeno, “[...] um conjunto de símbolos que representa conhecimento empírico” (ZINS, 2007, p. 335) e que provocam uma transformação na experiência do usuário

Outro destaque deve ser evidenciado quanto à aproximação da informação com a comunicação, áreas que se interligam, e que Barreto (1994) analisa a partir dos vários conceitos apresentados em estudos ao longo dos anos e que se concentram em descrever o início e o fim do processo de comunicação entre emissor e receptor da mensagem:

Contudo, são as definições - que relacionam a informação à produção de conhecimento no indivíduo - as que melhor explicam a natureza do fenômeno, em que termos finalistas, associando-o ao desenvolvimento e à liberdade do indivíduo, de seu grupo de convivência e a da sociedade como um todo. Aqui a informação é qualificada como um instrumento modificador da consciência e da sociedade como um todo. Aqui a informação é qualificada como um instrumento modificador da consciência do homem e de seu grupo. Deixa de ser uma medida de organização para ser a organização em si; é o conhecimento, que só se realiza se a informação é percebida e aceita como tal e coloca o indivíduo em um estágio melhor de convivência consigo mesmo e dentro do mundo em que sua história individual se desenrola (BARRETO, 1994, p. 1).

Tem-se, assim, a relação direta entre Informação e Comunicação, não obstante suas especificações particulares, mantêm um caminho paralelo. Destarte, Lima-Marques (2011) afirma que o conhecimento aparece da relação entre o sujeito e o objeto. A função do sujeito é apreender do objeto em termos de suas propriedades (a imagem do objeto) e a função do objeto é ser apreendido pelo sujeito: “O conhecimento é diferente do sujeito e do objeto. O conhecimento aparece como um terceiro elemento que através da correlação se conecta com esses dois elementos formando assim uma trindade” (Lima-Marques (2011, p. 14, tradução nossa).

Em determinado contexto, a representação dos dados suporta as atividades humanas de interesse (CAPURRO; HJØRLAND, 2007), o que implica dizer que os dados adquirem valor para quem o acessa, transformando em informação que permitem a composição do conhecimento:

[...] Se o leitor processa esses dados e os considera relevantes, torna-se informação; essa informação, então, será inferida e se tornará conhecimento que, com o tempo se tornará sabedoria. E através dos processos cognitivos realizados pela mente humana, essa informação é captada e armazenada na memória, a qual permite recuperá-la quando há necessidade. A partir dessa percepção e de posterior categorização da realidade através de conceitos, é possível transmitir conhecimento (pela fala, escrita ou quaisquer meios, dentre eles os informatizados), já que foram explicitados e são passíveis de serem comunicados (NOVAES, 2011, 101).

É significativo esclarecer que, Na CI, os fenômenos informação e conhecimento estão firmemente relacionados ao processo de comunicação como demonstram os vários conceitos até este ponto apresentados. A teoria da comunicação proposta por Shannon e Weaver (1949) discutiu a transmissão da informação descrita em um canal ruidoso que poderia envolver perda de informação. O autor introduziu a “codificação” na comunicação de mensagens como a tecnologia possível ao receptor da mensagem para detectar e corrigir os erros.

Para Shannon e Weaver (1949), a codificação envolve a introdução de uma redundância, a seguir codificar a mensagem antes da sua transmissão (ao invés de transmitir no formato original). A mensagem codificada é enviada pelo canal ruidoso. O receptor decodifica a mensagem, e, em razão da redundância de eventuais erros, estes podem ser detectados e corrigidos obtendo um bom equilíbrio entre redundância e capacidade de recuperar erros.

Dessa forma, a mensagem é, portanto, o conteúdo significativo e tratado da informação (PINHEIRO, 2003) e o conhecimento é o processo social e cognitivo alcançado pela passagem ou transmissão da mensagem informacional. A tecnologia é, reconhecidamente, uma aliada diante das constantes ofertas e buscas de informação, o imperativo tecnológico pressiona o fenômeno da informação e o processo comunicacional sob o aspecto da produção da informação quando favorece a criação e a disseminação de conteúdos informacionais (SARACEVIC, 1996).

2.1.2 Ciência da Informação: conceitos e princípios

Para início do entendimento do que sejam Sistemas de Organização da Informação e Sistemas de Organização do Conhecimento, faz-se necessário considerar o que Documentação. Assim, Paul Otlet, durante seus estudos, adotou a palavra documentação no sentido do processo de fornecimento de documentos ou referências dos mesmos àqueles que precisam da informação que eles contêm. Refere-se a um corpo de conhecimento denominado ciências bibliográficas definido como: produção, fabricação de material, distribuição, registro, estatística, conservação e utilização, por esta razão incluindo compilação, impressão, publicação,

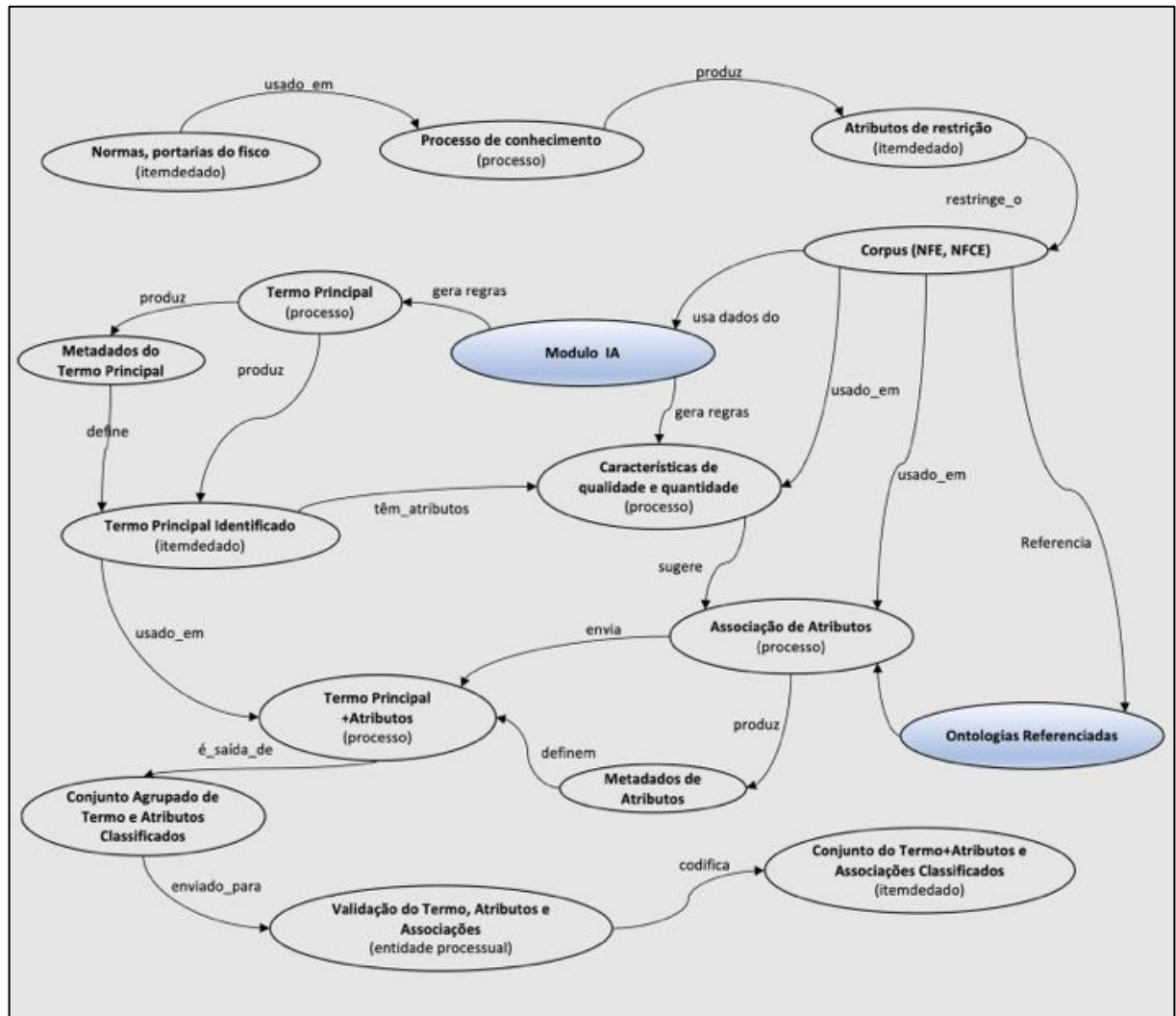
venda, bibliografia e Biblioteconomia considerando objeto construído histórica e conceitualmente (OTLET,1934).

Já Theodore Besterman, em 1945, entendeu que “[...] documentar [servia] para possibilitar a disseminação do conhecimento”. Assim, “[...] Qualquer coisa em que conhecimento é registrado é um documento, e documentação é todo processo que serve para tornar-se um documento disponível para alguém que busca conhecimento” (WOLEGDE, 1983, p. 270).

São exemplos de canais de documentação que guiam o conhecimento até quem o solicita: biblioteconomia e organização de serviços de informação, bibliografia e catalogação, resumo e indexação, classificação e arquivamento, métodos fotográficos e mecânicos de reprodução (WOLEGDE, 1983).

O planejamento para elaboração de um SOC no domínio Nota Fiscal Eletrônica (NFC-e), apresentado em Gorayeb e Gottschalg-Duque (2022), foi expandido com a inclusão das etapas do **Modelo AI** e das **Ontologias Referenciadas** para extrair conjunto dos termos e atributos classificados e associações mais significativas em um processo semiautomatizado sem a interferência de especialistas. É extraído diretamente da base de dados NFC-e, termos que se repetem e tem relevância para o processo de descrição do produto:

Figura 1 – Planejamento expandido para elaboração do SOC



Fonte: Gorayeb; Gottschalg-Duque (2022).

Considerando os tipos de estruturas que se ofertam aos usuários, a representação do conhecimento pode ser classificada, seguindo os preceitos de Brachman (1979, *apud* CARLAN, 2010) em quatro níveis: lógico, epistemológico, ontológico e conceitual. Detalhando cada um deles, tem-se:

- O nível lógico, considerado como nível da formalização, em que não existe preocupação com a semântica em relação aos conceitos e às relações. Neste caso, o foco está direcionado para uma dada “sintaxe”;

- O nível epistemológico, em que a noção genérica de um conceito se dá considerando os fundamentos de estruturação de conhecimento, isto é, especifica-se a estrutura dos conceitos e seus relacionamentos;
- O nível ontológico, cujo objetivo é “limitar o número de possibilidades de interpretação do conceito em um determinado contexto, usando-se de um formalismo para representar o conteúdo do conceito”. Assim sendo, volta-se para o processo de organização e de “classificação de um dado domínio, trabalhando com definições de conceitos que nele estão inseridos”; e
- O nível conceitual, quando os conceitos possuem a priori uma interpretação definida (CARLAN, 2010, p. 25).

Quando estudos são realizados com base nos SOC's, os níveis que interessam são aqueles balizados na estrutura do conhecimento, de modo que possa ser sistematizado e representado a partir de contextos específicos. Neste caso, consideram-se os níveis epistemológicos e ontológicos. Ressalta-se que esses níveis estão “[...] contemplados nos estudos da CI, Teoria do Conceito, Terminologia, Teoria da Classificação e dos SOC's”. A partir de estudos nesses níveis, têm-se “[...] a sistematização de conhecimentos a partir de definições conceituais, representadas por signos linguísticos ou não linguísticos” (CARLAN, 2010, p. 26)

Há de se considerar, ainda, que o termo SOC provém do original inglês “*Knowledge Organization System*” (KOS). Esse termo foi perspectivado na primeira Conferência da *ACM Digital Libraries* em 1998, Pittsburgh, Pennsylvania, quando do *Networked Knowledge Organization Systems Working Group*.

A sigla KOS, amplamente utilizada na literatura, é também denominada com seu correspondente SOC - em português. Os SOC's ou esquemas de representação do conhecimento, como denominado por alguns autores, são de domínio das

[...] áreas de Ciência da Informação, Biblioteconomia e Documentação para designar instrumentos que fazem a tradução dos conteúdos dos documentos originais e completos, para um esquema estruturado sistematicamente, que representa esse conteúdo, com a finalidade principal de organizar a informação e o conhecimento e, conseqüentemente, facilitar a recuperação das informações contidas nos documentos (CARLAN, 2010, p. 28-29).

A infraestrutura é elemento importante ao desenvolvimento de um SOC, pois é ela que dá o suporte. Por isso, há de se fazer “[...] uma análise das necessidades dos usuários dos sistemas; a identificação do tipo de SOC apropriado e o desenvolvimento do hardware e do software adequado à arquitetura de rede, sua integração e manutenção”. Entende-se que o SOC é uma denominação relativamente nova para as linguagens documentárias, que trazem grande potencial para agregar “[...] elementos incorporados nas inovações tecnológicas da era digital” (CARLAN, 2010, p. 28-29).

Para Hodge (2000, *apud* CARLAN, 2010, p. 29), a abrangência dos SOC's se estende a:

[...] todos os tipos de instrumentos usados para organizar a informação e promover o gerenciamento do conhecimento, incluindo os esquemas de classificação que organizam materiais em nível geral e os cabeçalhos de assunto que oferecem o acesso mais detalhado, os catálogos de autoridade, que controlam versões variantes de informação fundamental (como nomes geográficos ou nomes de pessoas) e outros esquemas, como as redes semânticas, tesauros, taxonomias e as ontologias. [...] os SOC são mecanismos para organizar a informação e constituem o “coração” dos Sistemas de Recuperação da Informação (SRI) das bibliotecas, museus e arquivos, no ambiente físico, e, principalmente, no ambiente web.

De tal forma, pode-se considerar como pertencentes à Lista dos SOC's:

- **Tesauros** – “conjunto de termos representando conceitos e as relações hierárquica, equivalente e associativa entre eles”.
- Redes semânticas – apresenta significativo desenvolvimento após o advento do processamento da linguagem natural. “Estrutura conceitos e termos em forma de rede ou teia, os conceitos são nós e os relacionamentos expandem-se a partir dos nós. Uma das redes semânticas usada em mecanismos de busca é a Princeton University's WordNet”;
- Ontologia – nova denominação incorporado aos SOC, “representa relacionamentos complexos entre objetos, incluindo regras de inferência e axiomas, não incluídos em nenhum outro tipo de SOC” (CARLAN, 2010, p. 36).

Para melhor esclarecer, Hjørland (2000, *apud* CARLAN, 2010, p. 29-30) considera os SOC's “[...] como ferramentas que apresentam a interpretação

organizada de estruturas do conhecimento, também chamadas de ferramentas semânticas”. Sintetizando:

Os SOC's são estruturas sistemáticas que visam à construção de modelos abstratos do mundo real, representando os conceitos de um domínio. Essas ferramentas semânticas são utilizadas para o tratamento da informação viabilizando a recuperação da informação, tanto no ambiente informatizado como no tradicional (CARLAN, 2010, p. 29-30).

2.1.3 Arquitetura da Informação

Richard Saul Wurman, em 1976, apresentou o termo arquiteto da informação como “o indivíduo que organiza os padrões inerentes aos dados”; “uma pessoa que cria a estrutura ou mapa de informação que permite aos outros encontrar os seus caminhos pessoais para o conhecimento”; “a ocupação profissional emergente do século 21 que atende às necessidades da época focada na clareza, na compreensão humana e na ciência da organização da informação” (WURMAN, 1997).

Considera-se uma associação à capacidade de lidar com o desenho da informação diante da sobrecarga informacional: “[...] um tsunami de dados é produzido cada vez mais rápido, com mais volume a cada dia e de todas as formas”, sem dúvida uma referência à percepção do tema AI traduzindo para a essência do problema e o contexto em que o tema está inserido (WURMAN, 1997, 58).

O autor destaca que o design deve ser envolvido para criar entendimento nas diversas formas de apresentação da informação (verbalmente, visualmente, numericamente). “Arquitetos devem criar princípios sistêmicos estruturados e ordenados para fazer algo funcionar [...] a criação de algum artefato, ideia ou política que informa porque é claro, objetivo e preciso” (WURMAN, 1997, 61).

A informação, para Wurman, está no sentido de informar, de comunicar algo. Um dicionário ou enciclopédia ou as várias coleções de dados do nosso mundo estão baseados na habilidade de achar algo, a habilidade de achar algo vai junto com a habilidade de organizar: “Precisa de fato pensar no princípio de organização antes de pensar na forma de comunicar” (WURMAN, 1997, p. 129).

Pode-se organizar por exemplo por grandes grupos de palavras, em categoria. Para o autor, a organização da informação é realizada pelos elementos: **Localiation** (localização), **Alphabet** (alfabeto), **Time** (tempo), **Category** (categoria) e **Hierarchy** (hierarquia) – LATCH.

A forma primária para organizar algo é feita decidindo como se deseja achar algo: “a construção da estrutura da informação que permite aos outros entendê-la”. Para Wurman (1997, p. 22), a “[...] estrutura da informação deve relacionar algo que já é compreensível para quem está sendo instruído de forma que seja possível a relação entre algo compreensível extensível a algo não conhecido”.

Zachman (1987), pioneiro da arquitetura aplicada à sistemas, derivou o conceito de arquitetura comparando os desenhos do arquiteto a uma transcrição dos requisitos preceptivos do cliente como: desenhos, plantas, planos de aquisição, planos de construção etc., estabelecendo uma relação com as fases do desenvolvimento de sistemas de informação.

Entende-se que a Arquitetura da Informação é a arte e a ciência de organizar informações desde o nível de infraestrutura até o nível da interface do usuário para que sejam localizáveis, gerenciáveis e úteis (BAILEY, 2002). Amplia-se essa definição quando se entende que a Arquitetura da Informação é um mapa das estruturas que suportam uma informação, um conjunto de elementos que auxiliam no encontro da necessidade do usuário com recursos informacionais. Toms (2002) destaca usuário, sistema e conteúdo como suportes do modelo de interação da informação.

Quanto ao objetivo da Arquitetura da Informação, considera-se projetar contexto com *hiperlinks*, isto é, a Arquitetura da Informação molda o contexto espacial semântico para necessidades específicas através da criação de limites e conexões, a conectividade, a associação (HINTON, 2009). Sem desconsiderar que há a organização da informação como um suporte informacional às organizações em uma perspectiva de gestão de arquitetura dos dados, metadados e gestão do conhecimento. O papel da tecnologia da informação, como base para a Arquitetura da Informação, deve abordar desde a infraestrutura utilizada, porque “[...] AI é a organização de informações para apoiar a localização, gerenciamento e utilidade

desde o nível de infraestrutura até o nível da interface do usuário” (DOWNEY; BANERJEE, 2010, p. 19).

Sendo assim, a Arquitetura da Informação é a atividade de organizar a informação através de mecanismos de rotulagem, estruturação, categorização com a finalidade de apoiar a navegação, localização e uso. Pode-se destacar a Arquitetura da Informação com potencial para: “[...] encontrar problemas de projeto antecipadamente, gerenciar e aproveitar a infraestrutura de software e hardware, identificar lacunas tecnológicas e permitir o uso melhor e mais produtivo dos ativos de uma organização, incluindo informações” (DOWNEY; BANERJEE, 2010, p. 21).

Também é importante ressaltar o papel da AI como meio de gerir a informação e torná-la útil à tomada de decisão. Inclui o projeto de espaços únicos ou múltiplos, agregando estrategicamente vários espaços informacionais, áreas, modalidades e plataformas: “[...] AI trata de organizar e simplificar informações, projetar, integrar e agregar espaços de informação; criar formas para as pessoas encontrarem, compreenderem, trocarem e realizarem gestão da informação para tomar decisões acertadas” (DING; LIN, 2010, p. 11, tradução nossa).

Quanto ao espaço, Lima-Marques (2011) afirma que não há espaço informacional sem Arquitetura da Informação e estes espaços são passíveis de mudança pela intencionalidade do usuário. Um espaço informacional é um conjunto de informações distintas em determinado período, espaço e conteúdo. No espaço informacional, a configuração dos elementos constituintes da informação e os atos de transformação dos espaços são fundamentos da Arquitetura da Informação, intimamente relacionados à natureza do conhecimento.

Considerando a relação sujeito e informação, Siqueira, (2012) destaca que a Arquitetura da Informação se dedica a estudar as relações sujeito e informação, pois é uma disciplina que tem por objetivo “[...] a investigação do Mundo considerando as suas manifestações, formas, contextos e significados do ponto de vista de um Sujeito”, ou seja:

[...] Um esforço sistemático de identificação de padrões e criação de metodologias para a definição de espaços da informação, cujo propósito é a representação e manipulação de informações; bem como a criação de

relacionamentos entre entidades linguísticas para a definição desses espaços da informação (SIQUEIRA, 2012, p. 124).

O autor explica que em ambientes organizacionais, o ciclo de tratamento da informação envolve elementos para “captura, representação, caracterização, significação, armazenamento, recuperação, comunicação, uso e descarte”, e seu desenvolvimento está ligado a AI quando se caracteriza por perceber, pensar, desenhar e habitar Espaços de Informação.

De modo mais detalhado, o *Information Architecture Institute* (2013) apresenta o conceito de AI por meio de várias visões:

1. O desenho estrutural de ambientes de informação partilhada.
2. A arte e a ciência de organizar e rotular sites, intranets, comunidades online e software para apoiar a usabilidade e a localização.
3. Uma comunidade emergente de prática focada em trazer princípios de design e arquitetura para o cenário digital. (Information Architecture Institute, 2013, tradução nossa).

Nesse segmento, Rosenfeld, Morville e Arango (2015) apresentam mais de uma maneira de definir Arquitetura da Informação:

1. O projeto estrutural de ambientes de informação compartilhada.
2. A síntese de sistemas de organização, rotulagem, pesquisa e navegação dentro de ecossistemas digitais, físicos e entre áreas.
3. A arte e a ciência de moldar produtos e experiências de informação para apoiar a usabilidade, a localização e a compreensão.
4. Uma disciplina emergente e uma comunidade de prática focada em trazer princípios de design e arquitetura para o cenário digital. (ROSENFELD; MORVILLE; ARANGO, 2015, tradução nossa).

Continuam os autores que uma AI eficiente considera que as relações entre **contexto**, **conteúdo** e **usuário** se estruturam por meio de quatro sistemas interdependentes: **sistemas de organização**, **sistemas de rotulagem**, **sistemas de navegação** e **sistemas de busca** (grifo nosso). Contexto, conteúdo e usuários formam ambientes informacionais que são alterados a partir da forma como se relacionam e também ao longo do tempo pela necessidade e uso.

Não se pode negligenciar o papel da Tecnologia da Informação como ferramenta para Arquitetura da Informação com enfoque na Recuperação da Informação, compartilhamento e reusabilidade dos recursos representacionais, descrição e semântica, e armazenamento de dados, pois:

[...] permite AI organizar e armazenar a informação estruturada, semiestruturada e não estruturada em repositórios informacionais (bancos de dados, sistemas de arquivos etc.) providos de consistência, compartilhamento, documentação, privacidade e recuperação eficaz de seus conteúdos (CARTAXO; DUQUE, 2016, p. 35).

A Arquitetura da Informação tem o objetivo de identificar as necessidades de informações em ambientes organizacionais, realizar a compreensão dos conteúdos e uma forma de organizá-los de maneira útil:

[...] Arquitetura da Informação (AI) se apresenta com um conceito que pode ser entendido como um *framework* para a representação, organização e armazenagem da informação em repositórios providos de consistência, compartilhamento, documentação, privacidade e recuperação eficaz de seus conteúdos. (CARTAXO; BASÍLIO; DUQUE, 2017, p. 39).

Há duas vertentes quanto à atuação da Arquitetura da Informação, proposta por Costa e Lima-Marques (2017, p. 13) como “[...] uma forma de perceber e manipular a informação nas mais variadas formas e manifestação” e como disciplina que investiga a estrutura da informação, suas características e relações em um domínio. De tal modo, Arquitetura da Informação projeta ambientes informacionais compartilhados a fim de facilitar tarefas, acesso à conteúdos e a apropriação desses espaços pelos usuários, oferecendo um conjunto de instruções para reunir, organizar e apresentar espaços informacionais (LUZ, 2020).

A partir dos conceitos apresentados, é possível entender que a elaboração de um artefato de natureza informacional em que um usuário possa realizar atividades de interesse como trabalhar, se informar, se relacionar, se divertir, conviver de forma ampla requer foco na gestão da informação para que seja útil, usável e aceitável. A Arquitetura da Informação dá sentido ao que se deseja encontrar, uma vez que oferece fundamentos para o levantamento da necessidade e a delimitação do escopo informacional, para o modelo da representação da informação (a descrição e a semântica), para o *design* e implementação do espaço informacional considerando como e onde projetar.

Uma associação importante é feita por Kuroki Junior e Duque (2023), quando associa Arquitetura da Informação à Multimodalidade, denominado Arquitetura da Informação Multimodal (MIA). Atuando na organização do conhecimento, produzindo

visões que expressam melhor e de forma mais eficaz o domínio e, com isso, facilitam reconhecimento de padrões. MIA, segundo os autores, é definida como: “[...] a construção e distinção de mundos arquitetônicos, por meio de pressupostos de modelos relacionais agrupados por contextos espaço-temporais de estados de informação correlacionados ou não”. Neste sentido, MIA está baseada em operadores de possibilidade e necessidade em que uma proposição é possível quando uma das configurações de domínio é verdadeira; e uma proposição é necessária quando todas as configurações de domínio são verdadeiras.

Figura 2 – MIA – Elementos relacionados à Arquitetura da Informação Multimodal

	Arquitetura			Informação	
Propriedade	4.1.1	4.1.2	4.1.3 e 4.1.4	4.2.1	4.2.2
Termos que representam a propriedade na definição	Construção e Distinção de Mundos Arquiteturais	por meio de Suposição de Modelos Relacionais	Agrupados por contextos espaço-tempo	de Estados de Informação	Correlacionados ou não.

Fonte: Kuroki Junior; Duque (2023).

Kuroki Junior e Duque (2023, p. 17) conclui que MIA se aplica em problemas de volume de dados. *Big Data, Deep Learning, Machine Learning* e se preocupa com:

[...] fornecimento de ferramentas para se construir modelos semânticos por meio de regras sintáticas aplicáveis a qualquer implementação linguística. Eis o equilíbrio singelamente proposto entre a Ciência da Informação e a Ciência da Computação em alguns de seus dilemas de fronteiras: a sintaxe é computacional e a semântica é informacional.

Portanto, o alcance da Arquitetura da Informação dentro da CI se constitui como uma disciplina dedicada a aproximar o sujeito do objeto, o pensamento da realidade e a necessidade à fonte informacional. A natureza abrangente dos objetivos da AI permite que seja ela seja pensada e planejada no processo de desenvolvimento do sistema computacionais também.

2.1.3.1 Propostas para o desenvolvimento de uma Arquitetura da Informação

Em geral, as propostas para a aplicação e desenvolvimento da Arquitetura da Informação apresentam uma hierarquia composta de etapas e processos cujos objetivos são organizar e descrever a informação por meio de padrões preocupando-se com estrutura e entrega da informação em canais interativos. Busca atender as expectativas dos usuários utilizando de uma forte infraestrutura tecnológica, algoritmos inteligentes, integração de plataformas de informação, desenvolvimento de *big datas*, e de outras formas de organização e representação do conhecimento, levando em consideração o contexto e os modelos de negócios de gestão. A seguir, apresentam-se algumas propostas:

- **Lima-Marques e Macedo (2006)** apresentam um estudo sobre a importância do conhecimento para gestão e processo de decisão e como ele está relacionado com produtividade e competitividade da organização.

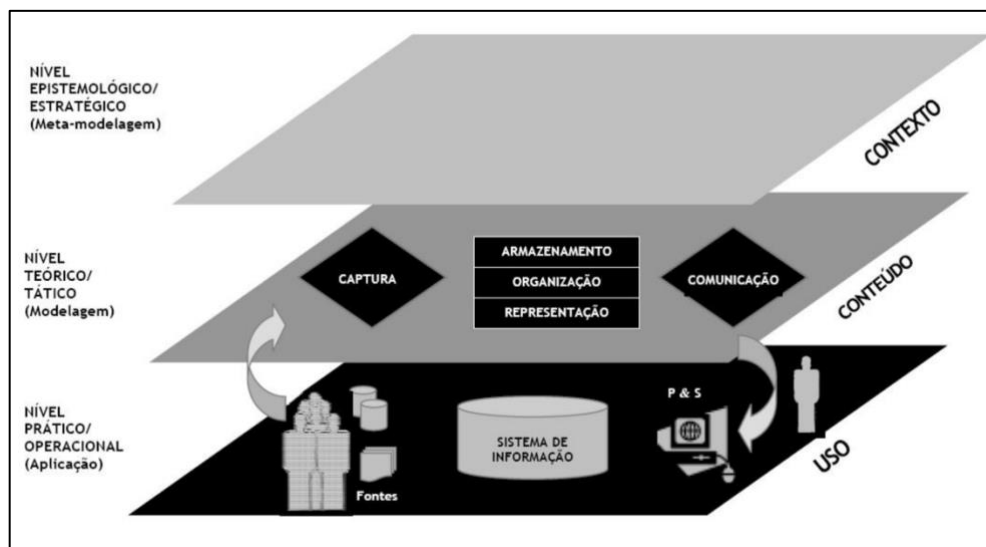
Os autores afirmam que o conhecimento é extraído de dados que caracterizam o negócio continuamente em evolução e acumulados em sistemas informacionais, espaços de atuação da Arquitetura da Informação.

Como sistemas de informação os autores descrevem como: conjunto de informações sistematicamente estruturado, servindo à propósitos bem definidos.

A Arquitetura da Informação apresenta instruções para reunir, organizar e apresentar a informação e é concebida a partir de modelos definidos por propriedades nas etapas do ciclo de vida: gênese, transformação, consolidação e comunicação.

Sendo assim, os autores propõem um modelo em três níveis: 1 – Nível de meta modelagem: para entendimento do contexto; 2 – Nível de modelagem: para identificação, captura, representação, organização e comunicação da informação; 3 – Nível de aplicação: para o uso da informação.

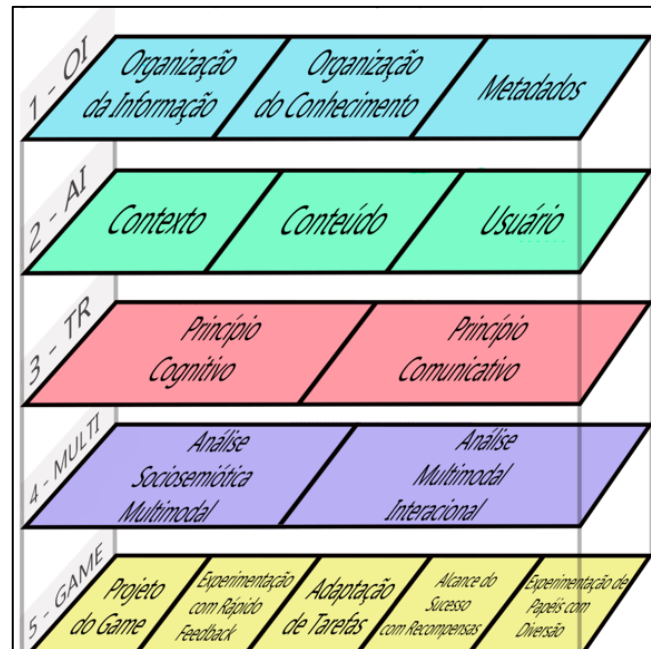
Figura 3 – Proposta de Modelo de Arquitetura da Informação



Fonte: Lima-Marques; Macedo (2006).

- Orlandi (2019) apresenta um modelo concebido para orientar arquitetos da informação no planejamento de espaços informacionais para capacitar profissionais de alto desempenho desenvolvido em 5 etapas:
 - 1 – Organização da informação: incluindo também nesta etapa organização do conhecimento e metadados;
 - 2 – Arquitetura da informação: incluindo percepção do contexto, conteúdo e usuário;
 - 3 – Teoria da relevância: e como o usuário percebe a informação pelos princípios cognitivo e comunicativo;
 - 4 – Multimodalidade: análise socio semiótica multimodal e análise multimodal interacional que, segundo o autor, busca objetos informacionais para melhor comunicar a informação a ser apresentada;
 - 5 – Gamificação: etapa de aplicação do modelo que utiliza jogos de aprendizagem para despertar interesse dos alunos pela informação disponibilizada no ambiente.

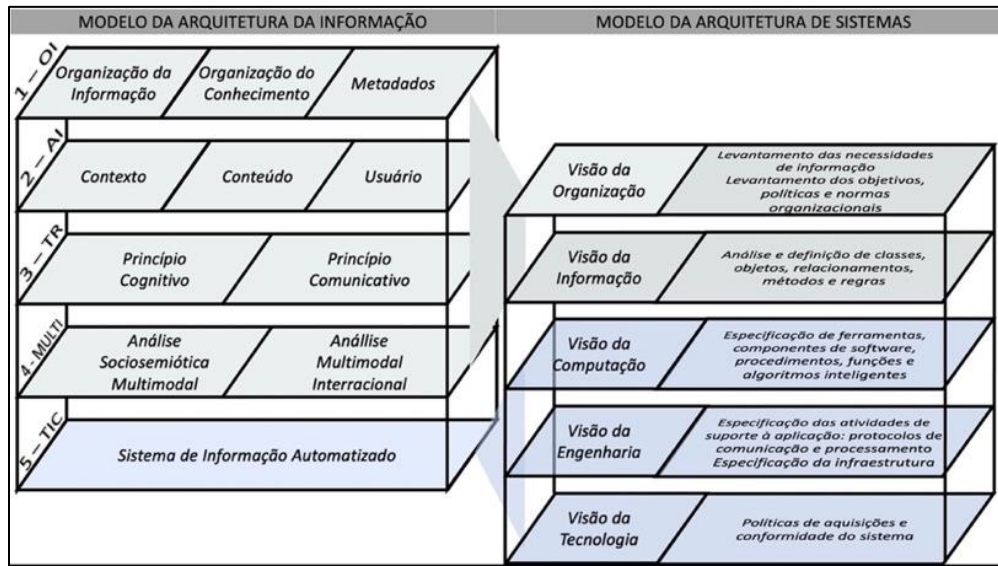
Figura 4 – Modelo de Arquitetura de Informação apoiado pela Multimodalidade



Fonte: Orlandi (2019).

O modelo Orlandi (2019) estendido com a 5ª Etapa sendo substituída para uma generalização: Tecnologia da Informação e Comunicação (TIC) com um único pilar Sistemas de Informação Automatizado proposto em Gorayeb e Gottschalg-Duque (2022). Os autores integram essa 5ª etapa ao modelo de Arquitetura de Sistemas descrito pela ISO 10.746:1998. A finalidade da integração é demonstrar a relação que a Arquitetura da Informação tem com a Ciência da Computação aplicando a necessária Arquitetura da Informação a toda e qualquer construção de sistemas.

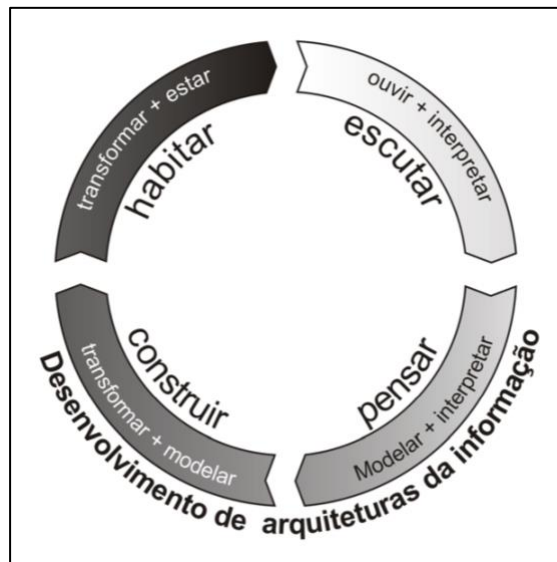
Figura 5 – Modelo de arquitetura da informação para sistemas automatizados



Fonte: Gorayeb, Gottschalg-Duque (2022).

- **Costa (2009)** apresenta um modelo para desenvolvimento de um espaço informacional com elementos escutar, pensar, construir e habitar, cuja ordem do momento de desenvolvimento está disposta na figura 6 abaixo:

Figura 6 – Representação gráfica do Método de Arquitetura da Informação Aplicada (MAIA)



Fonte: Costa (2009).

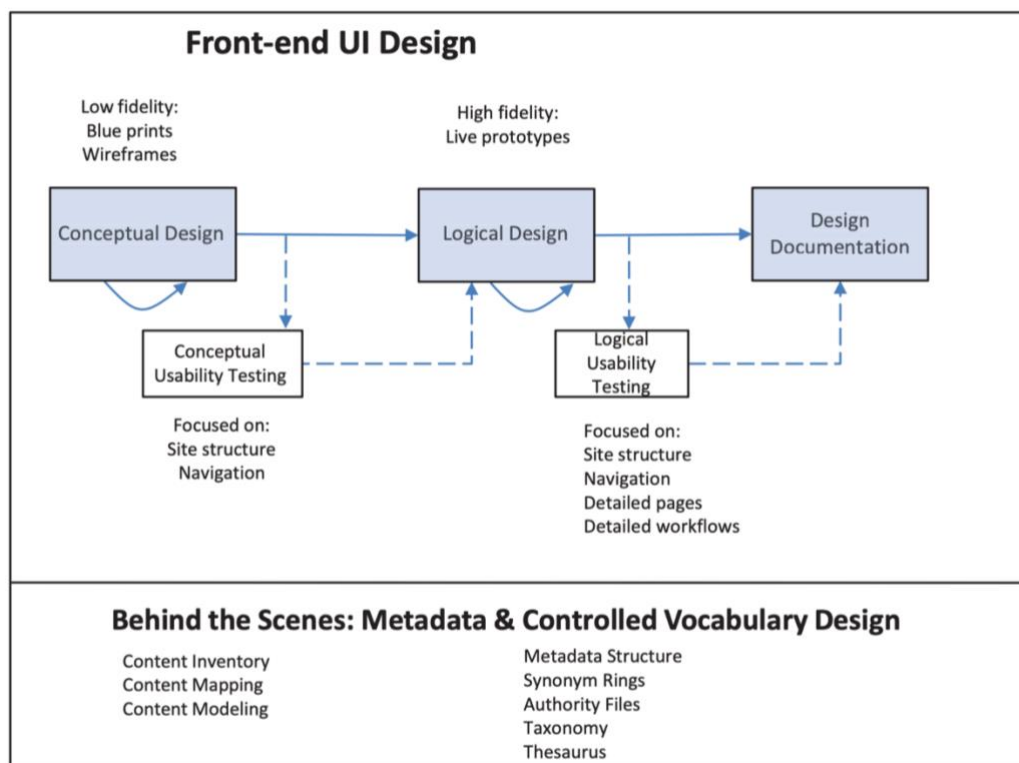
O autor destaca que o Escutar, o Pensar, o Construir e o Habitar são momentos de atuação do sujeito sobre o espaço de informação.

O Escutar e o Pensar são momentos voltados para os aspectos abstratos deste espaço. O Construir e o Habitar são momentos voltados para os aspectos concretos [...] o escutar concentra as percepções do espaço de informação; o pensar representa a modelagem hermenêutica do espaço de informação; o construir reúne as ações de manipulação dos elementos do espaço de informação; o habitar é o momento no qual o sujeito usa o espaço de informação. a configuração dos elementos em um espaço de informação é denominada arquitetura da informação (AI). (COSTA, 2009, p. 93).

A evolução do modelo MAIA foi tratada em Costa; Lima-Marques, (2017) os autores destacam que o modelo é uma ferramenta de configuração de espaços de informação, mas também pode ser aplicado como um modelo de investigação científica e como um modelo gerador de ontologias. Neste caso, Beira *et al.* (2017) destacam as contribuições da Arquitetura da Informação enquanto disciplina científica e a recomendação do MAIA para a construção de ontologias de domínio. Os autores ressaltam a importância de gestão do conhecimento na organização para elaborar diretrizes na condução e definição de ontologias. O estudo de caso baseou-se na construção de ontologia de domínio do produto “GovContaCaixa” e a contribuição da Arquitetura da Informação do ponto de vista filosófico, científico e tecnológico.

- **Ding e Lin (2010)** apresentam um processo de projeto de AI com objetivo de conciliar a necessidade do usuário e os objetivos organizacionais, o Design Centrado no Usuário (UCD). O processo integra as práticas da ISO 13407/1999 Human-Centered Design Process for Iterative Systems ou processo de projeto centrado no usuário para sistemas interativos e possui duas áreas: Front-end UI Design ou interface - projeto de interação da informação e usuário; Behind the Scenes ou estrutura e esquema de metadados - representando os mecanismos de classificação, vocabulários controlados, mecanismos de busca e modelo do conteúdo da informação, Figura 7:

Figura 7 – Visão detalhada do processo de projeto da Arquitetura da Informação



Fonte: Ding e Lin (2010).

Os autores trabalham em conjunto com o processo geral de desenvolvimento de sistemas. Os conceitos do projeto de AI são usados para orientar as atividades de coleta de requisitos em Arquitetura de Sistemas (ArS) e os requisitos, por sua vez, ajudam a refinar o projeto arquitetural em um ciclo iterativo e incremental.

- **Downey e Banerjee (2010)** os autores propõem na forma de uma lista de atividades a revisão de Arquitetura da Informação a fim de ajudar as organizações a alcançar a informação. A lista de atividades verificáveis inclui:
 1. Organização das informações no nível de estruturar e compor um repositório que coleciona informações, Corpus;
 2. Utilização de mecanismos inteligentes de descrição e de semântica para enriquecer conteúdo (metadados, categorização e relacionamentos);
 3. Acesso às informações na forma de consultas e Recuperação da Informação (RI) obtendo resultados correspondentes às expectativas, inclui navegabilidade e escolha de tecnologia de pesquisa; e

4. Encontrabilidade, a qualidade do ambiente prover informação, localizável e navegável. Incluía quatro áreas principais: organização da informação, acesso à informação, governança e experiência do usuário.

5. A governança foi o quinto item da lista para apoiar, manter e controlar as mudanças da Arl, prevê um conjunto de perguntas usadas para revisar o projeto dentro lista de verificação.

A Arquitetura da Informação pode consolidar essa relação com argumentos da classificação da informação adquirida, da estruturação de metadados, da modelagem de ontologias auxiliando a gestão do conhecimento sobre os ativos organizacionais multimodais (BASTOS, 2022). As diversas literaturas sobre Arquitetura da Informação demonstram a fundamentação de visões arquiteturais que exibem pontos de vistas distintos e complementares da Arquitetura da Informação:

Enquanto disciplina, a Arquitetura da Informação (SIQUEIRA, 2012) ajuda a identificar os interesses e necessidades informacionais dentro de um contexto; como princípio e procedimentos a Arquitetura da Informação ajuda no gerenciamento da complexidade enfrentada pelas partes interessadas e na compreensão dos sistemas informacionais (ORLANDI, 2019); enquanto modelo arquitetural, contribui na construção de artefatos que transformam a abstração em conceitos concretos e formais, regras, justificativa e associações, melhora a documentação (BEIRA *et al.*, 2017), reduzindo o risco de desenvolvimento desde sua estrutura até a apresentação e posterior manutenção.

2.1.4 Ontologia

A ontologia descreve explicitamente uma conceitualização compartilhada, uma interpretação estruturada de uma parte do mundo que as pessoas usam para pensar e se comunicar (GRUBER, 1993). Fornece contexto e significado aos dados e é essencial na extração e na reutilização do conhecimento, permite uma solução robusta para o problema da interoperabilidade sintática e semântica que dificulta a troca de informações em sistemas heterogêneos (HARIDY *et al.*, 2023).

De modo geral, a ontologia é utilizada por sistemas de gestão baseados em conhecimento, para dar suporte aos processos organizacionais usando ferramentas e métodos em computador (MIRANDA; MARCELINO; SILVA, 2023). As ontologias são compostas de “conceito”, um modelo; uma expressão humana do mundo real, semelhante ao significado de “classe” na orientação a objetos; um “relacionamento”, um conceito entre conceitos ou uma associação de classes; “instância”, o elemento básico do conceito ou um exemplo concreto; “função”, uma descrição abstrata do método; “axioma”, um fato reconhecido ou regra de inferência (YANG, 2020).

A engenharia de ontologia engloba todas as atividades do processo de desenvolvimento da ontologia desde o ciclo de vida, incluindo metodologias, ferramentas e linguagem de construção (KORST, 1997). Para construir uma ontologia, diversas metodologias foram desenhadas e conceberam os alicerces das fases e tarefas de definir o escopo, especificar requisitos, modelagem, formalizar, implementar, validar e evoluir o modelo. O quadro abaixo resume ciclo de vida de construção de uma ontologia compartilhadas por várias metodologias que serão discutidas a seguir:

Quadro 2 – Ciclo de vida da ontologia

Etapa	Objetivo	Descrição
Escopo, Requisitos e Conhecimento especializado	Plano de Projeto	Fornecer orientação do projeto e tomada de decisão.
	Questões de Competência (QC)	Estabelecem uma caracterização rigorosa dos problemas do domínio e das tarefas que devem ser atendidas pelo modelo e são formuladas a partir das atividades, recursos, estados dos recursos, quantidade, custos e tempo dos vários componentes do sistema (Grüninguer; Fox 1995).
Modelagem	Conceitualização	Trata de uma interpretação estruturada de uma parte do mundo, “daquilo que existe e que pode ser quantificado” (Quine, 1961 <i>apud</i> Kors, 1997). Conceito é uma expressão inequívoca e explícita do domínio com propriedades, relacionamentos, cardinalidade, valores, afirmações e pré-condições (Vizcaino <i>et al.</i> , 2004).
Formalização e Implementação	Linguagem lógica e computacional	É a “identificação e representação formal e implementação computacional dos conceitos de uma determinada perspectiva” (Fox <i>et al.</i> , 1993).
Validação	Generalidade	Que tarefas e soluções compartilham do modelo? Que conceitos abrange? (Fox <i>et al.</i> , 1993)

	Competência	Até que ponto apoia a resolução dos problemas no domínio? (Fox <i>et al.</i> , 1993)
	Eficiência	O modelo suporta um raciocínio eficiente? Fox <i>et al.</i> , 1993)
	Perspicuidade	O modelo é facilmente compreendido pelos usuários? (Fox <i>et al.</i> , 1993)
	Transformabilidade/ Extensibilidade	O modelo pode ser facilmente adaptado para responder aos novos problemas e tarefas? (Fox <i>et al.</i> , 1993)
	Granularidade	O modelo suporta raciocínio em vários níveis e facetas? (Fox <i>et al.</i> , 1993)
	Escalabilidade	O modelo alcança respostas aos grandes projetos no domínio? (Fox <i>et al.</i> , 1993)
	Integração	O modelo pode ser utilizado em conjunto com outros modelos publicado ou outros sistemas tecnológicos? (Fox <i>et al.</i> , 1993)
Evolução	Manutenção	Trata das atualizações do modelo e/ou incluir novos conceitos e relações (Haridy <i>et al.</i> , 2023).

Fonte: Adaptado de Grüninguer; Fox (1995), Quine (1961 apud Kors, 1997), Vizcaino et al. (2004), Fox et al. (1993) e Haridy et al. (2023).

Segundo os autores Haridy *et al* (2023), não existe “a melhor metodologia de desenvolvimento da ontologia”, mas sim uma mais adequada a partir do objetivo, das características do domínio e da fonte de conhecimento, essa metodologia adequada pode vir de uma combinação de duas ou mais metodologias ou adaptações de uma única com um número maior ou menor de etapas resgatando assim, os melhores recursos de cada uma delas. Por isso, existem inúmeras metodologias que apresentam diferentes perspectivas do processo e do foco combinando as fases descritas acima, expandindo-as com novas, incluindo tarefas e ferramentas de gestão e de Arquitetura da Informação. A seguir, foram explorados em fontes de pesquisa acadêmicas, exemplos de estudos com metodologias relevantes e aprimoradas com instrumentos tecnológicos, como por exemplo algoritmos de de Arquitetura da Informação: redes neurais, *clusters*¹, regras de associação, PLN, entre outros.

Quadro 3 – Metodologias para construção de ontologias

Metodologia	Desenvolvimento
-------------	-----------------

¹ Categorização (*clustering*) é uma das técnicas mais utilizadas no processo de mineração de dados para descobrir grupos e identificar distribuições de padrões ocultos em uma base de dados. Disponível em: https://www2.dbd.puc-rio.br/pergamum/tesesabertas/0116381_03_cap_03.pdf

<i>Methontology,</i> Fernández; Gómez Pérez; Juristo (1997)	<p>1º. Definição da proposta: definição do escopo, especificação dos requisitos, identificação dos recursos informacionais e ativos organizacionais que contém conhecimento específico;</p> <p>2º. Modelagem: possui atividades de levantamento, definição e classificação dos termos de interesse, desenvolvimento das relações entre os termos, definição de axiomas e regras do domínio e por fim criação das instâncias no modelo;</p> <p>3º. Formalização do modelo para fins de implementação e manutenção;</p> <p>4º. Avaliação: um exame técnico do modelo em relação a especificação dos requisitos; e</p> <p>5º. Documentação: os artefactos criados devem documentar todas as atividades do desenvolvimento.</p>
<i>101 Method ou</i> <i>Ontology</i> <i>Development 101,</i> Noy; McGuinness (2001)	<p>Apresenta uma prática iterativa incremental, com os seguintes passos para o desenvolvimento:</p> <p>1º. Determinar domínio, escopo, requisitos e usuários;</p> <p>2º. Verificar o reuso de outras ontologias já publicadas em repositórios;</p> <p>3º. Determinar termos relevantes: coleta e análise de termos do domínio e agrupá-los de acordo com sua representação do domínio.</p> <p>4º. Definir a hierarquia das classes;</p> <p>5º. Definir as propriedades das classes: definir atributos e relações;</p> <p>6º. Definir características das propriedades das classes: valores permitidos, tipos de dados etc.;</p> <p>7º. Criar instâncias: identificar a que classe cada instância pertence preenchendo as propriedades da classe de forma distinta.</p>
<i>On-to-Knowledge,</i> Sure; Studer (2003)	<p>Apresenta um ciclo de vida da fase inicial à fase de manutenção utilizando instrumentos de gestão e da engenharia da computação:</p> <p>1ª. Fase: estudo da viabilidade da ontologia com a definição do domínio, escopo, forma de aquisição do conhecimento, seleção de ferramentas, identificação de pessoas envolvidas;</p> <p>2ª. Fase: (<i>Kickoff</i>) buscar por de ontologias já publicadas para reuso, buscar fontes de conhecimento e definir requisitos, especificar questões de competência;</p> <p>3ª. Fase: refinamento e reprodução da ontologia em linguagem formal;</p> <p>4ª. Fase: avaliação e julgamento da ontologia e sua especificação (requisitos, questões de competência, documentação etc.</p>
<i>OntofoInfoScience,</i> Almeida (2020)	<p>1ª. Fase: especificação com definição do domínio, escopo, propósito geral e requisitos em questões de competência;</p>

	<p>2ª. Fase: aquisição do conhecimento com ativos organizacionais, entrevistas com especialistas, análise textual de referências e extração terminológica;</p> <p>3ª. Fase: fundamentação ontológica: busca por ontologias que possam servir como ponto de partida;</p> <p>4ª. Fase: organização taxonômica para criar uma estrutura de classificação dos termos da ontologia e seus significados;</p> <p>5ª. Fase: formalização em linguagem lógica, descrição formal do domínio;</p> <p>6ª. Fase: avaliação a partir de critérios de validação e verificação.</p>
NeOn-Ontology, Suárez-Figueroa; Gómez-Pérez; Fernández-López (2015)	<p>Prioriza reuso dos recursos e evolução contínua do produto:</p> <p>1º. Cenário desenvolve a ontologia sem utilizar recursos ontológicos existentes considerando as fases de: aquisição do conhecimento, especificação dos requisitos, conceitualização, formalização, implantação;</p> <p>2º. Os demais cenários buscam, inicialmente e em diferentes níveis, identificar recursos ontológicos ou não que possam colaborar efetivamente por meio de reuso, reengenharia e reestruturação para construção da ontologia.</p>
ON-ODM – <i>Ontology Development Methodology</i>, Haridy et al., (2023)	<p>Melhorias de qualidade na fase de conceitualização:</p> <p>1º. Aquisição dos Requisitos: identificação do domínio, análise e especificação em QC;</p> <p>2º. Conceitualização: especificação dos conceitos, definição em diagrama de classes da UML, formalização em linguagem lógica OWL, implementação com PROTÉGÉ;</p> <p>3º. Enriquecimento por meio de PLN:</p> <ul style="list-style-type: none"> a) Utilização de um Corpus texto para candidatos a novos termos; b) Segmentação de frases e busca das classes; c) Tokenização das frases e atribuição de tipos sintáticos aos tokens; d) Extração dos verbos para futuros relacionamentos entre classes; e) Lista final de candidatos para enriquecer a ontologia. <p>4º. Avaliação: verificação baseada em QC e em métricas (precisão, coesão, compreensão, concisão);</p> <p>5º. Publicação;</p> <p>6º. Manutenção;</p> <p>7º. Documentação.</p>
AHIDO – <i>Atlas of Human Infectious Disease Ontology</i>, Ghazi et al., (2023)	<p>Prioriza a interoperabilidade semântica dos dados. Os parágrafos longos são transformados em frases-chaves e os metadados são explorados na definição semiestruturada do <i>Atlas of Human</i>: nome,</p>

	classificação, sinônimo, agentes, reservatório, vetor, transmissão, período de incubação, testes, terapia, prevenção entre outros. 1º. Passo: extração das frases-chaves do texto; 2º. Passo: recuperação dos termos e relações taxonômicas (subclasses e superclasses) ou não taxonômicas (relação literal); 3º. Passo: construção da ontologia; 4º. Passo: enriquecimento com outras ontologias já firmadas.
--	--

Fonte: Adaptado de Fernández; Gómez Pérez; Juristo (1997), Noy; McGuinness (2001), Sure; Studer (2003), Almeida (2020), Suárez-Figueroa; Gómez-Pérez; Fernández-López (2015), Haridy *et al.*, (2023) e Ghazi *et al.*, (2023).

Ainda existem muitas possibilidades de métodos de construção em aberto quando se fala da colaboração entre AI e ontologias, como uma associação do raciocínio indutivo do AM com o raciocínio dedutivo das ontologias e resulta em um paradigma híbrido da Arquitetura da Informação, apresenta uma capacidade aprimorada de dedução, expressividade e decidibilidade das ontologias e da interpretabilidade da Arquitetura da Informação (HARIDY *et al.*, 2024). Isso facilita no mecanismo de conceitualizações compartilhadas, interoperabilidade dos dados e construção de um modelo formal e abrangente de regras.

A ontologia fornece uma compreensão compartilhada de um domínio que pode ser comunicado através de sistemas heterogêneos, distribuídos ou semiestruturados e, assim, facilita o compartilhamento e a reutilização de conhecimento entre esses sistemas (CHOUKRI, 2014, HAMOUDA; CHOURABI; BOUGHZALA, 2016).

Chang, *et al.* (2020) afirma que o conceito de Ontologia está ligado à:

1. Compartilhar conhecimento;
2. Recuperar informação;
3. Integrar informação;
4. Gerir o conhecimento;
5. Apoiar o processo de decisão; e
6. Aumentar a usabilidade da Arquitetura da Informação.

A ontologia, como instrumento para validar e compartilhar o conhecimento, integrar bases informacionais com estruturas e tecnologias diferentes e, pode-se entender que até mesmo para integrar termos de domínios diferentes, torna-se um

apoio versátil, eficaz e seguro para a Arquitetura da Informação dispor dos elementos necessários à satisfação e atendimento às necessidades do usuário.

Da literatura analisada destacam-se três ontologias na área de cerveja e comercio a seguir detalhadas:

- **Standaert; Yaroslaski; Castro (2021)** apresentam uma ontologia de tipos de cerveja que trata a discrepâncias na descrição e nos rótulos das cervejas e oferece recomendações de cerveja a partir de preferências de teor alcoólico, amargor, doçura, cor e ingredientes fornecidos por especialistas em cervejas.
- **Warren (2024)** em *The Beer Ontology* também apresenta uma ontologia de cervejas com estilos, legislação, fabricação, recipiente de fabricação e recipientes de embalagem das cervejas, resultando em um inventário para ajudar na tomada de decisão para seleção de cervejas e importante para esse trabalho. Esta ontologia definiu uma superclasse para *Packing* que estudou e definiu classes de tipos de embalagens como ‘barril’, ‘garrafa’, ‘lata’ e instâncias com suas volumetrias e empacotamentos: ‘12 garrafas de 355ml’ ou ‘12 latas de 355ml’. Essas descrições são encontradas em muitas transações das NFC- e analisadas neste trabalho.
- **Schulze et al. (2021)** trata de uma ontologia de compra e venda de produtos que acompanha o processo do pedido de compra do produto, emissão de nota fiscal, envio do produto, recepção e pagamento. Sua relevância para este trabalho é a conceitualização dos campos da nota fiscal como item do produto, preço, códigos, quantidade, descrição etc. e que serão formalizadas como classes na ontologia, o trabalho também apresenta uma lista de QCs significativa para o contexto da fiscalização como: qual número da nota fiscal; quais itens são listados na nota fiscal; qual quantidade e preço dos itens; e quais atributos do item produto.

O Quadro resume ontologias analisadas neste trabalho:

Quadro 4 – Endereços eletrônicos de ontologias relacionadas com a pesquisa

Endereço eletrônico	Fonte
---------------------	-------

https://beer-advisor--rpi-ontology-engineering.netlify.app/oe2020/beer-advisor/usecase	Schulze <i>et al.</i> (2021)
https://purl.org/p2p-o	Standaert; Yaroslaski; Castro (2021)
https://rdf.ag/o/beer-en.html	Warren (2024)

Fonte: Dados da pesquisa (2024).

Tendo em vista a necessidade de organizar e representar conhecimento, o desenvolvimento de um modelo requer identificar ontologias existentes no domínio de estudo e reconhecer os padrões de projeto úteis, com classes e instâncias, para serem reutilizados em um novo modelo de ontologia. Os recursos coletados serão aplicados fazendo adaptações e extensões para aprimorar a conceitualização, lista de termos padronizados e vocabulário sobre conceitos, seus relacionamentos, atividades e regras existentes considerando o domínio, o idioma de desenvolvimento da ontologia, o nível de formalismo e a integridade do conhecimento compartilhado.

Neste estudo, que perpassa por um cenário envolvendo muitos sistemas de informação de NFC-e, uma enorme massa de dados armazenados, formas diferentes de definição e expressão de um produto, falta de metadados, falta de recursos materiais e humanos para tratar os desafios dentre outros, a abordagem ontológica pode colaborar no objetivo da pesquisa que é o auxílio aos processos de recuperação da informação e representação do conhecimento no campo da auditoria e fiscalização da SEFAZ/AM com:

1. Ampliação do escopo dos produtos descritos na NFC-e;
2. Promoção de um mecanismo de acesso à informação tornando acessível e inteligível um vocabulário complexo; e
3. Promoção de melhoria na transparência da informação.

2.2 Bases teóricas da Ciência da Computação

Nesta etapa do trabalho, busca-se resguardar a importância da Ciência da Computação na construção da ontologia. Sabe-se que o termo ontologia na referida

ciência pode se tratar de: (a) um vocabulário expresso em uma linguagem de Rede de Computadores (corresponde à um software, um artefato computacional), ou (b) a uma teoria em que os fenômenos são explicados considerando fatos e regras (mantém a noção filosófica, um inventário de coisas do mundo e relações entre elas em um domínio particular). Nesta discussão, buscar-se-á esclarecer as bases da linguagem de redes de computadores de relevância para se alcançar os objetivos desta tese.

2.2.1 Linguística: base para a Ciência da Computação

Busca-se nessa sessão estabelecer um cenário geral da linguística para que se adentre mais adiante nas especificações linguísticas no campo computacional. Inicialmente, a linguagem natural é síntese do pensamento humano que busca por meio de padronização e univocidade a representação da realidade e intermediação do processo de comunicação. Pode-se afirmar que:

A linguagem é o fenômeno da comunicação entre os seres humanos em geral [...] compreende as palavras, sua pronúncia e os métodos de combiná-las usados e compreendidos por uma comunidade, expressa ideias e sentimentos pelo uso de sinais, gestos, sons ou marcas impregnadas de significados. (HARALAMBOUS, 2024, p. 25, tradução nossa).

A linguística, por sua vez, é a ciência para o estudo da linguagem e para sua expressão nas diferentes línguas do mundo, as línguas naturais (CHOMSKY, 1986, COSERIU, 1986). Nessa perspectiva, a língua é uma manifestação concreta da linguagem e é considerada um sistema de signos cujas partes possuem conceitos (significados) em conjunto, ou seja, inter-relacionados (MACULAN; LIMA, 2017).

A CI tem estreita ligação com a linguística pela intermediação da análise documentária, que se utiliza de métodos e processos para descrever o conteúdo dos documentos (MENDONÇA, 2000, p. 51). O autor afirma que as relações construídas da análise documentária e linguística aplicada são subsídios para a Terminologia como subárea e o estudo da construção de vocabulários para fins de documentação, sendo assim “[...] um campo da linguística que se ocupa de conceitos, de termos e denominações, estruturas de representações linguísticas”.

Terminologia identifica termos próprios de uma área, isto é, estuda termos especializados (de uma área particular) (MACULAN; LIMA, 2017). Definir um termo é ter domínio da especialidade que ele abrange. Considera-se, de acordo com Lara, (2004), que “é um signo linguístico que difere da palavra, unidade da língua geral, por ser qualificado no interior de um discurso de especialidade”. Dessa forma, um termo será considerado de acordo com o uso, o emprego dele no contexto específico e considerando sua especialidade.

[...] a Terminologia trabalha com o estudo científico dos conceitos e respectivos termos, que constituem um conjunto expressivo e comunicativo, possibilitando a transferência do conhecimento especializado. Neste sentido, os conceitos não existem isoladamente, mas sempre uns em relação aos outros (MAIMONE; SILVEIRA; TÁLAMO, 2011, p. 22).

A palavra, segundo Maculan e Lima (2017), é a unidade léxica que pode ser analisada sob o ponto de vista da expressão linguística – o termo, e do seu conteúdo – o conceito. Quanto ao conceito, ou noção, determina o conteúdo semântico do termo que o designa, de modo que:

[...] O conceito é convencionado no contexto de uso, articulado pela comunidade que o compartilha, para, depois, se proceder à sua designação por meio de um termo que o represente, numa monossignificação ou monorreferencialidade (um termo só pode representar um conceito), típicas do discurso científico. (MACULAN; LIMA, 2017, p. 61).

Na definição do termo, é preciso superar problemas como heterogeneidade, ambiguidades e polissemias já que toda forma de representação, por mais aprimorada que seja a linguagem, não é igual a realidade, sendo necessário, portanto, buscar uma linguagem mais apropriada, de forma que a sua precisão aproxime a representação da realidade.

Nesse sentido, o estudo da estrutura hierárquica do conceito tem o objetivo da precisão e de aproximação, uma vez que tenta reproduzir o processo cognitivo de interpretação que o ser humano faz para assimilar informação e transformá-la em conhecimento. Na estruturação do conceito, a reorganização do pensamento é estimulada a compreender o conteúdo apresentado pelo conceito, na medida que a sintaxe descreve a estrutura da linguagem e a semântica descreve o comportamento e o significado (NOVAES, 2011).

Segundo os autores Gonzalez e Lima (2003, p. 7), o significado é entendido como:

[...] o sentido da linguagem corrente, como sentido intuitivo, e podem ser relacionadas três funções da informação semântica codificável em enunciados linguísticos: (a) o significado descritivo, que pode ser objetivamente verificado; (b) o significado social, a partir das relações sociais; e (c) o significado expressivo, dependente do locutor.

Os objetos não são, por si só, informativos, pois dependem das pretensões do conhecimento dos sistemas informacionais. Segundo Capurro e Hjørland (2007) é necessário que outras linguagens, como as palavras, sejam utilizadas para ajudar a definir ou representar algo que se pretende alcançar. O campo lexical ou conjunto de palavras e o campo semântico, emprego das palavras, abrangem e estruturam um conceito.

Para que se tenha o uso real de um conceito, é preciso que a definição e as relações entre eles sejam bem definidas. Um conceito pode ser caracterizado por um conjunto de atributos de definição, propriedades definidoras dos termos. Abbagnano (2015, p. 803) define propriedade como “[...] uma determinação que pertence a *toda* uma classe de objetos, pertencendo sempre e somente a essa classe, mesmo que não faça parte de sua definição”.

Termo, então, seria um conjunto de propriedades essenciais que determinam a aplicabilidade de um termo, (NOVAES, 2011). Complementa a autora: “[...] conceito é uma noção abstrata contida nas palavras de uma língua para designar de modo generalizado e de certa forma estável as propriedades e características de uma classe, seres, objetos ou entidades abstratas” (NOVAES, 2011, p. 241).

Na Teoria do Significado de Wittgenstein (1958), a definição de termos é feita a partir de um contexto determinado pelo uso e pelo emprego na realidade das pessoas, ou seja, do emprego deles no cotidiano. Para o autor, as propriedades necessárias e suficientes para definição do termo não delimitam a existência de outras, pois nem todos os membros de uma família compartilham de todos os seus atributos.

O uso de um termo faz inferir a respeito de novas descobertas e consequentemente novas propriedades, e é possível agrupar características comuns

a um conjunto de seres da mesma família, diferenciando-os de seres pertencentes a outras famílias. Para Novaes (2011), o conceito representa um objeto por meio da reunião de características gerais e é expresso por termos ou predicados. Pelo conceito é possível agrupar características comuns a um conjunto de seres da mesma espécie, diferenciando-os dos seres pertencentes a outra(s) espécie(s). Ainda sobre tal questão a autora enfatiza:

Todas as vezes que vemos alguma coisa com um tipo de coisa, ou como parte de alguma coisa, estamos categorizando. Isso ocorre, principalmente, pelas características oriundas das similaridades e diferenças existentes entre conceitos, dentro de determinado contexto (NOVAES, 2011, p. 241).

A notação classificatória, hierárquica, a organização em categorias, relações lógicas de parte-todo e gênero-espécie, é parte do processo de formação dos conceitos: a identificação de conceitos e a sua estruturação e ordenação em classes, de acordo com as categorias existentes em um dado domínio. Categorizar é um mecanismo que simplifica a atividade de compreensão e aprendizado, memorização e armazenamento, recuperação da informação e compartilhamento do conhecimento.

Dentro de um determinado contexto, a análise conceitual faz parte de um processamento linguístico natural, essencial para representação do conhecimento, envolvendo seleção, extração, análise e organização de documentação, critérios dependentes do uso.

Há, portanto, entre a Linguística e a CI uma interseção de recursos utilizados sob três aspectos, principalmente na forma textual: registro pela linguagem, organização e representação da informação e ponto de partida para gerar, descobrir e disseminar o conhecimento (BAPTISTA, 2015). Entende-se, desse modo, que:

[...] O texto pode ser entendido como uma macro unidade, composta de informações de diversas naturezas, presente na estrutura de uma língua natural [...] ele apresenta outras informações, como por exemplo, o material necessário para a produção de representações mentais na forma de letras e palavras [...] as informações ativadas no acesso lexical permitem ao leitor a construção da estrutura sintática das frases, orações e período. A capacidade de relacionar essas informações faz parte do **processo de compreensão da linguagem** (DUQUE, 2005, p. 19, grifo nosso).

A intensificação de trabalhos nestas áreas para analisar e interpretar a capacidade cognitiva e outras atividades específicas do ser humano resultam em

estudos de formalização das línguas naturais próximas à Lógica e Matemática em áreas como cibernética e inteligência artificial com base no desenvolvimento de processamento automático da linguagem natural, a Linguística Computacional (LC) (FARIAS, 1998).

Essa evolução contribuiu com sistemas cada vez mais adaptados as necessidades de produção e busca de informação e com ampliadas possibilidades de compreensão, visto que o processo de cognição passa não somente pelo contexto, mas pelo valor que a informação possui e pela forma de expressar essa informação, da linguagem textual por exemplo.

Na atualidade, grande parte das diferentes formas de organização de banco de dados na informática visa armazenar material – muitos deles, linguísticos – e disponibilizá-lo ao usuário/autor, a fim de que tenha uma gama de recursos para suas produções na constituição do seu discurso (NOVAES, 2011, p. 239).

A LC busca uma forma de padronizar o termo e de transformar a linguagem natural em um artefato que possa ser processado, leia-se compreendido e produzido, na forma computacional de maneira útil e simples (OLIVEIRA, 2020), formando estruturas gramaticais (sintáticas e semânticas) para caracterizar línguas naturais, visando à implementação computacional com competências semelhantes à humana no diálogo, na aquisição e na recuperação da informação e obtenção de conhecimento.

Sendo assim, a Linguística Computacional pode ser didaticamente dividida em duas subáreas: a Linguística de Corpus e o PLN (OTHERO, 2006). Sobre as subáreas, o autor afirma que Linguística de Corpus se preocupa com a coleta e exploração de conjunto de dados que contenham amostras de linguagem natural como falada, escrita literária etc., geralmente em formatos eletrônicos, coletados com o propósito de servirem para o estudo de determinados fenômenos linguísticos e sua ocorrência em uma língua ou variedade linguística. Já PLN dedica-se a construção de softwares capazes de interpretar e/ou gerar informações em linguagem natural, exigindo vários subsistemas para abranger os diferentes aspectos da língua: sons, palavras, sentenças e discurso nos níveis estruturais, de significado e de uso.

2.2.2 Processamento de Linguagem Natural

PLN é um ramo da linguística que estuda a geração e recepção automática de texto, fazendo com que máquinas sejam capazes de ler, escrever, traduzir e interpretar textos (DUQUE, 2005). O PLN utiliza conjunto de técnica computacional para processamento de texto, permitindo a comunicação entre pessoas e máquinas, envolvendo áreas da linguística como fonologia, morfologia, sintaxe, semântica, pragmática e a área da inteligência computacional (CHANDRA *et al.*, 2022), revelando a significância contextual das palavras usadas no documento (MBOLI *et al.*, 2021).

Algumas técnicas de PLN já desenvolvidas estão implementadas em código na forma de “funções” e disponíveis em bibliotecas acopladas a *frameworks*² e utilizadas para desenvolvimento de sistemas inteligentes ou aplicação direta em massa de dados de diversos formatos (sons, imagens, palavras, sentenças, discursos etc.).

Várias funções foram desenvolvidas para dar suporte ao comportamento da atividade humana em relação ao conteúdo textual. Em geral, o significado pode ser construído considerando sua composicionalidade do texto, as expressões idiomáticas, o contexto e o uso (HARALAMBOUS, 2024). A avaliação pode ser feita sob diversos pontos de vista: fonético, morfológico, sintático, semântico e até mesmo pragmático. A estrutura textual é regida pelas regras gramaticais e categorias morfológicas do léxico, a seguir, a sentença é processada morfo-sintaticamente e semanticamente e cada palavra é identificada separadamente e relacionamentos semânticos possíveis são identificados entre palavras e frases (GONZALEZ; LIMA, 2003).

O PLN apoia-se na estratégia do processamento, buscando obter uma forma lógica adequada para representar o conhecimento, no campo linguístico as estratégias são:

- **Etiquetagem do texto:** são técnicas de reconhecimento de palavras relevantes *Relevant Words Recognition* (RWR) consistem em identificar e classificar entidades nomeadas e outros componentes com valor sintático e/ou semântico

² Coleção de componentes de software reutilizáveis que tornam mais eficiente o desenvolvimento de novas aplicações. Disponível em: <https://aws.amazon.com/pt/what-is/framework/>

significativo em textos e pode ser dividido em *Named Entity Recognition* (NER), *Named Entity Desambiguation* (NED), *Shallow Parsing* e *Parts-Of-Speech (POS) Tagging* (SORATO, *et al.* 2016). Para melhor compreensão, segue a especificidade de cada uma no processo.

- Reconhecimento de Entidade Nomeada ou *Named Entity Recognition* (NER) é uma tarefa de PLN que consiste reconhecer e extrair menções a entidades significativas em passagens de texto. Baseiam-se em regras específicas para classificação em conjunto de categorias sintáticas ou semânticas apoiadas por recursos de dicionários (SCHEIDER *et al.*, 2020, ALBUQUERQUE *et al.*, 2023).
- Desambiguação de Entidade Nomeada ou *Named Entity Disambiguation* (NED) é a desambiguação de sentido de uma palavra (do inglês *word sense disambiguation*), é a atividade de encontrar, dada a ocorrência em uma palavra ambígua no texto, o sentido específico daquela ocorrência de acordo com o contexto em que ela ocorre (SORATO *et al.* 2016; FACELI *et al.* 2021), detecta a diferença de significado das palavras com base no contexto (MBOLI *et al.* 2021).
- *Parts-Of-Speech (POS) Tagging* é a estratégia de PLN envolvendo conhecimento linguístico na busca de significado, no tratamento da ambiguidade e outros desafios. Estabelece classificações da estrutura gramatical dos termos e permite compreensão não apenas de termos individuais, mas também as relações entre eles dentro da sentença, uma vez que foi treinado com base em uma vasta quantidade de dados linguísticos (CAMBRIA; WHITE, 2014). Esta estratégia alcança o etiquetador morfológico *DefaultTagger*, etapa básica do POS e usa classes gramaticais simplesmente para atribuir uma classificação às palavras (SORATO *et al.* 2016), identifica a constituição de palavras ou grupos de palavras que formam elementos de expressão de uma língua: substantivos, verbos etc. (GONZALEZ; LIMA, 2003); e o etiquetador sintático que identifica os vários modos de combinar regras gramaticais com a finalidade de gerar as possíveis estrutura sintáticas do texto.
- O analisador *Shallow Parsing* é o procedimento que avalia os vários modos de como combinar regras gramaticais, com a finalidade de gerar uma estrutura de

árvore que represente a estrutura sintática da sentença analisada. Se a sentença for ambígua, o analisador sintático (*parser*) irá obter todas as possíveis estruturas sintáticas que a representam (FACELI *et al.*, 2021).

- Existem outras abordagens para aprimorar a classificação dos termos com o *Rule-Based Tagging* e *Transformation-Based Tagging* (TBT), com marcações baseadas em regras pré-determinadas que podem, inclusive, alterar a classificação de um termo dependendo das informações contextuais, e o *Statistical POS Tagging*, técnica de linguística computacional que coloca categorias gramaticais em palavras do texto baseada em modelos probabilísticos e aprendizado de máquina com algoritmos como *Conditional Random Fields* (CRF) e *Hidden Markov Models* (HMMs) (CAMBRIA; WHITE, 2014).
- **Normalização de variáveis linguísticas:** *Stemming* e *Lemmatization* são filtros mais comuns que permitem a extração dos afixos das palavras reduzindo-as às suas raízes (MBOLI *et al.* 2021). Destacam-se a seguir suas funções:
 - Estemização ou *Stemming* explora as similaridades morfológicas inferindo proximidades conceituais, conseguindo reduzir todas as palavras com mesmo radical a uma forma denominada *stem* (similar ao próprio radical), eliminando afixos de derivação e flexão (HARALAMBOUS, 2024).
 - Lematização ou *Lemmatization* reduz à forma canônica, como por exemplo, reduz verbos ao infinitivo e adjetivos e substantivos à forma masculina e singular (HARALAMBOUS, 2024).
- **Eliminação de Stop Words:** filtro para palavras irrelevantes na interpretação textual chamadas de palavras funcionais: artigos, preposições, conectivos etc. (GONZALEZ; LIMA, 2003). Com a eliminação das *Stop Words* corre-se o risco de eliminar a estrutura composicional da expressão, mas quando se busca termos específicos de um domínio e suas relações, a eliminação de ruídos é uma etapa importante da PLN.

- **Métodos estatísticos:** é uma abordagem quantitativa que anexa informações numéricas às variáveis linguísticas. Quaisquer que sejam os critérios para representação e classificação de conteúdos, ao final a necessidade é obter ou recuperar a informação, entender como ela se apresenta através de consultas a um conjunto de dados. A representação de conteúdos no formato computacional de uma única palavra, chamadas de unigramas, independente da sequência de palavras, é chamada de *bag of words* (AGGARWAL, 2014). Os valores para representar o termo em um documento geralmente estão associados na frequência de ocorrência do termo mudando apenas a estratégia de atribuição de peso aplicada (FACELI *et al*, 2021):
- **Extração de *tokens* ou segmentar dados linguísticos** em *tokens* é a divisão ou decomposição de dados textuais em componentes menores e significativos, unidades de palavras, um processo que usa autômatos finitos e elimina informações e caracteres desnecessários. (MBOLI *et al.*, 2021). Os *tokens* são sequências de grafemas separadas por espaços em branco (HARALAMBOUS, 2024), *tokenização* é equivalente à detecção de limites entre *tokens*, a segmentação.

Além disso, é bom destacar alguns pontos importantes em relação ao PLN, a Lei de Zipf, por exemplo, estabelece a constante de rank-frequency obtida em um texto pelo produto $\log(f_T) \times K_t$, onde f_T é o número de vezes que o termo T ocorre no texto e K_t é a posição do termo T em relação a todos os termos do texto, em uma análise morfológica e semântica, ordenados pela frequência de ocorrência (GONZALEZ; LIMA, 2003, HARALAMBOUS, 2024). A extração do Gráfico de Luhn é uma forma útil de verificar a frequência de ocorrência de um termo e analisar a sua expressividade. Termos que se repetem com muita frequência ou termos que nunca se repetem não necessariamente representam uma expressão de interesse, sobrando como expressivas palavras com frequência intermediária.

Colocação ou *Collocations* é uma sequência de palavras que aparecem juntas, utilizadas em algoritmos de análise de termos frequentes e relações entre eles como

do tipo: *if-then*³, *is-a*⁴, obtendo uma medida para a força da colocação, isto é, probabilidade dos pares de palavras aparecerem em determinado texto e se a segunda palavra do par é dependente ou não da primeira. Técnicas como *N-grams* generalizam *Collocations* para análise de 'n' termos frequentes e são utilizadas para busca de um conjunto de palavras que se repetem propondo uma palavra consequente a partir das palavras que a antecedem (AGGARWAL, 2014).

Semântica vetorial ou *Vector Semantics* é um meio algébrico que calcula o relacionamento semântico entre duas palavras, a probabilidade condicional de uma combinação, de qualquer complexidade, das palavras que ocorrem em contextos semelhantes possuírem significados relacionados como “carro-gasolina”, o parentesco semântico é obtido pelo produto escalar dos vetores e é um tipo de representação de conteúdo distributiva, também conhecida como modelo espaço-vetorial, cada termo pode representar uma ou mais palavras do texto. Quando cada um deles representa um par de palavras do texto, eles são chamados de bigramas (FACELI *et al.*, 2021).

A tarefa de descobrir e classificar termos de interesse no documento, no contexto e para uso específico e que se manifestam de formas distintas ou semelhantes prepara o texto na linguagem computacional permitindo a troca de informações na forma de consultas, a organização da informação, o registro e a comunicação na forma de conhecimento.

2.2.3 Aprendizado de Máquina

Segundo Faceli *et al.* (2021) algoritmos baseados em AM são um ramo da Inteligência Artificial utilizados em diversas tarefas Preditivas e Descritivas. Tarefas

³ Testa uma condição, depois retorna um valor com base no resultado dessa condição. Disponível em: <https://support.zendesk.com/hc/pt-br/articles/4408838560922-Uso-da-fun%C3%A7%C3%A3o-IF-THEN-ELSE>

⁴ Quando uma classe "é" um tipo de outra classe. Por exemplo, um gato é um animal. Disponível em: <https://www.dio.me/articles/entendendo-a-relacao-is-a-e-has-a-em-java-quando-usar-heranca-ou-composicao>

Descritivas extraem padrões de um conjunto de dados, por exemplo: buscando grupos de objetos similares entre si (agrupamento); buscando padrões frequentes de associação entre objetos encontrados (associação); ou buscando uma descrição simples e compacta para esse conjunto de dados (sumarização). A forma de representação da associação busca identificar as relações que existem ou devem existir entre os dados, seguindo a premissa de “[...]encontrar elementos que implicam na presença de outros em uma mesma transação” (SCHUNEIDER, 2002).

Em síntese, a associação procura padrões frequentes de associação entre objetos encontrados (FACELI *et al.*, 2021). Destaca-se, também, que os métodos associativos funcionam como tarefas descritivas, interativas, repetitivas e incrementais, cuja realização pode ser feita por meio de algoritmos de AM.

Os métodos associativos utilizam o paradigma dos algoritmos não supervisionados, ou seja, não dependem de um elemento externo para conduzir o aprendizado na extração de um modelo com boa capacidade descritiva. O aprendizado é dirigido aos dados e o algoritmo de AM busca aprender um modelo, uma regra, e esta regra, por sua vez, deve alcançar e ser válida para novos objetos do mesmo domínio em uso e que não fizeram parte do conjunto de dados utilizados no treinamento. Esta característica é a generalização, capacidade de perpetuar modelo ou regra para novos dados.

Recursos não supervisionados permitem o reconhecimento e a classificação de termos, expressões linguísticas e unidades de informação relevantes no texto e são uma sub tarefa da extração da informação (NADEAU; SATOSHI, 2007). Atividades não supervisionadas permitem, por exemplo, identificar e reunir o termo e similaridades em grupos dentro de um contexto (identificação por dedução e inferência), estabelecer a correlação entre o número de vezes que o termo apareceu (número de ocorrências) e por fim, medir a dependência entre as expressões que o termo aparece (relevância da regra).

Termo importante nesta pesquisa, o Apriori é um algoritmo de associação eficiente, responsável pela mineração de itens frequentes (*ItemSet*) para descoberta do conhecimento em base de dados que contém várias transações e cada uma delas

suportando um conjunto de itens frequentes. O Suporte de um *ItemSet* é a fração de transações que o contém, representado por:

$$\sigma_T(I) = \frac{1}{n} |K_T(I)|$$

A partir de conjunto de itens frequentes, é possível derivar regras de associação entre eles, de natureza probabilística, apresentada na forma *se* antecedente *então* consequente. O grau de incerteza da regra é dado pela confiança da regra:

$$confiança(A \rightarrow B) = suporte(A \cup B) / suporte(A)$$

Para otimizar o grande número de regras de associação, é extraído um padrão de interesse e independência da regra para selecionar os mais relevantes, baseados no princípio que padrões que ocorrem aleatoriamente não são de interesse, são eles o *lift* e a convicção:

$$lift(A \rightarrow B) = confiança(A \rightarrow B) / suporte(B)$$

$$convicção(A \rightarrow B) = [1 - suporte(B)] / [1 - confiança(A \rightarrow B)]$$

2.2.4 Mineração de Dados e Mineração de Texto

Uma grande quantidade de dados organizacionais disposta em sistemas informacionais automatizados é incorporada inicialmente a partir de uma finalidade, contexto e necessidade de seus usuários. Contudo, a dinâmica informacional que esses dados proporcionam quase sempre estão além do planejado e a expectativa criada sobre a recuperação e fluxo informacional provoca uma busca por tecnologias computacionais que dissolva a distância entre a fonte e a necessidade de informação.

Técnicas de Mineração de dados (MD) e Mineração de Texto (MT) permitem a descoberta de conhecimento em bases de dados ou *Knowledge Discovery in Databases* (KDD), isto é, descobrir informações ocultas e associação de padrões desconhecidos ou regras de interesse em uma grande quantidade de dados. A MD é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A partir deles,

é possível a interpretação de dados, itens elementares da informação, dentro de determinado contexto e revestidos de significados para extração do conhecimento.

Dessa forma, a MD é um dos componentes da Descoberta do Conhecimento em Bases de Dados, conhecida como KDD e consiste na aplicação de algoritmos de inteligência artificial para exploração de quantidades massivas de dados (AFONSO; DUQUE, 2020, p. 2). A aquisição do conhecimento para auxílio na tomada de decisão passa pela etapa de busca da informação com entendimento e interpretação dos dados coletados em determinado domínio envolvendo três etapas operacionais do KDD: pré-processamento, Mineração de Dados e pós-processamento:

A etapa de **pré-processamento**, que compreende as funções relacionadas à captação, à organização e ao tratamento dos dados com o objetivo de preparar os dados para os algoritmos da etapa seguinte, a **Mineração de Dados**. Durante a etapa de **Mineração de Dados**, é realizada a busca efetiva por conhecimentos úteis no contexto da aplicação de KDD. A etapa de **pós-processamento** abrange o tratamento do conhecimento obtido na Mineração de Dados. Tal tratamento, nem sempre necessário, tem como objetivo viabilizar a avaliação da utilidade do conhecimento descoberto (FAYYAD *et al.*, 1996^a, grifo nosso).

As técnicas de MD se baseiam em algoritmos de Inteligência Artificial como Redes Neurais ou Aprendizado de Máquina (supervisionado ou não supervisionado) e usa Modelos Estatísticos e Probabilísticos para o tratamento dos dados (GOLDSCHIMIDT; PASSOS, 2005). As tarefas da MD são predição e descrição na busca por padrões de interesse e análise de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). As tarefas se dão na forma representacional de um modelo do tipo classificação, regressão, agrupamento (*clustering*), associação, sumarização, modelagem de sequência, dependência e análise de linhas de tendências.

A MD é obtida pelo conjunto da forma representacional de um modelo, do critério de preferência para sua representação e do método ou algoritmo de busca. Para se realizar a mineração dos dados, Fayyad, Piatetsky-Shapiro e Smyth (1996) indicam seis tarefas:

- Classificação: momento em que se faz a descoberta de uma função que faça o mapeamento (classificação) de um item de dados em um conjunto de classes pré-definidas;

- Regressão: quando se faz a descoberta de uma função que mapeie um item de dados em uma variável de predição de valor real;
- Agrupamento: identificação de um conjunto finito de categorias (*clusters*) que descrevam os dados;
- Sumarização: momento da busca de uma descrição compacta para um subconjunto de dados;
- Modelagem de dependência: busca de um modelo que descreva as dependências mais significativas entre as variáveis;
- Detecção de mudança e desvio: quando se faz a descoberta das mudanças mais significativas nos dados a partir de valores normativos ou previamente medidos.

Já a Mineração de Textos é considerada como variação da MD, realizada em documentos não estruturados ou semiestruturados com o fito de descobrir padrões e associações relevantes (GOLDSCHIMIDT; PASSOS, 2005).

Nos estudos de Moraes e Ambrósio (2007, p. 2), há destaque para as denominações MT como sinônimo de descoberta de conhecimento em textos. Os autores evidenciam que nomenclaturas como Mineração de Dados em Textos (*Text Data Mining*) ou Descoberta de Conhecimento a partir de Bancos de Dados Textuais (*Knowledge Discovery from Textual Databases*), são termos encontrados na literatura, haja vista o processo de mineração de textos ser realizado, também, a partir de técnicas de Descoberta de Conhecimento em Bancos de Dados (KDD) aplicadas sobre dados extraídos a partir de textos. Outros termos utilizados como sinônimo de mineração de textos podem ser: Busca de Informação (*Information Seeking*); Conhecimento Público não Descoberto (*Undiscovered Public Knowledge*); Recuperação de Conhecimento (*Knowledge Retrieval*).

Definida as possíveis nomenclaturas, a definição de mineração de textos pode ser:

Considerada uma evolução da área de Recuperação de Informações (RI), Mineração de textos (*Text Mining*) é um Processo de Descoberta de Conhecimento, que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando

métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado (MORAIS; AMBRÓSIO, 2007, p. 2).

Como principais contribuições desta área estão: busca de informações específicas em documentos, análise qualitativa e quantitativa de grandes volumes de textos, melhor compreensão de textos disponíveis em documentos. Tais textos podem “[...] estar representados das mais diversas formas, dentre elas: e-mails; arquivos em diferentes formatos (pdf, doc, txt, por exemplo); páginas Web; campos textuais em bancos de dados; textos eletrônicos digitalizados a partir de papéis” (MORAIS; AMBRÓSIO, 2007, p. 6).

2.2.5 Metadados

Importante categoria para a descrever os dados, o metadado se configura como uma metalinguagem⁵. Seu conceito não é de simples estabelecimento, mas se define, de maneira geral, como “dados sobre dados”. Tendo sua origem no latim *metá*, que significa “além”, “através de” ou “sobre”, muitas definições são consideradas pelos estudiosos da área porque compreendem que o objetivo de uso do metadados pode especificar seu conceito. Assim, não há uma definição ampla o suficiente para abranger todo seu significado.

Nessa perspectiva, Mori e Carvalho (2004, p. 2) apresentam alguns conceitos válidos para metadados a partir de alguns estudiosos:

“Metadados consistem em dados que descrevem todos os outros dados em um banco de dados.

Metadados são dados que descrevem atributos de um recurso. Eles suportam um número de funções: localização, descoberta, documentação, avaliação, seleção etc.

Metadados fornecem o contexto para entender os dados através do tempo.

⁵ Linguagem de descrição de linguagens, “linguagem (natural ou formalizada) que serve para descrever ou falar sobre uma outra linguagem, natural ou artificial”). Disponível em: <https://www.google.com/search?client=opera&q=metalinguagem+significado&sourceid=opera&ie=UTF-8&oe=UTF-8>.

Metadados são dados associados com objetos que ajudam seus usuários potenciais a terem vantagem completa do conhecimento da sua existência ou características.

Metadados são o instrumental para transformar dados brutos em conhecimento”.

Em relação aos tipos de Metadados, podem ser classificados como tipo estrutural ou tipo semântico. O metadado estrutural “[...] representa a informação que descreve a organização e estrutura dos dados gravados”; já os metadados semânticos, “[...] fornecem informações sobre o significado dos dados disponíveis e seus relacionamentos semânticos” (MORI; CARVALHO, 2004, p. 15). São exemplos de metadados semânticos:

[...] dados que descrevem o conteúdo semântico de um valor de dado (como unidades de medida e escala), ou dados que fornecem informações adicionais sobre sua criação (algoritmo de cálculo ou derivação da fórmula usada), linhagem dos dados (fontes) e qualidade (atualidade e precisão) (MORI; CARVALHO, 2004, p. 15).

Para elaborar um modelo de metadados, Mori e Carvalho (2004, p. 15) entendem que é preciso descrever o contexto da informação de maneira que não seja nem ambígua, nem redundante. Por isso, é aconselhável uma conceitualização de um domínio específico de problema ou ontologias, pois assim ocorrerá um acordo comum de vocabulários de modo que os dados sejam referenciados. Considera-se que determinado “[...] objeto semântico representa um item de dado, junto com sua base de contexto semântico” que, por sua vez, “[...] consiste de um conjunto flexível de meta-atributos que explicitamente descrevem a compreensão implícita sobre o significado do item de dado (MORI; CARVALHO, 2004, p. 15).

Convém destacar que cada objeto semântico detém um rótulo de conceito associado a ele. Esse rótulo de conceito (adquirido de uma ontologia) aponta o relacionamento entre o objeto e os aspectos do mundo real descritos por ele. “A detecção e resolução destas heterogeneidades semânticas, obviamente, requerem conhecimento sobre a exata base semântica dos dados representados” (MORI; CARVALHO, 2004, p. 15). Para ter a correta interpretação dos metadados disponíveis, um domínio específico de ontologias pode ser utilizado.

Considera-se, assim, que os conceitos específicos da “ontologia fornecem vocabulário comum para que nenhuma negociação adicional seja necessária”.

Também, pode-se afirmar que “[...] a ontologia fornece informação sobre a representação do dado descrito sobre a base do modelo”. Em síntese, a situação ideal é quando todas as instâncias de uso dos dados e metadados de um determinado domínio assintam à ontologia correspondente (MORI; CARVALHO, 2004, p. 16).

2.2.5.1 ISO 23081 – Informação e Documentação – Metadados para Registro

Para a normalização de metadados, a *International Organization for Standardization* (ISO) define a norma ISO 23081 – Registro de Metadados. A série ISO 23081 é composta por três documentos publicados em anos diferentes, inicialmente 2006, 2009 e 2011, que atualmente duas partes possuem versões mais recentes (2017 e 2021), conforme apresentadas a seguir:

- ISO 23081-1:2017 Processos de gerenciamento de registros - Metadados para registros - Parte 1: Princípios

Publicada inicialmente em 2006, a atual **ISO 23081-1:2017 Informação e documentação - Processos de gerenciamento de registros - Metadados para registros - Parte 1: Princípios**, é uma norma que “[...] abrange os princípios que sustentam e regem os metadados de gerenciamento de registros” (ISO, 2017, p. 3). A ISO 23081-1:2017 – parte 1 está direcionada aos registros e seus metadados, bem como todos os processos que os afetam, a qualquer sistema em que residam e qualquer organização responsável pela sua gestão (ISO, 2017).

- ISO 23081-2:2021 Metadados para gerenciamento de registros — Parte 2: Questões conceituais e de implementação

A segunda norma da série para metadados é a ISO 23081-2:2021 Informação e documentação — Metadados para gerenciamento de registros — Parte 2: Questões conceituais e de implementação. Teve sua primeira versão publicada em 2009 e foi atualizada em 2021. O documento define métodos práticos de implementação para metadados, além de opções de gerenciamento e um modelo conceitual para definir elementos de metadados para registros.

A **ISO 23081-2:2021 - parte 2** tem como escopo: permitir descrição padronizada de registros e entidades contextuais críticas para registros; fornecer um entendimento comum dos pontos fixos de agregação para permitir a interoperabilidade de registros e informações relevantes para registros entre sistemas organizacionais; e permitir a reutilização e padronização de metadados para gerenciamento de registros ao longo do tempo, espaço e entre aplicativos (ISO, 2021, p. 2).

Além disso, também estão incluídos nos objetivos desta norma, a identificação de alguns dos pontos de decisão críticos que precisam ser abordados e documentados para permitir a implementação de metadados para a gestão de registros, tais como os diversos caminhos para a tomada de decisão implementação de metadados para gestão de registros (ISO, 2021).

- ISO/TR 23081-3:2011 - Gerenciamento de metadados para registros - Parte 3: Autoavaliação

Por fim, o terceiro documento da série metadados é a norma **ISO/TR 23081-3:2011 – Informação e Documentação - Gerenciamento de metadados para registros - Parte 3: Autoavaliação**, publicado em 2011. Essa norma traz orientações sobre a criação, captura e controle de registros nos processos de autoavaliação de metadados. O documento define como vantagens do processo de autoavaliação a identificação do estado atual da captura e da gestão de metadados dentro ou entre organizações; a identificação de prioridades sobre o que trabalhar e quando; a identificação dos principais requisitos das duas partes anteriores da série metadados ISO 23081; a avaliação do progresso no desenvolvimento de um quadro de metadados para a implementação de sistemas e projetos específicos; além da prontidão na avaliação do sistema e do projeto (passar para a próxima fase de um sistema ou projeto) ao incluir a funcionalidade de metadados de registros em um sistema. Uma avaliação da prontidão dos metadados dos registros é fornecida para as principais etapas, desde o início do projeto até a fase de implementação/manutenção (ISO, 2011).

Dessa forma, o conjunto das três partes dessa série de normas se faz essencial para o desenvolvimento e para a implantação de projetos e aplicações envolvendo o registro de metadados nos mais diversos contextos. Tais padrões podem ser aplicados em qualquer ambiente de implementação e buscam garantir características de usabilidade, integridade, conveniência e confiabilidade no registro de metadados.

Assim, na linha de registro de metadados, e considerando o objeto informacional como um **módulo de conteúdos** Garshol (2004), Rosenfeld; Morville; Arango (2015), Victorino, Pinheiro, Santos, (2015), Alves (2010) apresentam metadados como os “atributos que representam o conteúdo na forma de **entidade** dentro de um sistema de informação”, ou seja, dados que representam e identificam um objeto do mundo real que seja de interesse do domínio. Para Carvalho (2013) esses metadados são representados por **termos** dentro da linguagem e se utilizam do alcance de uma correspondência entre objetos para classificá-los como idênticos ou análogos (ISO 25964-1, 2011a, p. 45, tradução livre). As equivalências acontecem nas seguintes situações: os termos são sinônimos; os termos são quase-sinônimos; o termo é considerado como desnecessariamente específico e é representado por outro termo com escopo mais amplo; e o termo é considerado como desnecessariamente específico e é representado por uma combinação de dois ou mais termo, conhecido como “equivalência composta”.

Na mesma intenção e forma de especificar o metadado para tratamento, uso e compartilhamento da informação, a definição de um objeto pode ser feita por uma associação de termos significativos que, uma vez relacionados, irão apresentar conteúdo suficiente para expressar a representação de algo que se necessita alcançar do mundo real na forma de um conceito estruturado do ponto de vista informacional e lógico (GORAYEB; DUQUE, 2024).

3 METODOLOGIA

A CI apresenta uma nova demanda de cientificidade, que é indício das mudanças na produção e na direção do conhecimento e levarão a um novo campo científico, além de colocar em xeque alguns pressupostos epistemológicos. Isso leva a resultados mais formalizados na produção de conhecimentos, considerando metodologias observacionais e quantitativas (GONZÁLES DE GÓMEZ, 2000).

Pode-se dizer que a metodologia da pesquisa em CI deve considerar seu caráter poliepistemológico, seja considerando o aspecto interdisciplinar ou multidisciplinar. Com efeito, A CI engloba fenômenos, processos e construções em diversos segmentos, dentre eles a linguagem (níveis sintáticos, semânticos e pragmáticos e suas plurais formas de expressão - sonoras, imagéticas, textuais, digitais/analógicas); “[...] os sistemas sociais de inscrição de significados (a imprensa e o papel, os meios audiovisuais, o *software* e o *hardware*, as infraestruturas das redes de comunicação remota)”; ainda, “[...] os sujeitos e as organizações que geram e usam informações em suas práticas e interações comunicativas” (GONZÁLES DE GÓMEZ, 2000, p. 4).

Cabe salientar, contudo, que a diversidade de condições epistemológicas não pode ser considerada indefinição metodológica eclética ou relativista. Não é como se a CI não tivesse bases próprias, com campos de atuação definidos, pelo contrário, é nas Ciências Sociais que a CI tem seu traço identificador. Dessa forma, os estudos metodológicos podem ser denominados como a “dupla hermenêutica” – política e epistemológica –, pois:

Seja qual for a construção do objeto da Ciência da Informação, ele deve dar conta do que as diferentes disciplinas, atividades e atores sociais constroem, significam e reconhecem como informação, numa época em que essa noção ocupa um lugar preferencial em todas as atividades sociais, dado que compõe tanto a definição contemporânea da riqueza quanto na formulação das evidências culturais (GONZÁLES DE GÓMEZ, 2000, p. 6).

Essa dupla hermenêutica política e epistemológica implica em que as pesquisas na área da CI não podem focar seus aportes teórico-metodológicos apenas em “[...] uma escola, uma teoria, uma técnica, uma temática, já que toda proposta se perfila num horizonte de demandas concorrenciais atualizadas permanentemente fora

do campo por macroprocessos econômicos e políticos” (GONZÁLES DE GÓMEZ, 2000, p. 6). De tal forma, quando se faz uma escolha metodológica, há um “[...] esforço de preenchimento do núcleo, como um espaço sempre em constituição que exige, caso a caso, uma nova justificativa”. (GONZÁLES DE GÓMEZ, 2000, p. 6).

Assim, nesse caminho interdisciplinar, este estudo está inserido na CI, portanto, seguirá as bases metodológicas das Ciências Sociais, com combinação da Ciência da Computação.

3.1 Contextualização e relação com os temas e conceitos da CI

Por questões que remetem à apresentação da metodologia, o cuidado, precisão e clareza dos temas faz-se necessárias algumas considerações a respeito do contexto desta pesquisa.

O primeiro tema a ser explicitado diz respeito à Nota Fiscal Eletrônica (NF-e) e Nota Fiscal de Consumidor Eletrônica (NFC-e), considerando que o resultado da pesquisa, a construção de um modelo de ontologia, partirá dessa temática em questão. Assim, busca-se elucidar o objeto para que se compreenda o ponto de partida da presente pesquisa.

A cobrança dos tributos sobre a comercialização de mercadorias e de serviços e a arrecadação é de competência dos Estados. Para a fiscalização sobre a arrecadação, o Estado do Amazonas atribui responsabilidade à SEFAZ/AM e mantém um sistema informatizado para emissão de NF-e, modelo 55, e a Nota Fiscal de Consumidor Eletrônica (NFC-e), modelo 65, instituindo em lei a obrigatoriedade do preenchimento do documento eletrônico e o pagamento do imposto, Convênios 57/95 e 58/05 e Legislação Superveniente de Ajuste SINIEF 07/05 associada ao Convênio ICMS 199/10.

A NF-e e NFC-e são documentos emitidos e armazenados eletronicamente que registram, para fins fiscais, as operações relativas aos produtos industrializados ou à circulação de mercadorias, transporte interestadual, intermunicipal e de comunicação, com validade jurídica garantida pela assinatura digital do remetente e pela recepção, pelo Fisco estadual.

Dentre outras exigências legais da NF-e e NFC-e estão: Convênio ICMS 199/10, Cláusula Terceira, Inciso I que diz: Os arquivos digitais da NF-e e NFC-e estão no padrão *Extended Markup Language* (XML) e possuem uma “chave de acesso” única de 44 dígitos para identificação, do tipo numérica composta por: modelo da nota, Unidade Federativa (UF), série, numeração sequencial única de 1 a 999.999.999 (que deve ser reinicializada quando atingido limite), CNPJ/CPF do remetente e tipo de emissão. Isto demonstra que os dados lógicos computacionais do documento fiscal estão em formato de metadados e podem ser reconhecidos e controlados pelos sistemas informatizados da SEFAZ/AM desde o emissor até destinatário final garantido o registro do fato gerador do ICMS e seu futuro recolhimento como receita do tesouro estadual.

Os campos de informação da NFC-e são logicamente estruturados de forma a conter as seguintes informações:

- Denominação: nota fiscal de venda ao consumidor;
- Número de ordem, série, subsérie e número da via;
- Data de emissão;
- Hora de emissão;
- Nome, endereço e os números de inscrição estadual e do CNPJ do estabelecimento emitente;
- CPF ou CNPJ do consumidor quando identificado;
- Destaque dos tributos estaduais; e
- Discriminação da mercadoria.

A discriminação das mercadorias comercializadas com NFC-e é feita por meio do seu correspondente código estabelecido na Nomenclatura Comum do Mercosul (NCM) (Frossard, 2011). O código do NCM foi criado em 1995 e deve ser utilizado por todos os países do Mercosul para classificar a mercadoria e definir a tributação de impostos. A tabela com NCM é mantida pela Secretaria de Fazenda da Receita Federal e segue orientação da metodologia internacional de classificação de

mercadorias conhecida como Sistema Harmonizado de Designação e de Codificação de Mercadorias (SH).

O código possui 8 dígitos com a finalidade de classificar e descrever as características de itens e subitens relacionados à mercadoria comercializada. Sua utilização busca padronizar as informações das mercadorias para fins de análise e acompanhamento tributário. Sendo assim, ao colocar o código NCM errado pode acontecer a cobrança indevida de impostos, perda de benefícios fiscais pela empresa e perda de arrecadação pelo Fisco.

Os campos de informação da NFC-e computacionalmente estruturados para a **identificação da venda e descrição da mercadoria**, ou seja, os metadados são: número do item, código do item, descrição do item, quantidade do item, unidade, valor unitário do item, valor total do item e valor total da operação ou prestação (somatória de todos os valores dos itens descritos no documento fiscal).

Especificamente sobre o campo que identifica a mercadoria, denominado como descrição do item (produto ou serviço), é um campo não estruturado do ponto de vista informacional, lógico e computacional. As normas de escrituração fiscal falam da “perfeita identificação” do produto, adequada ao ponto de identificar as características do produto vendido, “[...] A discriminação do item deve indicar precisamente o mesmo, sendo vedadas discriminações diferentes para o mesmo item ou discriminações genéricas” (SEFAZ/AM, 2011), “[...] discriminação das mercadorias ou dos serviços, tais como quantidade, marca, tipo, modelo, espécie, qualidade e demais elementos que permitam a sua perfeita identificação” (Amazonas, 2010), sem contudo especificar com clareza como escrever esses atributos, a ordem que eles se apresentam e quais são elementos e/ou propriedades que os descrevem.

A seguir, nas figuras abaixo, alguns exemplos de NFC-e emitidas em compras do produto “cerveja” para análise deste trabalho. As compras foram realizadas em vários estabelecimentos comerciais na Cidade de Manaus (AM) a partir da escolha de marcas de cervejas diversas apontadas no campo “DESC” de Descrição:

Figura 8 – NFC-e (exemplo 1)

Telefone: (92) 3652 2600 DATA: 04/06/2024 - 10:30
LOJA=0210 PDV=022 Seq=112392

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 269ML				
	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.an.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



Fonte: Acervo do autor (2023).

A figura 8 apresenta a compra de 3 produtos diferentes, o primeiro corresponde a “1 unidade da lata de cerveja Brahma com volume 350ML”, o segundo corresponde a “1 unidade da garrafa tipo Long Neck de cerveja Heineken com volume 330ML” e o terceiro produto corresponde a “1 unidade da lata de cerveja Heineken com volume 269ML”.

Como visto, os três produtos são iguais, porém de marcas, embalagens e volumes diferentes. O produto é “cerveja”, mas o emitente da NFC-e o descreve na forma de palavras abreviadas e diferentes: “CERV” e “CERVEJ”; na primeira linha além do nome do produto abreviado, a descrição apresenta o nome da marca “Brahma”, o volume “350ML”, mas não apresenta a embalagem onde esse volume está acondicionado, a embalagem está descrita como “La” (possível Lata) em outro campo diferente, campo “UN” de “Unidade”; na segunda linha, o produto tem o nome

da marca Heineken, possui volume “330ML”, também não apresenta a embalagem nem na descrição do produto nem na unidade do produto; essa mesma forma de descrição se repete na terceira linha apesar de, como visto anteriormente, a abreviação do produto ser diferente “CERVEJ”, sem contudo apresentar a embalagem garrafa.

A figura 9 apresenta outro exemplo de NFC-e e a dificuldade de reconhecer a descrição do produto cerveja:

Figura 9 – NFC-e (exemplo 2)

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

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	4,39	(1,60)	4,39
002	7897395040246	LONGNECK ITAIPAVA 25	1	un	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 8505 3000 0860 8010 3095 8603
 CONSUMIDOR NAO 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

Fonte: Acervo do autor (2023).

Essa segunda NFC-e corresponde à compra de 2 produtos, o primeiro corresponde a “1 unidade da garrafa tipo Long Neck de cerveja Itaipava 100% puro malte com volume 330ML”, e o segundo produto corresponde a “1 unidade da garrafa tipo Long Neck de cerveja Itaipava com volume de 250ML”.

Entretanto, a primeira linha do campo descrição apresenta o produto cerveja associado ao nome da marca abreviada “ITA 100 MALT” sem descrever volume ou embalagem do produto; já na segunda linha não há referência ao nome do produto cerveja, neste caso há descrição da embalagem “LONGNECK” associada ao nome da marca descrita de forma completa “ITAIPAVA” sem, contudo, descrever corretamente o volume (existe o numeral 25, mas não se pode afirmar que corresponde a 250ML).

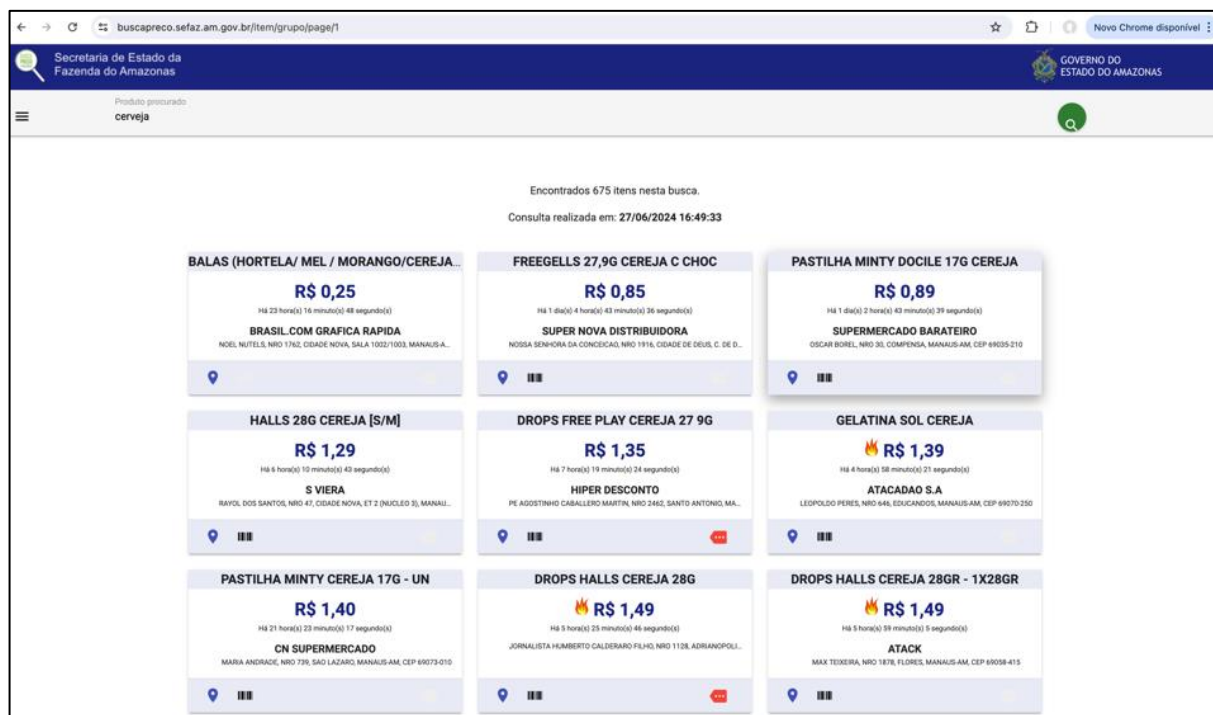
Existe à disposição da SEFAZ/AM uma grande quantidade de dados sobre produtos nas bases da NFC-e, porém, a falta de clareza na descrição, na organização da informação e dos meios de recuperação confiáveis que permitam à fiscalização a percepção do produto, do seu significado em determinado contexto e uso de diferentes formas faz com que a extração para as diversas finalidades da fiscalização seja basicamente manual quando necessário, somente sobre o que já é conhecido, limitando as ações de controle da arrecadação e o combate aos atos fraudulentos. Quando na necessidade a informação é extraída com ferramentas de *Big Data* e *Business Intelligence* produzindo conteúdos menos flexíveis quando comparado ao número de produtos nas bases da NFC-e, mais dependentes da equipe de tecnologia e que nem sempre reconhecem efetivamente o produto que se deseja alcançar informação, ou que nem sempre recuperam toda a informação necessária deste produto.

Um exemplo da dificuldade de recuperar informação está disponível ao público, a SEFAZ/AM disponibiliza em sítio eletrônico⁶ a consulta para buscar valores de produtos, a finalidade é orientar o consumidor do melhor valor e local da venda do produto buscado. Os dados selecionados são os mesmos da base de NFC-e e os mesmos entregues para esta pesquisa e são tratados do ponto de vista do pré-processamento lógico computacional.

O resultado da consulta para o nome “cerveja” pode ser visualizado na figura abaixo:

6 Disponível em: www.buscapreco.sefaz.am.gov.br

Figura 10 – Tela sítio eletrônico da SEFAZ/AM - consulta para o nome “cerveja”.



Fonte: SEFAZ/AM (2023).

Os nove primeiros resultados apresentados na figura acima não correspondem ao produto pesquisado, apresentando produtos variados como “Balas (hortelã/mel/morango/cereja...”, “Halls 28g cereja (S/M)” ou “gelatina sol cereja”. É possível perceber que, para recuperação da informação, são utilizadas palavras indexadas, como em outras ferramentas de busca, mas com baixo grau de precisão. Disso resulta uma representação inadequada do objeto sem uso para a necessidade da busca. Não é somente a baixa precisão da ferramenta utilizada para a recuperação da informação, mas também as diversas formas de descrever o produto interferem na expectativa de obter uma informação de qualidade no processo de recuperação.

As diversas formas de “comunicar o produto”, os valores e características, informações do emitente e destinatário etc. são consequências diretas da necessidade de registrar o fato gerador de impostos no ato da compra/venda (documento fiscal das figuras 8 e 9) e de recuperar a informação (exemplo da figura 10) independente da finalidade da SEFAZ/AM, se para apresentar ao consumidor na forma de consulta ao menor preço, localização do produto ou para fins de

acompanhamento tributário, comportamento dos preços para composição de preço médio para substituição tributária ou para auditoria em uma fiscalização.

Neste talante, é importante, também, estabelecer o escopo de atuação da SEFAZ/AM para que o objetivo deste trabalho se torne mais esclarecido. A SEFAZ é a sigla para sintetizar a Secretaria da Fazenda, que é um órgão público governamental, atuante nos âmbitos federal, estadual e municipal. É de sua responsabilidade a administração das finanças públicas, incluindo a arrecadação de impostos, a fiscalização tributária, a gestão orçamentária, o controle de gastos públicos e outras atribuições que mantenham relação com a área financeira (MARTINS, 2003).

Pode-se dizer que tudo que esteja ligado às leis e regras fiscais deve ser gerenciado e administrado pela SEFAZ de cada esfera e do ponto de vista da federação, os estados, as Secretarias de Fazendas mantêm controle e fiscalização adequada para cada caso. Quanto ao seu objetivo, considera centralizar a gestão financeira e fiscal e promover a arrecadação de impostos. De tal forma, é por meio de sua atuação que o Estado se mantém funcionando, com desenvolvimento econômico do país (MARTINS, 2003).

Quanto às funções exercidas pela SEFAZ, há um número expressivo de funções, dentre elas, Martins (2003) destaca:

- Arrecadação de impostos: responsável por receber os impostos, taxas e contribuições pagas pelos cidadãos e empresas. É o órgão que verifica se as obrigações fiscais estão sendo cumpridas corretamente, emite guias de pagamento e realiza a cobrança dos valores devidos.
- Controle fiscal: fiscaliza e monitora as atividades econômicas para garantir que as empresas e contribuintes estejam cumprindo suas obrigações tributárias e seguindo as leis fiscais. É por meio desse controle que se combate a sonegação fiscal e outras irregularidades, de modo que todos contribuam de forma justa.
- Gestão financeira e orçamentária: encarregada de planejar, executar e controlar o orçamento público. Por isso, acompanha as receitas (dinheiro

arrecadado) e as despesas (gastos públicos), garantindo equilíbrio das contas do governo, de modo a utilizar os recursos de maneira eficiente.

- Políticas econômicas e fiscais: contribui para a definição e implementação de políticas que impactam a economia e os impostos. Por tanto, realiza estudos e análises com o fito de entender os efeitos de possíveis mudanças na legislação tributária. Também propõe medidas para incentivar o desenvolvimento econômico, atrair investimentos e promover a geração de empregos.
- Cadastro e registro de empresas: Em alguns casos, é responsável por cadastrar e registrar as empresas, emitindo inscrições necessárias para o funcionamento das atividades econômicas e o cumprimento das obrigações fiscais.
- Concessão de incentivos fiscais: pode conceder benefícios fiscais, como isenções e reduções de impostos, além de oferecer incentivos para determinadas atividades econômicas. Por meio dessas medidas, busca estimular setores específicos, promover o desenvolvimento regional e atrair investimentos.
- Análise e controle do comércio exterior: Em alguns estados, é de sua responsabilidade controlar e fiscalizar as operações de comércio exterior. Como exemplo desse ato, tem-se: verificar o cumprimento das obrigações aduaneiras, emitir documentos fiscais para importação e exportação, além de aplicar normas relacionadas ao comércio internacional.

A SEFAZ é parte importante da economia do país, configurando-se como um órgão responsável por atividades inseridas nas organizações brasileiras, prestando serviços como:

- Registro do ICMS;
- Consulta de cadastro do ICMS e ITCMD;
- Emissão de guias de pagamento;
- Verificação do protocolo integrado;
- Verificação da nota fiscal do consumidor, a NFC-e

- Verificação do CT-e (Conhecimento de transporte eletrônico);
- Verificação da NF-e (Nota Fiscal Eletrônica);
- Consultas relacionadas à receita do Estado;
- Impressão das guias;
- Emissão da Escrituração Fiscal Digital ou SPED (Sistema Público de Escrituração Digital) fiscal;
- Consulta da Declaração de Importação (DI);
- Acesso às certidões negativas, débitos automáticos ou certidão de transferências voluntárias;
- Realização do pagamento do IPVA (Martins, 2003).

Em particular, a Secretaria de Fazenda do Estado do Amazonas (SEFAZ/AM) possui uma organização interna que contempla uma área específica para as atividades de acompanhamento fiscal e tributário denominada de Secretaria Executiva da Receita (SER) que tem por finalidade a supervisão da execução das atividades do Centro de Estudos Econômico-Tributários – CEET, da Central de Atendimento ao Contribuinte – CAC, do Núcleo de Educação Fiscal – NEF, da Unidade de Inteligência Fiscal – UNIF, e dos Departamentos de Análise e Revisão da Ação Fiscal – DEARF, de Arrecadação – DEARC, de Controle de Entrada de Mercadorias – DECEM, de Fiscalização – DEFIS, de Informações Econômico-Fiscais – DEINF e de Tributação – DETRI. Dentre outros objetivos, compete a estas unidades, conforme Decreto Estadual nº 44.753 de 27/10/2021: “[...] I - **promover** a orientação normativa e a **coordenação dos sistemas de arrecadação**, cadastro, desembaraço de documentos, **fiscalização** e tributação”. Desta forma, é sob a responsabilidade da SER como órgão gestor da arrecadação e fiscalização que se promovem as **ações de fiscalização, inteligência fiscal e análise de dados fiscais** baseadas nas informações lógica e computacionalmente formatadas nos sistemas informatizados disponíveis que reconhecem a movimentação de mercadorias e serviços por meio dos documentos fiscais eletrônicos e que permitem acompanhar o fato gerador e recolhimento do tributo para o tesouro estadual.

A **auditoria** é outra temática que merece ser destacada para complementar a inteligência fiscal, haja vista que a pesquisa tem como finalidade melhorar este processo, tornando-o mais efetivo e produtivo. Assim, serão destacados alguns elementos relevantes:

- **Conceito:** no âmbito da receita estadual, a auditoria governamental, desempenhada por profissionais de auditoria, ocorre com o levantamento, análise imparcial e a independente avaliação da informação, devidamente consubstanciada em evidências e seguindo critérios previamente estabelecidos, como legalidade, legitimidade, economicidade, eficácia, ética e transparência (NAG-1102.1, 2011).
- **Competência de auditoria:** no âmbito da receita estadual alcança as áreas de arrecadação, fiscalização e tributação para o tesouro estadual (FROSSARD, 2011), visando:

Desenvolver estudos e realizar projeções sobre o comportamento da arrecadação de receitas de competência do Estado, adotando medidas que propiciem o seu incremento;

Planejar e coordenar as atividades de natureza econômico fiscal; coordenar o planejamento estratégico, estudos, pesquisas e projetos, bem como estabelecer diretrizes gerais e específicas de sua área, objetivando o aprimoramento da gestão da Receita Estadual e a política fiscal.

Exercer a fiscalização do trânsito e da circulação de mercadorias, bens e serviços, bem como a de outros tributos que não os instituídos pelo Estado, cuja competência lhe seja delegada por ente tributário, mediante convênio;

Exercer as atividades de auditoria digital e propor, implementar, controlar e avaliar instrumentos e sistemas de informática a serem utilizados no planejamento, desenvolvimento, execução, acompanhamento, controle e avaliação dos programas, projetos e ações de controle fiscal (FROSSARD, 2011).

- **Auditoria e o controle da informação armazenada nas bases de dados:** o departamento de fiscalização da SEFAZ possui competência para fazer a gestão de documentos eletrônicos recebidos compartilhando informações pelos Estados Federados por meio do Sistema Integrado de Informações sobre Operações Interestaduais com Mercadorias e Serviços – SINTEGRA, além de:

Realizar o tratamento das informações contidas nos arquivos apresentados pelos contribuintes comparando-os com arquivos de

sistemas internos e externos da SEFAZ/AM para subsidiar planejamento e formatação das ações fiscais;

Executar as atividades de administração e controle relacionados ao uso, alteração de uso ou desistência de uso de Equipamentos Emissores de Cupom Fiscal – ECF e do Sistema Eletrônico de Processamento de Dados PED, para emissão de documentos fiscais e escrituração de livros fiscais; Autorizar a confecção de lacre de segurança do ECF, mediante solicitação das empresas credenciadas;

Executar ações de orientação e controle junto aos contribuintes emissores da Nota Fiscal de Consumidor Eletrônica – NFC-e e demais documentos fiscais eletrônicos;

Analisar as informações contidas na base de dados da NFC-e e elaborar levantamentos sobre o movimento econômico dos contribuintes emissores para subsidiar às demais gerências do DEFIS no planejamento e formatação das ações fiscais, Decreto N.º 44.753, (2021).

- Auditoria e a inteligência fiscal: possui a finalidade de produzir conhecimento por meio de:

Busca e análise de fatos, indícios, denúncias e informações; Apurações e levantamentos de interesse da ação fiscal; Monitoramento eletrônico de contribuintes; Cruzamento de dados oriundos dos sistemas internos da SEFAZ e de fontes externas; Identificação e mapeamento de focos e formas de fraudes fiscais, Decreto N.º 44.753, (2021).

- Auditoria e aspectos técnicos das ferramentas e dos sistemas de informação automatizados: existem atualmente ferramentas de consultas de *Business Intelligence* que produzem consultas específicas baseadas em mineração de dados e textos sem contudo, contar com mecanismos de interpretação semântica e que facilitem a recuperação da informação na forma mais precisa; outras ferramentas de linguagem de consulta estruturada ou *Structured Query Language* (SQL) para consultas diretas as bases de dados da NF-e e NFC-e realizadas principalmente por técnicos da área de tecnologia da informação e comunicação (TIC) da SEFAZ/AM.

A partir dos conceitos, processos e ferramentas de auditoria e as necessidades de evolução apresentados até aqui e, considerando a massa de dados originada na emissão do documento fiscal NFC-e é possível estimar que os fatores relevantes para esta pesquisa são: (a) na **representação adequada do objeto** de interesse e do seu uso tanto para o consumidor quando para fiscalização que na concepção estão associados ao projeto e à gestão de informação; (b) **necessidade de organizar a**

informação, escolha e seleção do meio associados à Arquitetura da Informação; (c) **necessidade de recuperar a informação** para apresentar informação em um processo de comunicação que pode ser entendido como a interface entre o autor/produtor de informação e o leitor/usuário de informação (DUQUE, 2005).

Aspectos como compreensão, uso e contexto ampliaram a capacidade de RI e modificaram o processo de comunicação da informação. Continua o autor, “[...] cria-se novas metodologias e técnicas de organização e recuperação da informação aproxima o usuário àquela informação desejada e necessária” (DUQUE, 2005, p. 13).

Na troca de informação e de conhecimento, a Arquitetura da Informação concilia os objetivos do negócio e da necessidade de informação do público, como disciplina científica a Arquitetura da Informação propõe um *framework* para capturar, organizar e representar a informação armazenando-a para uso, projetando a informação acumulada e aumentando a capacidade de gestão diante da velocidade e do volume de dados gerados por diferentes fontes.

Aliada à expansão das tecnologias para soluções computacionais, a Arquitetura da Informação inclui na proposta o uso de programas codificados passo a passo, associados ao uso de ferramentas mais sofisticadas, autônomas e mais independentes da intervenção humana para aquisição do conhecimento, baseadas na AM, PLN e SOCs que permitem avanços na disponibilidade de infraestrutura de coleta, armazenamento, processamento e distribuição dos dados (FACELI *et al.*, 2021). Assim, se tornam mais acessíveis e permitem mais facilmente representar o conhecimento contido neles (BRACHMAN; LEVESQUE, 1982).

Quanto ao processo de criação de ontologias, o objetivo é garantir a validade da informação a partir da forma como elas descrevem a realidade por meio da modelagem de entidades e de processos de interesse, e de como formam um repositório de conhecimento especializado que se relacionam por regras para interagir com usuários na forma de consultas e recomendações justificadas (PROMKOT; ARCH-INT; ARCH-INT, 2019) e (SOBHANI; IZQUIERDO; PIATRIK, 2017).

Na prática, modelos são construídos para desenhar objetos de interesse sobre o mundo e demonstrar como é a influência de uns em relação a outros, bem como apresentar seus relacionamentos com os atores que participam da mesma realidade.

O modelo construído na Engenharia de Ontologias possui padrão formal para organizar e representar conceitos e relações em linguagem de representação lógica para ser manipulada por mecanismos de inferência (GRUBER, 1993), enquanto artefato da Arquitetura da Informação (BEIRA *et al.*, 2017). Semelhantes ao processo de construção da Engenharia de Sistemas (SILVA, 2008), seus resultados influenciam a produção e uso da informação de qualidade, disponível e compartilhada adequadamente para fins de fiscalização.

Diante do exposto, o desafio é, portanto:

1. Identificar o conteúdo no campo não estruturado do documento fiscal eletrônico buscando nos dados a descrição do nome e das características do produto, que, por vezes, confundem marca, tipo e aspectos como tamanho, cor, textura, embalagem e volumetria dificultando:
 - a. A rastreabilidade do produto na cadeia de venda;
 - b. O controle da arrecadação quando há divergência entre código do produto e descrição dele;
 - c. A definição da alíquota do Imposto sobre a Circulação de Mercadorias e Serviços (ICMS) para substituição tributária; o cálculo do Preço Médio Ponderado ao Consumidor Final (PMPF); e estimativas de preço em licitações governamentais entre outras finalidades que dependem exclusivamente de pesquisa sobre os valores praticados no mercado, em determinado momento.
2. Superar o volume com um grande número de documentos e imensa massa de dados a serem analisados;
3. Superar a inconsistência e a falta de metadados na descrição do item;
4. Superar a heterogeneidade, complexidade e a existência de polissemia e falta de padronização da terminologia para a categorização na descrição dos produtos (nomes e atributos) com abreviações e descrições ambíguas, o que demanda esforço para compreensão e controle da informação; e
5. Comunicar informação de qualidade para fiscalização do produto “cerveja de malte”.

Da exposição do contexto acima, o problema da pesquisa, reforça o escopo deste trabalho considerando a heterogeneidade, ambiguidade e má qualidade dos dados que estão dispostos no campo descrição do item produto e que não permitem facilmente a percepção do significado do produto, a classificação e a organização da informação visando à recuperação e à apresentação da informação para a fiscalização em Notas Fiscais de Consumidor Eletrônicas.

3.2 Contextualização da pesquisa científica: configurações metodológicas

Os termos (conceitos e categorias) desta pesquisa em Ciência da Informação (CI) tem enfoque em: Sistema de Organização do Conhecimento (SOC); Arquitetura da Informação (AI); e Ontologia. Na área da Ciência da Computação (CC), os termos são: Processamento de Linguagem Natural (PLN); Aprendizado de Máquina (AM); Mineração de Dados (MD) e Mineração de Texto (MT); e Metadados.

O Estruturalismo é o **método de procedimento** desta pesquisa (LAKATOS; MARCONI, 2017, p. 118) posto que pretende:

- Realizar uma decomposição analítica, “[...] já que, para entender um fenômeno, é mister desmontá-lo em suas partes; e isto é precisamente análise”;
- A decomposição analítica vai apresentar “[...] que a complexidade do fenômeno é uma percepção superficial”;
- Considera que “explicar é escavar a subjacência, porquanto a superfície varia, não o fundo, que invaria”; e
- Considera que “[...] o fenômeno é simplificável em modelos estruturais, revelando a ordem interna subjacente, ao contrário da visão de superfície” (Demo, 1985, p. 106).

Para Lakatos e Marconi (2017, p. 118), o Estruturalismo parte de um fenômeno concreto, chega ao nível do abstrato, viabilizado pela “[...] constituição de um modelo que represente o objeto de estudo”, para, então, retornar ao concreto “[...] como uma realidade estruturada e relacionada com a experiência do sujeito social”. Assim sendo, o método estruturalista faz um intercâmbio do concreto para o abstrato e vice-versa e,

na segunda etapa, apresenta um modelo em que se pode analisar a realidade concreta dos diversos fenômenos (LAKATOS; MARCONI, 2017, p. 118). Em síntese:

[...] toda análise deve levar a um modelo, cuja característica é a possibilidade de explicar a totalidade do fenômeno, assim como a sua variabilidade aparente, porque, por intermédio da simplificação (representação simplificada), o modelo atinge o nível inconsciente e invariante: resume o fenômeno e propicia sua inteligibilidade. Utilizando-se o método estruturalista, não se analisam os elementos em si, mas as relações que entre eles ocorrem, pois somente estas são constantes, ao passo que os elementos podem variar. Dessa forma, não existem fatos isolados passíveis de conhecimento, pois a verdadeira significação resulta da relação entre eles.

No conseqüente deste caminho metodológico, problema e hipótese serão detalhados, considerando o que ensina Lakatos e Marconi (2017, p. 136-137). Quanto ao problema “[...] consiste em um enunciado claro, compreensível e operacional, cujo melhor modo de solução ou é uma pesquisa, ou pode ser resolvido por meio de processos científicos”; já hipótese é “[...] uma suposta, provável e provisória resposta a um problema”, cuja adequação (comprovação = sustentabilidade ou validade) será verificada através da pesquisa.

Do ponto de vista da sua **natureza**, a proposta desta pesquisa se assenta como uma **pesquisa aplicada** (GIL, 2002, p. 27) cujos resultados possuem fins práticos revertidos como novos conhecimentos. Enseja “[...] gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos” (PRODANOV; FREITAS, 2013, p. 51), no caso, servirá como um modelo para auxiliar as auditorias SEFAZ/AM.

Quanto à **finalidade** da pesquisa, entendendo que há um objetivo final de toda e qualquer pesquisa, este estudo se enquadra na **pesquisa descritiva**: quando o pesquisador registra e descreve os fatos observados sem interferir neles. Nesse tipo de pesquisa, o objetivo é descrever as características de determinada população ou fenômeno ou o estabelecer quais as relações entre variáveis. Envolve o uso de técnicas padronizadas de coleta de dados: questionário e observação sistemática. Assume, em geral, a forma de Levantamento (PRODANOV; FREITAS, 2013, p. 51). No caso desta pesquisa, o levantamento de dados não será feito por questionário ou observação, mas se dá a partir de mineração de texto, usando elementos estruturais da AI.

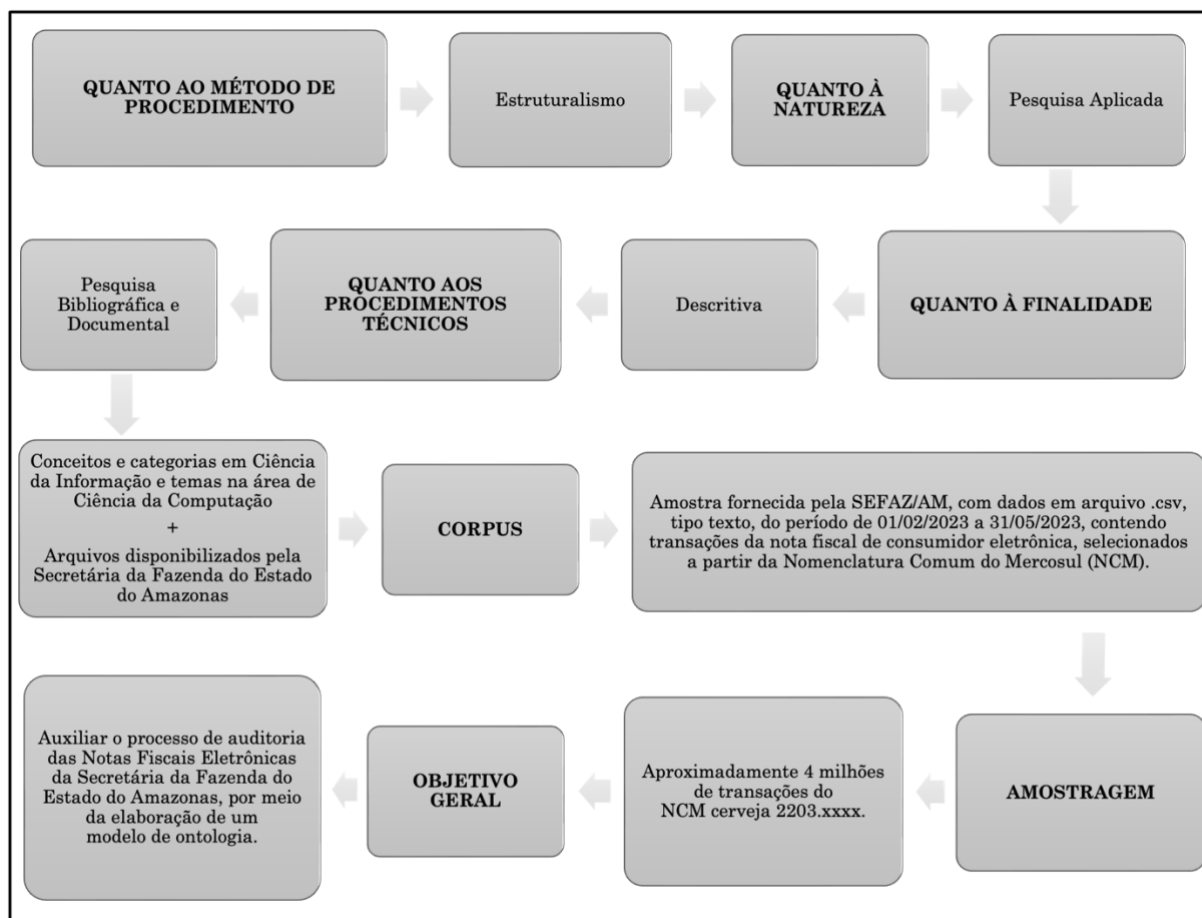
Quanto aos procedimentos técnicos, ou seja, a maneira pela qual os dados serão levantados, o que responde à questão: como os dados necessários serão obtidos? Nesta pesquisa, os procedimentos são **pesquisa bibliográfica** por fazer um levantamento teórico de estudos nas áreas desta pesquisa já realizados anteriormente e por sistematizar conceitos e teorias de autores que debatem as temáticas deste estudo; **e documental**, pois os dados não receberam ainda um tratamento analítico e provêm de **arquivos**, neste caso as notas fiscais eletrônicas são documentos disponibilizados pela SEFAZ/AM – conforme termo de arquivo de domínio público da esfera estadual.

É importante considerar que a pesquisa documental, embora seja parecida com a pesquisa bibliográfica, não podem ser confundidas, considerando a natureza das fontes de ambas as pesquisas. No caso, a pesquisa bibliográfica se utiliza fundamentalmente das contribuições de vários autores sobre determinado assunto, já a pesquisa documental tem sua natureza em documentos de arquivos públicos ou privados (PRODANOV; FREITAS, 2013, p. 55).

Quanto à abordagem do problema, a pesquisa se apresenta como **pesquisa quantitativa**, pois “[...] considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las”. A pesquisa quantitativa faz “[...] uso de recursos e de técnicas estatísticas (percentagem, média, moda, mediana, desvio-padrão, coeficiente de correlação, análise de regressão etc.)” (PRODANOV; FREITAS, 2013, p.69).

O **Corpus** da pesquisa em questão é uma amostra fornecida pela SEFAZ/AM, com dados em arquivo .csv, tipo texto, do período de 01/02/2023 a 31/05/2023, contendo transações da nota fiscal de consumidor eletrônica, selecionados a partir do NCM. A **amostragem** não considerará a nota fiscal – sequer sua identificação, considerando o caráter sigiloso desse documento –, será constituída por mais de 4 milhões de transações do NCM cerveja 2203 comercializadas no período acima. Esclarece-se que houve autorização da SEFAZ/AM para que os dados fossem disponibilizados (anexos B e C). A título de síntese, a figura 11 apresenta a trilha metodológica desta pesquisa:

Figura 11 – Percurso metodológico da pesquisa



Fonte: Dados da Pesquisa (2024).

3.3 Configurações do procedimento metodológico da Ontologia

É importante situar a ontologia e diferenciá-la de outros tipos de SOC. Primeiro, há de se considerar que diversos tipos de estruturas são utilizados na organização da informação. Seguindo, as estruturas organizadas a partir da utilização de termos são denominados de arquivos de autoridade, de glossários e de dicionários. Além disso, as estruturas que se organizam com base na classificação e na criação de categorias são estabelecidas como cabeçalhos de assunto e esquemas de classificação (ou taxonomias). Por fim, as estruturas que se organizam partindo de conceitos e de seus relacionamentos são designadas de ontologias, tesouros e redes semânticas (ALMEIDA; BAX, 2003, p. 07).

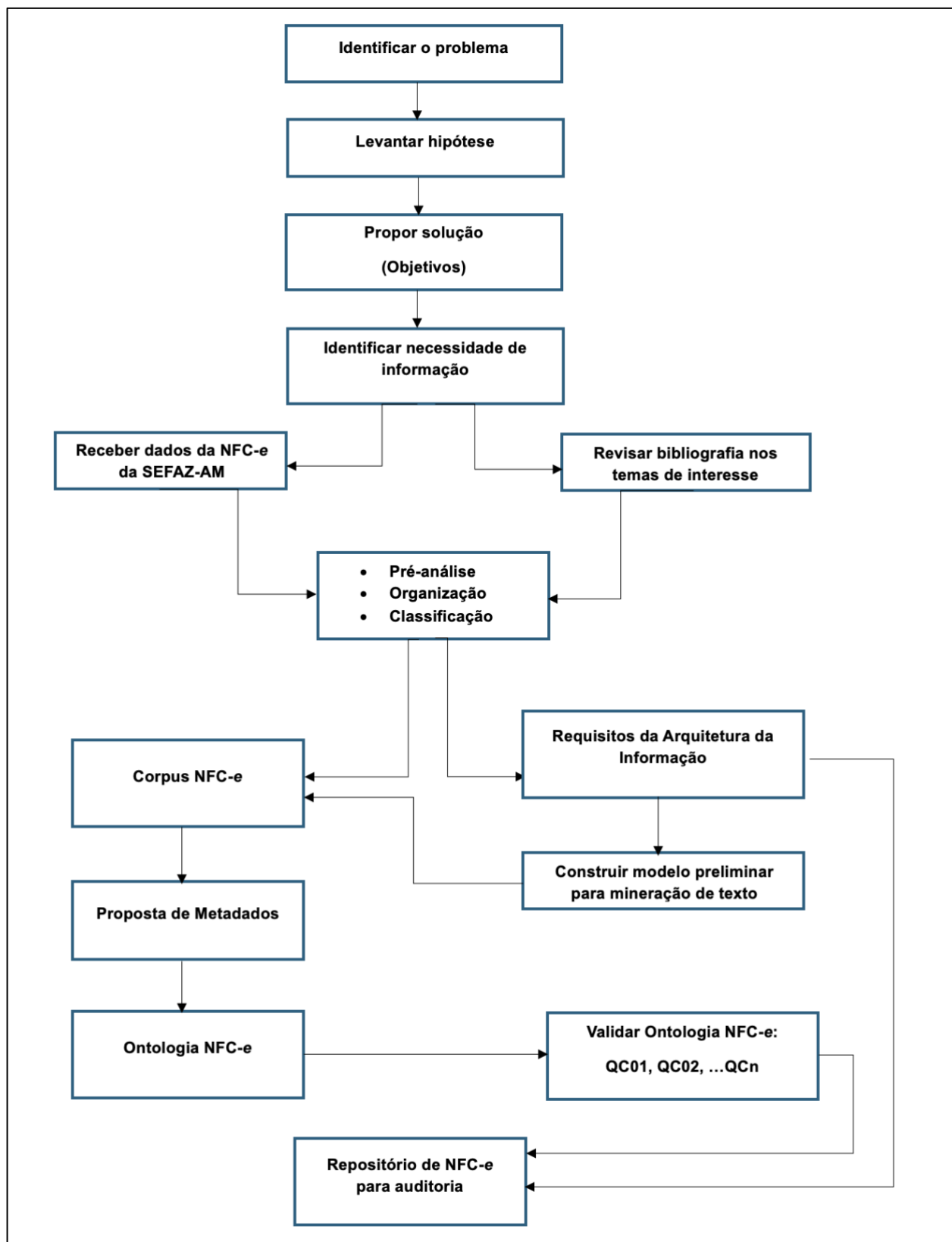
Assim, parte-se dos conceitos e daquilo que se relaciona a ele para se formular a ontologia, sendo criada por especialistas para definir as regras que regulam a combinação entre termos e relações em um domínio do conhecimento. Para fazer uso da Ontologia, os usuários usam conceitos definidos na ontologia. Desse modo, “o que se busca, em última instância, são melhorias nos processos de recuperação da informação” (ALMEIDA; BAX, 2003, p. 7).

Neste caso, a ontologia deverá favorecer o trabalho de auditoria da SEFAZ/AM e seguirá as etapas abaixo descritas:

- Realizar o pré-processamento dos dados recebidos da SEFAZ/AM e preparação do Corpus;
- Realizar as pesquisas bibliográficas necessárias ao desenvolvimento dos requisitos da Arquitetura da Informação, proposta de metadados e do modelo de ontologia;
- Construir proposta de modelo preliminar da Arquitetura da Informação para aplicar com mineração de texto e alcançar a seleção de termos de interesse para ontologia;
- Proposta de metadados para construção da ontologia em substituição aos especialistas;
- Construção da ontologia a partir do modelo preliminar e do Corpus.
- Validação por meio das Questões de Competência para alcançar informação útil.
- Disposição da informação no repositório de NFC-e para auditoria.

Para melhor visualização das etapas, a figura 12 demonstra os passos a serem seguidos:

Figura 12 – Procedimentos metodológico de construção do modelo de ontologia.



Fonte: Dados da pesquisa (2024).

De modo resumido o quadro 5 sintetiza os processos e os resultados para que se efetive o modelo da ontologia:

Quadro 5 – Processos e resultados para efetivação da ontologia

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: <ul style="list-style-type: none"> • 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: <ul style="list-style-type: none"> • 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 fiscalização.
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 - <i>Data Property Assertions</i> de instâncias das classes para validação dos dados.
Incorporar requisitos da Arquitetura da Informação	Modelo de repositório para auxiliar a auditoria.

Fonte: Dados da pesquisa (2024).

Os principais benefícios relacionados à utilização de ontologias voltam-se para três aspectos, que se conformam ao desenvolvimento desta pesquisa:

- **Comunicação** - permitem a comunicação entre pessoas sobre determinado conhecimento, de modo a favorecer raciocínio e entendimento sobre um domínio. Por meio dessa relação, pode-se chegar à obtenção de consenso, em especial sobre termos técnicos, entre comunidades profissionais, de pesquisa etc.
- **Formalização** – relaciona-se à especificação da ontologia, o que viabiliza a eliminação de contradições e de inconsistências na representação de conhecimento, levando-a, também, a não ser ambígua. Outrossim, há possibilidade dessa especificação ser testada, validada e verificada.
- **Representação de Conhecimento e Reutilização** - formam um vocabulário de consenso, com potencial para representar conhecimento de um domínio em seu nível mais alto de abstração, possuindo, desta forma, potencial de reutilização (MORAIS; AMBRÓSIO, 2007, p. 4-5).
- Tomando como referência a proposta de metodologia para criação de ontologia

Os autores Almeida e Bax (2003, p. 10) apresentam trabalhos no uso da ontologia destacando projetos, metodologia, ferramentas, linguagem e métodos de avaliação. Os autores afirmam que nestes termos a ontologias regulam a combinação de termos e relações em um domínio de conhecimento melhorando o processo de recuperação da informação e classificam os tipos de ontologia pela função, grau de formalismo, aplicação, estrutura e conteúdo da ontologia proposta para este trabalho, assim descrito:

- **Quanto à função** - Ontologia de Domínio – já que descreve conceitos e vocabulários relacionados a domínios particulares. Construída para representar um “micro-mundo” (MORAIS; AMBRÓSIO, 2007, p. 6).
- **Quanto ao grau de formalismo** - Ontologia Semiformal – por ser expressa em uma linguagem artificial definida formalmente.
- **Quanto à aplicação** - Ontologia como Especificação - é criada para um domínio, que será usada para documentação e manutenção no desenvolvimento do software

- **Quanto à estrutura** - Ontologia de Domínio - descreve o vocabulário relacionado a um domínio.
- **Quanto ao conteúdo** - abrange a ontologia de informação, pois especifica a estrutura de registro do banco de dados; a ontologia de modelagem do conhecimento, porque especifica conceitualizações do conhecimento, tem uma estrutura interna semanticamente rica e são refinadas para uso no domínio do conhecimento que descrevem; a ontologia de aplicação, considerando que contém as definições necessárias para modelar o conhecimento em uma aplicação; por fim, a ontologia de domínio, que expressa conceitualização específica para um determinado domínio do conhecimento (ALMEIDA; BAX, 2003, p. 10).

De modo sintético, a figura 13 apresenta a abordagem e a classificação da ontologia proposta nesta tese:

Figura 13 – Abordagem e classificação da ontologia



Fonte: Dados da pesquisa (2024).

Respondidos alguns questionamentos: quais procedimentos? Quais ferramentas? quais caminhos seguir? Ou seja, estabelecidas as bases metodológicas, pode-se “fazer ciência”, cuidar da parte prática, posto que é “um erro superestimar a metodologia, no sentido de cuidar mais dela do que de fazer [ciência]” (Demo, 1985, p.19). É necessário chegar ao ponto final da pesquisa (mesmo que para outros pesquisadores seja um ponto de interrogação e dela, quiçá, provenham outros estudos). “A pergunta pelos meios de como chegar lá é essencial, mas é especificamente instrumental” (Demo, 1985, p.19)., pois quando o caminho já foi traçado, cabe ao cientista aplicar a metodologia e obter resultados, que neste trabalho é a produção de modelo de ontologia.

4 RESULTADOS ESPERADOS

Do ponto de vista científico, a formulação de uma proposta metodológica utilizando a Arquitetura Da Informação na mineração de texto de termos relevantes para construção de ontologia como artefato para repositórios de informações de qualidade propicia:

1. Melhor da consistência do dado, tanto na descrição quando análise semântica, garantindo melhor compreensão e interpretação das informações;
2. A interoperabilidade dos dados pelos padrões criados a partir de metadados e relações semânticas;

Do ponto de vista fiscal, a importância e o esforço de superar a velocidade e volume com que os dados se acumulam nas bases de dados da NF-e e NFC-e da SEFAZ/AM bem como o problema da ambiguidade na descrição dos produtos para:

1. Aumentar o alcance dos produtos fiscalizados de forma eficiente possivelmente ampliando as áreas de maior impacto justificado pela precisão da comunicação da sentença completa que descreve o produto,
2. Aumentar da base de consulta e formulação do Preço Médio Ponderado ao Consumidor Final;
3. Aumentar o número de contribuintes sujeitos envolvidos nos processos de fiscalização reduzindo a desigualdade no sistema tributário;
4. Incentivar uma política de padrão mais adequado a descrição do produto de acordo com a produção de descrições completas mais precisas incluindo metadados para organização e representação da informação
5. Melhorar o uso de recursos públicos destinados à fiscalização, implementando técnicas assertivas para otimizar o uso desses recursos e maximizar os resultados, fortalecer a confiança
6. Fortalecimento da Confiança Pública com transparência adequada à sociedade que busca informações de um produto específico;

Nessa abordagem, é possível revelar que o objetivo geral, auxiliar o processo de auditoria das notas fiscais da SEFAZ/AM, por meio da elaboração de um modelo de ontologia já se encontra em processo. Isso é evidente pelos testes já realizados que se configuraram em três artigos escritos. Além disso, os passos em direção à elaboração da ontologia também já estão iniciados.

Dos objetivos específicos, estes a seguir já foram consolidados: (a) Descrever a relevância dos SOC's, especificamente, da ontologia para os processos de Organização e Recuperação da Informação; (b) Definir as principais informações extraídas quando aplicada a mineração de dados nas Notas Fiscais da SEFAZ/AM; (c) Identificar possíveis requisitos de uma Arquitetura da Informação para a mineração de texto em Notas Fiscais de Consumidor Eletrônicas. Resta concluir o último objetivo específico, qual seja, (d) Apresentar um modelo de ontologia para orientar um padrão para as auditorias da Notas Fiscais da SEFAZ/AM.

5 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

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						

REFERÊNCIAS

- ABBAGNANO, N. Dicionário de Filosofia. 5. ed. [S.l.]: Martins Fontes, 2015.
Disponível em: <https://marcosfabionuva.com/wp-content/uploads/2012/04/nicola-abbagnano-dicionario-de-filosofia.pdf>. Acesso em: 13 jan. 2024.
- AFONSO, A. R.; DUQUE, C. G. Mineração de textos aplicada a postagens do Twitter sobre Coronavírus: uma análise na linha do tempo. **Liinc em Revista**, Rio de Janeiro, v. 16, n. 2, e5325, dez. 2020. DOI 10.18617/liinc.v16i2.5325. Acesso em: 2 ago. 2023.
- AGGARWAL, C. C. **Data Classification**: Algorithms and Applications. CRC Press, New York, 2014.
- ALBUQUERQUE, H. O.; SOUZA, E. P. R.; GOMES, C.; PINTO, M. H. de C.; FILHO SILVA, R. P.; COSTA, R.; LOPES, V. T. de M.; SILVA, N. F. F. da; CARVALHO, A. C. P. L. F. de; OLIVEIRA, A. L. I. de. Named Entity Recognition: a survey for the Portuguese language. **Procesamiento del Lenguaje Natural**, [S. l.], n. 70, pp. 171-185, mar., 2023. Disponível em: [https://repositorio.usp.br/result.php?filter\[\]=author.person.name:%22Filho,%20Ricardo%20P.%20S%22](https://repositorio.usp.br/result.php?filter[]=author.person.name:%22Filho,%20Ricardo%20P.%20S%22). Acesso em: 13 jan. 2024.
- ALMEIDA, M. B.; BAX, M. P. **Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção**. Ci. Inf., Brasília, v. 32, n. 3, p. 7-20, set./dez. 2003. Disponível em: <https://www.scielo.br/j/ci/a/LR68syZsPSSmwvPHrNXmC8N/abstract/?lang=pt>. Acesso em: 22 jan. 2024.
- ALMEIDA, M. B. **Ontologia em Ciência da Informação**: teoria e método. Curitiba: CRV, 2020. (Coleção Representação do Conhecimento em Ciência da Informação – volume 1). Disponível em: <https://ncor-brasil.org/wp-content/uploads/2023/04/Vol-1-OCI.pdf>. Acesso em: 4 ago. 2024.
- ALMEIDA, M. B. Uma abordagem integrada sobre ontologias: Ciência da Informação, Ciência da Computação e Filosofia. **Perspectivas em Ciência da Informação**, v.19, n.3, p.242-258, jul./set. 2014. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22953/18537>. Acesso em: 4 ago. 2024.
- ALPAYDIN, E. **Introduction to Machine Learning**. 3. ed. Massachusetts: MIT Press, 2014.
- ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 132 f. Universidade Estadual Paulista “Júlio Mesquita Filho”, 2010.

AMAZONAS. SEFAZ. **Lei Complementar N.º 84, de 29 de dezembro de 2010.** Modifica dispositivos da Lei Complementar n.º 19, de 29 de dezembro de 1997, que institui o Código Tributário do Estado do Amazonas, e dá outras providências.

Disponível em:

https://online.sefaz.am.gov.br/silt/Normas/Legisla%E7%E3o%20Estadual/Lei%20Complementar%20Estadual/Ano%202010/Arquivo/LCE%20084_10.htm#:~:text=SEFAZ%2FAM%20%2D%20Lei%20Complementar%20Estadual%20084_10&text=Publicada%20no%20DOE%20de%2029.12,Amazonas%2C%20e%20d%C3%A1%20outras%20provid%C3%AAs. Acesso em: 28 jul. 2023.

AMAZONAS. SEFAZ. **Resolução N.º 0028 de 23 de novembro de 2023.**

Modifica a Resolução nº 011/2019-GSEFAZ, que estabelece o valor do preço médio ponderado a consumidor final - PMPF para cálculo do ICMS devido por substituição tributária nas operações com cervejas. Disponível em:

https://online.sefaz.am.gov.br/silt/Normas/Legisla%C3%A7%C3%A3o%20Estadual/Resolu%C3%A7%C3%A3o%20GSEFAZ/Ano%202023/Arquivo/RG%200028_23.htm. Acesso em: 13 dez. 2023.

AMAZONAS. SEFAZ. **Decreto N.º 44.753, de 27 de outubro de 2021.** Aprova o Regimento Interno da Secretaria de Estado da Fazenda - SEFAZ. Disponível em:

https://legisla.imprensaoficial.am.gov.br/diario_am/41538/2021/10/9866. Acesso em 30 set. 2024.

ARAÚJO, C. A. A. **Ciência da informação como ciência social.** Ciência da Informação, Brasília, v. 32, n. 3, p. 21-27, set./dez. 2003. Disponível em:

<https://www.scielo.br/j/ci/a/DZcZXSqTbWHpF6fhRm8b9fP/?format=pdf&lang=pt>. Acesso em: 22 jan. 2024.

BAILEY, S. Do you need a taxonomy strategy?. **Inside Knowledge**, [S. l.], v. 5, n. 5, 2002. Disponível em: <http://www.ikmagazine.com/>. Acesso em: 7 mar. 2024.

BAPTISTA, D. M. A relevância do texto na organização e representação da informação. In: BAPTISTA, D. M.; ARAÚJO JUNIOR, H. (org). **Organização da informação: abordagens e práticas.** Brasília, DF: Thesaurus Editora de Brasília Ltda, 2015. p. 21-43.

BARRETO, A. de A. A questão da informação. **Revista São Paulo em Perspectiva**, São Paulo, v. 8, n. 4, 1994. Disponível em:

<https://bibliotextos.wordpress.com/wp-content/uploads/2012/03/a-questao-da-informac3a7c3a3o.pdf>. Acesso em: 11 jun. 2023.

BASTOS, G. G. **Arquitetura da informação multimodal como suporte ao processo de governança nas organizações.** 2022. Tese (Doutorado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2022.

BEIRA, S. de C. P. de; SIQUEIRA, A. H.; FERNEDA, E.; PRADO, H. A. do. Ontologia como um artefato da arquitetura da informação para a representação do conhecimento organizacional. **Perspectiva em Gestão & Conhecimento**, Joao Pessoa, v. 7, n. 2, p. 122-159, 2017. Disponível em: <https://periodicos.ufpb.br/ojs/index.php/pgc/article/view/26837>. Acesso em: 11 jun. 2023.

BELKIN, N. J. Information concepts for information science. **Journal of Documentation**, [S.l.], v. 34, n. 1, pp. 55-85, 1978. DOI 10.1108/eb026653. Acesso em: 7 mar. 2024.

BLUMER, H. Natureza do interacionismo simbólico. *In*: MORTENSEN, C. D. (org). **Teoria da comunicação: textos básicos**. Tradução C. David Mortensen. São Paulo: Mosaico, p. 119-137, 1980.

BORKO, H. Information Science: What is it?. **American Documentation**, v.19, n.1, p.3-5, Jan. 1968. jan.1968. Disponível em: https://edisciplinas.usp.br/pluginfile.php/1992827/mod_resource/content/1/Borko.pdf. Acesso em: 7 mar. 2024.

BRACHMAN, R. J.; LEVESQUE, H. J. **Competence in Knowledge Representation**. Second Nacional Conference on Artificial Intelligence. Pennsylvania, 1982. Disponível em: <https://cdn.aaai.org/AAAI/1982/AAAI82-045.pdf>. Acesso em: 11 jun. 2023.

BRASIL. Receita Federal. **NMC**. 2019. Disponível em: <https://www.gov.br/receitafederal/pt-br/assuntos/aduana-e-comercio-exterior/classificacao-fiscal-de-mercadorias/ncm>. Acesso em 04 fev. 2024.

BROOKES, B. C. The foundations of information science - part I. **Philosophical aspect. Journal of Information Science**, London, n. 2, p. 125–133, jun 1980. Disponível em: <https://journals.sagepub.com/doi/abs/10.1177/016555158000200302>. Acesso em: 11 jun. 2023.

BUCKLAND, M. Information as Thing. **Journal of the American Society for Information Science**, [S.l.], jun 1991. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-4571%28199106%2942%3A5%3C351%3A%3AAID-ASI5%3E3.0.CO%3B2-3>. Acesso em: 18 set. 2023.

CAMBRIA, E.; WHITE, B. Jumping NLP Curves: A Review of Natural Language Processing Research. **IEEE Computational Intelligence Magazine**, [S.l.], v. 9, n. 2, pp. 48-57, 2014. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6786458>. Acesso em: 11 jun. 2023.

CAPURRO, R; HJORLAND, B. O conceito de informação. **Perspectivas em Ciência da Informação**, Belo horizonte, v. 12, n.1, p. 148-207, jan./abr. 2007. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22360>. Acesso em: 7 mar. 2024.

CARLAN, E. **Sistemas de Organização do Conhecimento**: uma reflexão no contexto da Ciência da Informação. 2010. Dissertação (Mestrado em Ciência da Informação) - Departamento de Ciência da Informação e Documentação, Universidade de Brasília, 2010. Disponível em: <https://core.ac.uk/download/pdf/11887470.pdf>. Acesso em: 11 jun. 2023.

CARTAXO, M. A.; DUQUE, C. G. Aspectos da Arquitetura da Informação envolvidos no mapeamento de processos em Organizações Militares sob a perspectiva semiótica. **Informação & Informação**, Londrina, v. 21, n. 1, p. 103 – 130, 2016. DOI 10.5433/1981-8920.2016v21n1p103. Acesso em: 7 mar. 2024.

CARTAXO, M. A.; BASÍLIO, F. A. C.; DUQUE, C. G. Arquitetura da informação para uma economia da informação. **Informação & Informação**, Londrina, v. 22, n. 1, p. 34 – 59, 2017 DOI 10.5433/1981-8920.2017v22n1p34. Acesso em: 7 mar. 2024.

CARVALHO, E. de O. **Uma proposta de interdisciplinaridade entre arquitetura da informação e ciência da computação**: linguagem SOWL para as ontologias da Web utilizando o formalismo dos grafos conceituais. 2013. 248 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, DF, 2013.

CASTELLS, M. A Sociedade em Rede: do Conhecimento à Política. *In*: CASTELLS, M.; CARDOSO, G. **A Sociedade em Rede**: Do Conhecimento à Ação Política. Brasília: Imprensa Nacional-Casa da Moeda, 2006.

CHANDRA, R.; A. SHUKLA, A.; TIWARI, S.; AGARWAL, S.; SVAFRULLAH, S.; ADIYARTA, K. Natural Language processing and Ontology based Decision Support System for Diabetic Patients. *In*: INTERNATIONAL CONFERENCE ON ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION, 9., 2022. **Anais** [...]. Indonesia, 2022. DOI [10.23919/EECSI56542.2022.9946601](https://doi.org/10.23919/EECSI56542.2022.9946601). Acesso em: 18 set. 2023.

CHANG, T.; KAO, H.; WU, J.; HSIAO, K.; CHAN, T. Integrated ontology based approach with navigation and content representation for health care website design. **Computer in Human Behavior**, [S.l.], v. 128, n. C, 2022. DOI 10.1016/j.chb.2021.107119. Acesso em: 20 maio 2024.

CHOMSKY, N. **Knowledge of language**: it's nature, origin, and use. New York: Praeger Publishers, 1986.

CHUNG, K.; YOO, H.; CHOE, D. Ambient context-based modeling for health risk assessment using deep neural network. **Journal of Ambient Intelligence and Humanized Computing**, [S.l.], v. 11, 2020. Disponível em: https://www.researchgate.net/publication/327650836_Ambient_context-based_modeling_for_health_risk_assessment_using_deep_neural_network. Acesso em: 20 maio 2024.

COSERIU, E. **Introducción a la Lingüística** - vol. 3. Madrid: Editora Gredos S. A., 1986. Disponível em: <https://books.google.com.br/books?id=JMEbAQAAIAAJ>. Acesso em: 4 nov. 2023

COSTA, I. de M.; LIMA-MARQUES, M. MAIA - Método de Arquitetura da Informação aplicada: constructo metodológico de tratamento da informação em contextos complexos. **Informação & Informação**, Londrina, v. 22, n. 1, p. 60 – 87, jan./abr., 2017. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/24008>. Acesso em: 20 maio 2024.

COSTA, I. M. **Um Método para Arquitetura da Informação**: Fenomenologia como base para o desenvolvimento de arquiteturas da Informação aplicadas. Dissertação (Mestrado em Ciência da Informação) – Universidade de Brasília 2009. Disponível em: <http://www.realp.unb.br/jspui/handle/10482/7087>. Acesso em: 20 maio 2024.

DEMO, P. **Introdução à metodologia da ciência**. 2. ed. São Paulo: Atlas, 1985.

DILLON, A. Information architecture in JASIST: just where did we come from?. **Journal of the American Society for Information Science and Technology**, [S.l.], v. 53, n.1, p. 821-823, 2002. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10090>. Acesso em: 23 maio 2024.

DILLON, A.; TURNBULL, D. Information architecture. In: Drake, M. (ed.). **Encyclopedia of Library and Information Science**: First Update Supplement. [S.l.]: Taylor & Francis, 2005.

DING, W.; LIN, X. **Information Architecture**: the design and integration of information space. [S.l.]: Springer Cham, 2010.

DOWNEY, L.; BANERJEE, S. Building an Information Architecture Checklist – Encouraging and Enabling IA from Infrastructure to the User Interface Architecture. **Journal of Information Architecture**, [S.l.], v. 2, n. 2, p. 27 – 46, 2010. Disponível em: <http://journalofia.org/volume2/issue2/03-downey/>. Acesso em 22 jul. 2024.

DUQUE, C. G. **SiRILiCO uma proposta para um sistema de Recuperação da Informação baseado em Teorias da Lingüística Computacional e Ontologia**. 2005. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005. Disponível em: <https://repositorio.ufmg.br/handle/1843/EARM-7HBND8>. Acesso em 22 jul. 2024.

EDITORA MELHORAMENTOS. Michelis Dicionário Brasileiro da Língua Portuguesa: Cerveja. **UOL**, 2024. Disponível em: <https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/cerveja/>. Acesso em 26 maio 2024.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. de L. F. de. **Inteligência artificial**: uma abordagem de aprendizado de máquina. Rio de Janeiro, Editora LTC, 2021.

FARIA, I. H.; PEDRO, E. R.; DUARTE, I.; GOUVEIA, C. A.M. (org.). **Introdução à Linguística Geral e Portuguesa**. Lisboa: Caminho. 1998.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. Da Mineração de Dados à Descoberta de Conhecimento em Bancos de Dados. **AI Magazine**, [S.l.], v. 17, n. 3, p. 37-42, 1996. DOI 10.1609/aimag.v17i3.1230. Acesso em: 12 ago. 2024.

FERNÁNDEZ, M.; GÓMEZ-PÉREZ, A.; JURISTO, N. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. **Proceedings of the Ontological Engineering AAAI97**, Palo Alto, 1997. Disponível em: <https://aaai.org/papers/0005-ss97-06-005-methontology-from-ontological-art-towards-ontological-engineering/>. Acesso em: 27 set. 2023.

FELIPE, E.; SOUZA, A.D de. KOS na recuperação da informação multilíngue em literatura biomedicina. In: ALMEIDA, M.B. (org). **Representação do conhecimento, ontologias e linguagem**. Curitiba, PR: Editora CRV, 2020. p. 83-112.

FOX, M. S. The TOVE Project Towards a Common-Sense Model of the Enterprise. In: BELLI, F.; RADERMACHER, F. J. (ed.) **Industrial and engineering applications of artificial intelligence and expert systems**. London: [S.n.], 1993. p. 25–34.

FROSSARD, D. **ICMS Genérico**. Rio de Janeiro, Editora Ferreira, 2011.

GARSHOL, L. M. **Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all**. In Journal of Information Science. v. 30, n. 4, p. 378-391, 2004.

GHIDALIA, S.; NARSIS, O. L.; BERTAUX, A.; NICOLLE, C. Combining Machine Learning and Ontology: A Systematic Literature Review. **ArXiv preprint**, [S.l.], 2024. DOI arXiv:2401.07744. Acesso em: 27 set. 2023.

GOLDSCHIMIDT, R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.

GONZÁLEZ DE GÓMEZ, M. N. Metodologia de pesquisa no campo da Ciência da Informação. **DataGramaZero - Revista de Ciência da Informação**, [S.l.], v.1 n.6, dez. 2000. Acesso em: 27 set. 2023. <https://ridi.ibict.br/bitstream/123456789/127/1/GomesDataGramaZero2000.pdf>. Acesso em: 27 set. 2023.

GONZALEZ, M.; LIMA, V. L. S. **Recuperação de informação e processamento da linguagem natural**. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23.v. 3, 2003. **Anais** [...]. Marília, SP, p.346-395, 2003. Disponível em: <https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/mri-06--gonzales-e-lima-2003.pdf>. Acesso em: 22 set. 2023.

GORAYEB, D. M. C.; DUQUE, C. G. Planejamento de um ambiente informacional automatizado para a extração de termos relevantes à fiscalização em nota fiscal eletrônica e a nota fiscal de consumidor eletrônica. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 22., 2022. **Anais** [...]. Porto Alegre, RS, 2022. Acesso em: 27 set.

2023. <https://enancib.ancib.org/index.php/enancib/xxii/enancib/paper/view/870>. Acesso em: 27 set. 2023.

GORAYEB, D. M. da C.; DUQUE, C. G. Proposta de metadados para descrição de produtos da Nota Fiscal de Consumidor Eletrônica (NFC-e) usando Apriori. **P2P & Inovação**, Rio de Janeiro, v. 11, n. 1, p. 1-24, e-7124, jul./dez.2024.

GRUBER, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal Human-Computer Studies**, [S.l.], v. 43, Padova, Italy, p. 907-928, 1993.

GRÜNINGER, M.; FOX, M.S. The Role of Competency Questions in Enterprise Engineering. In: ROLSTADÅS, A. (ed.) **Benchmarking** — Theory and Practice. Boston: IFIP Advances in Information and Communication Technology, 1995. p. 22-31.

HARALAMBOUS, Y. **A course in Natural Language Processing**. France: Springer, 2024.

HARIDY, S.; ISMAIL, R. M; BADR, N.; HASHEM, M. An Ontology Development Methodology Based on Ontology-Driven Conceptual Modeling and Natural Language Processing: Tourism Case Study. **Big Data and Cognitive Computing**, ano 7, v. 101, n. 2, p. 1-23, 2023. DOI [10.3390/bdcc7020101](https://doi.org/10.3390/bdcc7020101). Acesso em: 7 maio 2023.

HINTON, A. The Machineries of Context: New Architectures for a New Dimension. **Journal of Information Architecture**, [S. l.], v. 1, iss. 1, p. 45-58, 2009. Disponível em: <https://journalofia.org/volume1/issue1/04-hinton/jofia-0101-04-hinton.pdf>. Acesso em: 13 maio 2023.

INFORMATION ARCHITECTURE INSTITUTE. **What is Information Architecture**. 2013. Disponível em: http://iainstitute.org/documents/learn/What_is_IA.pdf. Acesso em: 22 jul. 2024.

ISO. **ISO 10746-1: 1998**: Information technology — Open Distributed Processing — Reference model: Overview. Genebra: ISO, 1998.

ISO. **ISO 23081-1:2017**: Information and documentation — Records management processes — Metadata for records — Part 1: Principles. Genebra: ISO, 2017.

ISO. **ISO 23081-2:2021**: Information and documentation — Metadata for managing records — Part 2: Conceptual and implementation issues. Genebra: ISO, 2021.

ISO. **ISO 23081-3:2011**: Information and documentation — Managing metadata for records — Part 3: Self-assessment method. Genebra: ISO, 2011.

ISO. **ISO 25964-1:2011**. Information and documentation — Thesauri and interoperability with other vocabularies: part 1: Thesauri for information retrieval. 1. ed. Genebra: ISO, 2011.

JORGE, E. M. F.; SANTOS, F. P. dos; CARNEIRO, B. P. B.; MACHADO, F. A. Arquitetura da informação analítica para integração de dados da pesquisa e pós-graduação: um estudo de caso da Universidade do Estado da Bahia. **Informação & Informação**, Londrina, v. 25, n. 1, p. 115–140, jan./mar. 2020. Disponível em: <https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/36009>. Acesso em 26 maio 2024.

KUROKI JÚNIOR, G. H.; DUQUE, D. G. Arquitetura da informação aplicada ao processamento de linguagem natural: uma proposta. Contribuições da Ciência da Informação no pré-processamento de dados para treinamento e aprendizagem de redes neurais artificiais. **RDBCI**, Campinas, SP, e.21, e023002, 2023. DOI 10.20396/rdbci.v21i00.8671396/30919. Acesso em: 14 jul. 2023.

LAKATOS, E. M.; MARCONI, M. de A. **Fundamentos de metodologia científica**. 8. ed. São Paulo: Atlas, 2017.

LÉLIS, C. A. S.; BRAGA, R.; ARAÚJO, M. A. P.; DAVID, J. M. N. ArchiRI – uma arquitetura baseada em ontologias para troca de informações de reputação. In: BRASILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 12., 2016. **Anais [...]**. Florianópolis, SC, 2016. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/5946/5844>. Acesso em: 14 jul. 2023.

LIMA-MARQUES, M. Outline of a theoretical framework of Architecture of Information: a School of Brasilia proposal. In: BÉZIAU, J.; CONIGLIO, M. E. (ed.). **Logic without Frontiers: Festschrift for Walter Alexandre Carnielli on the occasion of his 60th Birthday**. London: College Publications, 2011.

LIMA-MARQUES, M.; MACEDO, F. L. O. Arquitetura da informação: base para a Gestão do Conhecimento. In: TARAPANOFF, K. O. (ed.). **Inteligência, informação e conhecimento**. Brasília, DF: IBICT, 2006. p. 241-255.

MACULAN, B. C. N. dos S.; LIMA, G. A. B. de O. Buscando uma definição para o conceito de “conceito”. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.22, n.2, p.54-87, abr./jun. 2017. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22503/18096>. Acesso em: 14 jul. 2023.

MAIMONE, G. D.; SILVEIRA, N. C.; TÁLAMO, M. DE F. G. M. Reflexões acerca das relações entre representação temática e descritiva. **Inf. & Soc.:Est.**, João Pessoa, v.21, n.1, p. 27-35, jan./abr. 2011. Disponível em: <https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/7367/5596>. Acesso em: 5 jul. 2023.

MARTINS, S. SEFAZ: saiba o que é, suas funções e mais. **Mainô Blog**, 4 jul. 2023. Disponível em: <https://blog.maino.com.br/sefaz/>. Acesso em: 14 jul. 2023.

MBOLI, J. S.; THAKKER, D.; MISHRA, J. L.; SIVARAJAH, S. Domain Experts and Natural Language Processing in the Evaluation of Circular Economy Business Model

Ontology. *In*: INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING (ICSC), 15., 2021. **Anais [...]**. Virtual Conference, 2021. DOI 10.1109/ICSC50631.2021.00069. Acesso em: 14 jul. 2023.

MENDONÇA, E. S. A linguística e a ciência da informação: estudos de uma interseção. **Ciência da Informação**, Brasília, v. 29, n. 3, p. 50-70, set./dez. 2000. Disponível em: <https://revista.ibict.br/ciinf/article/view/873/907>. Acesso em: 5 jul. 2023.

MENDONÇA, F. M. **OntoForInfoScience**: metodologia para construção de ontologias pelos cientistas da informação – uma aplicação prática no desenvolvimento da ontologia sobre componentes no sangue humano (HEMONTTO). Tese (Doutorado) - Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brasil, 2015.

MIRANDA, S. K. O.; MARCELINO, M. J.; SILVA, O. A. **An Ontology to Trace the Computer Science Student Profile**. 2023. IEEE Frontiers in Education Conference (FIE), College Station, TX, USA, 2023. p. 1-5.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Ontologias**: conceitos, usos, tipos, metodologias, ferramentas e linguagens. Relatório Técnico. Goiás: Instituto de Informática - Universidade Federal de Goiás: 2007. Disponível em: https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-07.pdf. Acesso em: 5 jul. 2023.

MORI, A.; CARVALHO, C. L. de. **Metadados no Contexto da Web Semântica**. Relatório Técnico. Instituto de Informática Universidade Federal de Goiás, 2004. Disponível em: https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_002-04.pdf. Acesso em: 4 ago. 2024.

MORI, A. **Modelagem do conhecimento baseada em ontologias aplicada às Políticas Públicas e Habitação**. 2009. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2009.

NADEAU, D.; SATOSHI S. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, v.30, p-2-36, 2007.

NOVAES, D. Reflexões linguísticas para a organização hierárquica de conceitos em tesouros. *In*: DUQUE, C. G (org). **Ciência da Informação Estudos e Práticas**. Brasília, DF: Thesaurus Editora de Brasília Ltda, 2011. p. 237-250.

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101: A guide to creating your first Ontology**. California: Stanford Knowledge Systems Laboratory Technical Report KSL-01-05; Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

OLIVEIRA, A. C. de S. **Linguística Computacional**: um mapeamento bibliográfico de 2000 a 2020. **Interface Tecnológica**, v. 17, n. 2, p. 376 – 385, 2020. DOI 10.31510/inf.v17i2.1056. Acesso em: 5 jul. 2023.

OTHERO, G. de A. Linguística computacional: uma breve introdução. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 341-351, jun. 2006.

OTLET, P. **Documentos e Documentação**. Tradução por Hagar Espanha Gomes. Paris, 1937. Disponível em: <http://www.conexaorio.com/bit/otlet/index.htm>. Acesso em: 30 jun. 2024.

PINHEIRO, L. V. R. Comunidades científicas e infra-estrutura tecnológicas no Brasil para uso de recursos eletrônicos de comunicação e informação na pesquisa. **Ciência da Informação**, Brasília, v. 32, n.3, p. 62-73, set./dez. 2003.

POPPER, K. R. **Lógica das ciências sociais**. Brasília: Universidade de Brasília, 1978.

PRODANOV, C. C.; FREITAS, E. C. de. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013.

PROMKOT, A.; ARCH-INT, S.; ARCH-INT, N. **The Personalized Traditional Medicine Recommendation System Using Ontology and Rule Inference Approach**. Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, p. 96–104, 2019.

RAMOS, A. L. T.; LORINI, F. **Architecture Information Context in a Design for Manufacturing (DFM) Framework**. 11. ed. Sao Paulo, SP: IFAC Workshop on Intelligent Manufacturing Systems., 2013.

ROSENFELD, L.; MORVILLE, P.; ARANGO, J. **Information Architecture**: for the web and beyond. 4. ed. Canadá: Ed. O'reilly Media Inc., 2015.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectiva em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22308>. Acesso em 28 jun. 2024.

SCHEIDER, E. T. R. BioBERTpt – A Portuguese neural language model for clinical named entity recognition 3. ed. **Clinical Natural Language Processing Workshop**, [on-line], 2020.

SCHULZE, M.; SCHRODER, M; JILEK, C.; ALBERS, T.; MAUS, H.; DENGEL, A. P2P-O: A Purchase-To-Pay Ontology for Enabling Semantic Invoices. **ESWC 2021: The Semantic Web Lecture Notes in Computer Science**, v. 12731. p 647–663, 2021.

SEFAZ/AM. **Guia Prático Escrituração Fiscal Digital - EFD – Versão 2.0.4**. Disponível em:

<http://dbcon.sefaz.am.gov.br/efd/arquivos/Guia%20Pr%C3%A1tico%20da%20EFD%20-%20Vers%C3%A3o%202.0.4.pdf>. Acesso em: 7 maio 2023.

SHANNON, C. E.; WEAVER, W. **The Mathematical Theory of Communication**. Champaign: University of Illinois Press, 1949.

SILVA, D. L. da. **Uma proposta metodológica para construção de ontologias: uma perspectiva interdisciplinar entre as ciências da informação e da computação**. 2008. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Minas Gerais, 2008.

SIQUEIRA, A. H. de. **Arquitetura da Informação: uma proposta para a fundamentação e caracterização de uma disciplina científica**. 2012. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, 2012.

SOBHANI F.; IZQUIERDO E.; PIATRIK T. **Ontology-based forensic event detection using inference rules**. 2017. International Conference on Engineering, Technology, and Innovation (ICE/ITMC), Portugal, pp. 584-591, 2017.

SORATO, D.; GOULARTE, F. B.; NASSAR, S. M.; FILETO, R. Analysis of Methods and Tools for Relevant Words Recognition in Microblogs. In: BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 12., 2016. **Anais [...]**. Florianópolis, 2016.

STANDAERT, L.; YAROSLASKI, A.; CASTRO, M. de. **Beer Advisor – A beer ontology**. Vancouver: Association for the Advancement to Artificial Intelligence, 2021.

SUÁREZ-FIGUEROA, M. C.; GÓMEZ-PÉREZ, A.; FERNANDÉZ-LÓPEZ, M. The NeOn framework: A scenario-based methodology for ontology development. **Applied Ontology**, [S.l.], v. 10 (n.2), pp. 105-145, 2015.

SURE, Y.; STUDER, R. A. A Methodology of Ontology-based Knowledge Management. In: DAVIES, J.; FENSEL, D.; VAN HARMELEN, F. (ed.). **Towards the Semantic Web: ontology-driven knowledge management**. Chichester: John Wiley & Sons, pp. 33-46, 2003.

TARUS, J. K.; NIU, Z.; MUSTAFA, G. Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning. **Artificial Intelligence Review**, [S.l.], v. 50, p. 21-48, 2018.

TOMS, E. G. Information interaction: providing a framework for information architecture. **Journal of the American Society for Information Science and Technology**, [S.l.], v. 53, n. 10, p. 855-862, 2002.

VIEIRA, A. da S. Caminhos transdisciplinares para a formação de bibliotecários. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 12, n. 2, p. 250-263, set.1983.

VICTORINO, M. de C.; PINHEIRO, M.S.; SANTOS, R. F. dos. Organização da informação e do conhecimento em sistemas de informação transacionais para o seu

reuso em sistemas de apoio à decisão. *In*: BAPTISTA, D. M.; ARAÚJO JUNIOR, H. (org). **Organização da informação: abordagens e práticas**. Brasília, DF: Thesaurus Editora de Brasília Ltda, 2015. p. 291-247.

WARREN, R. **The Beer Ontology**. 2024. Disponível em: <https://rdf.ag/o/beer-en.html>. Acesso em 4 mar. 2024.

WERSIG, G. Information Science: the study of postmodern knowledge usage. **Information Processing & Management**, [S.l.], v. 29, n. 2, p. 229-239, 1993.

WITTGENSTEIN, L. **Investigações filosóficas**. Tradução Marcos G. Montagnoli. 9. ed. São Paulo: Editora Vozes, 2014.

WOLEDGE, G. Bibliography and Documentation: words and ideas. **Journal of Documentation**, v. 39, n. 4, p. 266-279, 1983. Disponível em: <https://pt.scribd.com/document/592386149/Woledge-G-1983-Historical-Studies-in-Documentation-Bibliography-and-Documentation>. Acesso em 4 mar. 2024.

WURMAN, R. S. **Information Architects**. [S.l.]: Graphis Inc, 1997.

YANG, B. Construction of logistics financial security risk ontology model based on risk association and machine learning. *Safety Science*, [S.l.], v. 123, 2020.

ZACHMAN, J. A. A framework for information systems architecture. **IBM Systems Journal**, [S.l.], v. 38, n. 2-3, p. 454-470 1999. a reprint of 1987 version.

ZINS, C. Conceptions of Information Science. *Journal of the American Society for Information Science and Technology*, [S.l.], p. 335-350, 2007.

ANEXO A – SOLICITAÇÃO PARA COLABORAÇÃO NA PESQUISA

Ao Ilustríssimo Senhor,

Secretário de Estado da Fazenda do Amazonas– SEFAZ/AM
Dr. Alex Del Giglio

Prezado Senhor,

Ao cumprimentá-lo respeitosamente, venho a V.Exa. para solicitar colaboração no sentido de **compartilhar informações sobre a base de dados da nota fiscal eletrônica**: Nota Fiscal Eletrônica (NF-e) e Nota Fiscal de Consumidor Eletrônica (NFC-e), tratadas neste lúdimo órgão de administração financeira do Estado do Amazonas.

Eu, Diana Maria da Camara Gorayeb, servidora pública do Estado do Amazonas, professora concursada na instituição de ensino superior, Universidade do Estado do Amazonas (UEA), lotada na Escola Superior de Tecnologia (EST) no núcleo de Engenharia da Computação e outras Ciências Tecnológicas, venho por meio deste esclarecer o pedido acima proposto nos fatos a seguir:

1. Desde maio de 2021 esta professora foi aprovada em edital público nacional para vaga de doutoramento pela Universidade de Brasília, buscando manter-se na excelência do quadro docente da UEA com o título de Dra. em Ciência da Informação;
2. O projeto de pesquisa proposto à tese de doutorado busca desenvolver um componente de software para auxiliar o **planejamento, supervisão e análise das informações dos produtos nas notas fiscais eletrônicas** utilizando algoritmos de inteligência artificial e outras tantas técnicas de organização e arquitetura e recuperação da informação e do conhecimento;
3. O resultado esperado é uma ferramenta computacional inteligente que ajude no **rastreamento dos produtos descritos em notas fiscais de entrada e saída com critérios definidos a partir da necessidade atuais e de conformidade com a legislação tributária competente**;
4. Para tanto, o algoritmo inteligente faria o estudo sobre **um recorte nos produtos das notas fiscais mantendo e preservando o sigilo fiscal sobre o contribuinte bem como sua relação com a atividade comercial ou industrial exercida por meio de mascaramento de informações sensíveis e de dados sigilosos**. O mascaramento não influenciará nos resultados do projeto ou prejuízos à pesquisa, já que o percentual de acerto na avaliação e validação do conteúdo pelo algoritmo inteligente é que demonstrará se a pesquisa é relevante ou não ou se os

resultados foram ou não alcançados. Neste sentido, esta pesquisadora fica a disposição para colaborar na definição das soluções adequadas ao projeto;

5. O projeto de pesquisa, em sua íntegra, está detalhado em documento complementar à este pedido.

Trata-se, portanto, ilustríssimo Senhor, de um projeto de pesquisa de grande relevância para a educação e ciência no âmbito nacional, cujos resultados serão apresentados por meio de artigo científicos e na tese de doutoramento. Reforço que todo o material de pesquisa resultante, bem como o componente inteligente de software será entregue à esta Casa para o uso na área de tecnologia, se assim desejado.

Por fim, acreditando no interesse deste órgão em investir em tecnologia, promover a melhoria das atividades fiscais, colaborar com a construção de ferramentas de repressão aos atos ilícitos tributários, e na forma cooperativa de um grande projeto de pesquisa que traga resultados aplicáveis futuramente em sistemas automatizados da SEFAZ/AM venho mui respeitosamente solicitar acesso as informações e aos dados necessários: **base de dados da nota fiscal eletrônica Nota Fiscal Eletrônica (NF-e) e Nota Fiscal de Consumidor Eletrônica (NFC-e).**

Nestes termos pedimos,
Deferimento.

Atenciosamente,

DIANA MARIA DA
CAMARA

GORAYEB:32155930259

Assinado de forma digital por
DIANA MARIA DA CAMARA
GORAYEB:32155930259

Dados: 2022.03.28 10:34:13 -04'00'

MSc. Diana Maria da Camara Gorayeb
Professora do Núcleo de Computação
Escola Superior de Tecnologia – EST.
Universidade do Estado do Amazonas – UEA.

ANEXO B – DESPACHO FUNDAMENTADO DA SOLICITAÇÃO PARA COLABORAÇÃO NA PESQUISA



PROCESSO Nº: 01.01.014101.1025392022-45
INTERESSADO(A): DIANA MARIA DA CAMARA GORAYEB
DO: DEPARTAMENTO DE TRIBUTAÇÃO – DETRI
À: SECRETARIA EXECUTIVA DA RECEITA – SER

DESPACHO FUNDAMENTADO

- 1 Trata-se de solicitação apresentada pela interessada (fls. 7-8) para ter acesso à base de dados da Nota Fiscal Eletrônica (NF-e) e Nota Fiscal de Consumidor Eletrônica (NFC-e), administrada pela SEFAZ.
- 2 Os dados acessados serão utilizados na pesquisa desenvolvida pela interessada para sua tese de doutoramento.
- 3 A interessada informa que o estudo será conduzido utilizando-se "um recorte nos produtos das notas fiscais, mantendo e preservando o sigilo fiscal sobre o contribuinte, bem como sua relação com a atividade comercial ou industrial exercida por meio de mascaramento de informações sensíveis e de dados sigilosos".
- 4 Nos termos do art. 198 da Lei nº 5.172/66, o Código Tributário Nacional (CTN), o dever de guardar sigilo pela Fazenda Pública e seus servidores incide sobre informações acerca da situação econômica ou financeira do sujeito passivo ou de terceiros e da natureza e do estado de seus negócios ou atividades.
- 5 Sendo assegurada a preservação dos dados resguardados pelo dever de sigilo, não se vislumbra óbice para o atendimento da solicitação, mediante manifestação do Departamento de Tecnologia da Informação (DETIN) quanto à adoção de medidas que atendam ao disposto nas alíneas "a" e "b" do subitem 7.1 da minuta de Termo de Acordo constante dos autos (fls. 13).

Manaus, 20 de junho de 2022.

Alan Cesar Monteiro Corrêa
Auditor Fiscal de Tributos Estaduais

Luiz Aurélio C. Leite
Chefe do DETRI

ANEXO C – DESPACHO DE AUTORIZAÇÃO PARA ACESSO AOS DADOS E INFORMAÇÕES DA BASE DE DADOS SEFAZ/AM

Página 1 de 1

 AMAZONAS <small>GOVERNO DO ESTADO</small>	
SECRETARIA DE ESTADO DA FAZENDA – SEFAZ SECRETARIA EXECUTIVA DA RECEITA – SER	
DESTINATÁRIO	GSEFAZ
PROCESSO Nº	102539/2022-45
INTERESSADA	DIANA GORAYEBE
ASSUNTO	ACESSO NOTA FISCAL ELETRONICA
DESPACHO	
<p>Ao GSEFAZ,</p> <p>Considerando que os dados solicitados possuem natureza sigilosa, e por conseguinte requerem uma análise pormenorizada antes de sua divulgação a terceiro, considerando ainda a atual carência no quadro de pessoal com know-how para o gerenciamento do acesso da Requerente, entendemos não ser, por ora, oportuno conceder o acesso na forma solicitada.</p> <p>Entretanto, considerando que os trabalhos da pesquisa proposta podem trazer subsídios para atividades futuras desta Sefaz, esta SER fica a disposição para atender a Requerente com os dados e informações que julgar necessário, resguardando os dados de natureza sigilosa, na forma prevista na legislação.</p>	
Gabinete do Secretário Executivo da Receita, em Manaus, 29 de agosto de 2022.	
DARIO JOSÉ BRAGA PAIM Secretário Executivo da Receita	

Avenida André Araújo, 150 - Aleixo
 Fone: (92) 2121-1600
 Manaus-AM - CEP 69060-000

Secretaria de
Fazenda

