UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

DIANA MARIA DA CAMARA GORAYEB

MINERAÇÃO DE TEXTO USANDO ARQUITETURA DA INFORMAÇÃO E ONTOLOGIA COMO MÉTODO PARA AUXÍLIO DE AUDITORIA EM DOCUMENTOS DIGITAIS

BRASÍLIA, DF

2025

DIANA MARIA DA CAMARA GORAYEB

MINERAÇÃO DE TEXTO USANDO ARQUITETURA DA INFORMAÇÃO E ONTOLOGIA COMO MÉTODO PARA AUXÍLIO DE AUDITORIA EM DOCUMENTOS DIGITAIS

Tese apresentada 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: Gestão, Tecnologias e Organização da Informação e do Conhecimento.

Orientador: Dr. Cláudio Gottschalg Duque

BRASÍLIA, DF

2025

Dados Internacionais de Catalogação na Fonte

G661 Gorayeb, Diana Maria da Camara.
Mineração de Texto usando Arquitetura da Informação e ontologia como método para auxílio de Auditoria em documentos digitais / Diana Maria da Camara Gorayeb. _ Brasília, DF, 2025.
201 f.: il., color.; 27 cm.

Orientador: Cláudio Gottschalg Duque.
Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Informação, Universidade de Brasília.

1. Ontologia – Nota Fiscal de Consumidor Eletrônica. 2. Mineração de texto. 3. Arquitetura da Informação. I. Duque, Cláudio Gottschalg. II. Univesidade de Brasília III. Título.

CDU – [003.68:004.912.37](811.3)(043.3)

Bibliotecária Responsável: Thaís Lima Trindade – CRB11/ 687

UNIVERSIDADE DE BRASÍLIA

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Ata Nº: 102

Aos três dias do mês de dezembro do ano de dois mil e vinte e cinco, instalou-se a banca examinadora de Tese de Doutorado da aluna **Diana Maria da Câmara Gorayeb**, matrícula 210002972. A banca examinadora foi composta pelos professores Dr. Clovis Carvalho Britto (Titular Interno, PPGCINF), Dra. Guilhermina de Melo Terra (Titular Externa à UnB, UFAM), Dr. Rômulo Ferreira dos Santos (Titular Externo à UnB, EB), Dr. Mac Amaral Cartaxo (Suplente externo brasileiro, SENAC DF) e Dr. Claudio Gottschalg Duque, orientador(a)/presidente. A discente apresentou o trabalho intitulado "Mineração de Texto Usando Arquitetura da Informação e Ontologia como Método para Auxílio de Auditoria em Documentos Digitais.

Concluída a exposição, procedeu-se a arguição do(a) candidato(a), e após as considerações dos examinadores o resultado da avaliação do trabalho foi:

(X) Pela aprovação do trabalho;

( ) Pela aprovação do trabalho, com revisão de forma, indicando o prazo de até 30 (trinta) dias para apresentação definitiva do trabalho revisado;

( ) Pela reformulação do trabalho, indicando o prazo de (Nº DE MESES) para nova versão;

( ) Pela reprovação do trabalho, conforme as normas vigentes na Universidade de Brasília.

Conforme os Artigos 34, 39 e 40 da Resolução 0080/2021 - CEPE, o(a) candidato(a) não terá o título se não cumprir as exigências acima.

Dr. Claudio Gottschalg Duque
(Presidente)

Dr. Clovis Carvalho Britto, PPGCINF/UnB
(Membro Titular Interno)

Dr.(a) Guilhermina de Melo Terra, UFAM
(Membra Titular Externo)

Dr.Rômulo Ferreira dos Santos, EB
(Membro Titular externo)

Dr.Mac Amaral Cartaxo, SENAC/DF
(Suplente)

Diana Maria da Câmara Gorayeb
(Doutoranda)

Documento assinado eletronicamente por **Claudio Gottschalg Duque**, **Pesquisador(a) Colaborador(a) Pleno(a) da Faculdade de Ciência da Informação**, em 09/12/2025, às 10:50, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Diana Maria da Camara Gorayeb**, **Usuário Externo**, em 09/12/2025, às 15:10, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Clovis Carvalho Britto**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 09/12/2025, às 15:34, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **GUILHERMINA DE MELO TERRA**, **Usuário Externo**, em 09/12/2025, às 15:35, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **ROMULO FERREIRA DOS SANTOS**, **Usuário Externo**, em 09/12/2025, às 19:48, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Ana Lucia de Abreu Gomes**, **Vice-Coordenador(a) da Pós-Graduação da Faculdade de Ciência da Informação**, em 10/12/2025, às 17:00, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **13479087** e o código CRC **6DDC1972**.

Referência: Processo nº 23106.116641/2025-69 SEI nº 13479087

AGRADECIMENTOS

Aos meus filhos queridos e amados por toda a vida que me nutrem de amor diariamente.

Aos meus pais Ernani Villar (*In memoriam*) e Dalva Maria e a todos da minha família que torcem e festejam as minhas conquistas.

Ao estimado professor, Dr. Cláudio Gottschalg Duque, expresso minha profunda gratidão pelas suas competência e generosidade que me conduziram ao longo de toda realização do doutorado.

Agradeço ao corpo docente e técnico do Programa de Pós-Graduação em Ciência da Informação (PPGCI) da Universidade de Brasília (UnB) pela excelência no ensino, pela dedicação e pelo suporte oferecido ao longo de toda minha trajetória no doutorado.

Agradeço à Universidade Federal do Amazonas (UFAM) pela promoção e realização do Programa de Doutorado Interinstitucional (DINTER), que possibilitou a concretização deste importante passo na minha formação acadêmica e o fortalecimento da pesquisa na região norte do país.

Agradeço à Secretaria da Fazenda do Amazonas (SEFAZ-AM) pelo apoio e incentivo durante o desenvolvimento desta pesquisa.

Aos colegas do R.E.G.I.I.M.E.N.T.O., que sempre promoveram e compartilharam ideias e soluções criativas e científicas na descoberta do conhecimento.

Aos amigos que participaram do DINTER no Amazonas, Daniele Pontes, Marcos Falcão, Thaís Trindade e Vanusa Jardim, agradeço o estímulo e carinho promovidos ao longo dos anos de estudo e que levarei por toda vida.

Aos colegas do Departamento de Polícia Técnico Científica (DPTC), meu profundo agradecimento pelo apoio, compreensão e espírito de equipe demonstrados durante os anos de formação deste doutorado que permitiram que eu pudesse me dedicar aos estudos sem comprometer o serviço público que compartilhamos.

Agradeço aos professores do Núcleo de Computação Aplicada (NUCOMP) da Universidade do Estado do Amazonas (UEA) pelo apoio humano prestado e disponibilidade demonstrada em diversos momentos ao longo da minha formação.

À professora, Dra. Guilhermina de Melo Terra, que aceitou o convite para participar da banca de defesa desta tese e que generosamente dedicou seu tempo para a leitura atenta e para apresentar valiosas contribuições.

Ao professor, Dr. Clóvis Carvalho Britto, que aceitou o convite para participar da banca de defesa desta tese e que generosamente dedicou seu tempo para a leitura atenta e para apresentar valiosas contribuições.

Ao professor, Dr. Romulo Ferreira dos Santos, que aceitou o convite para participar da banca de defesa desta tese e que generosamente dedicou seu tempo para a leitura atenta e para apresentar valiosas contribuições.

Por fim, à Nossa Senhora por interceder por mim ao Deus Único, jamais me abandonar e manter meu coração aquecido na fé, eu Vos consagro neste dia e para sempre o meu coração e inteiramente todo o meu ser e saber.

.

DEDICATÓRIA

Ao meu amor Marcel Gomes de Carvalho, você é o meu lar, meu coração, meu lugar seguro, meu melhor amigo. Sem palavras para expressar o tudo.

RESUMO

Este trabalho, cujo objetivo geral é auxiliar o processo de auditoria em Documentos Fiscais Eletrônicos por meio da elaboração de um modelo de ontologia a partir da Arquitetura da Informação e da Mineração de Texto para validar a informação de descrição e venda do produto, se configura como uma pesquisa na área da Ciência da Informação. No que tange aos objetivos específicos, delimitou-se em: Identificar possíveis requisitos de uma Arquitetura da Informação para a Mineração de Texto em Notas Fiscais de Consumidor Eletrônicas e Notas Fiscais Eletrônicas no modelo de ontologia para o produto cerveja; Definir as principais informações extraídas do produto quando aplicada à Mineração de Texto nas Notas Fiscais; Descrever a relevância dos sistemas de organização do conhecimento, especificamente, da ontologia para os processos de organização e recuperação da informação. Apresenta como base teórica o enfoque nos conceitos e nas 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 proposta de 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 e exploratória; 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 Secretaria de Fazenda do Estado do Amazonas 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 e Notas Fiscais Eletrônicas, selecionadas a partir da Nomenclatura Comum do Mercosul para o produto cerveja. Para a elaboração da ontologia, o conhecimento será um conjunto de padrões cuja formulação pode envolver e relacionar dados e informações, cuja 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. Os resultados foram apresentados na forma de respostas às Questões de Competência que retomam consultas no formato *DLQueries* e identificam corretamente o produto interpretando e compondo elementos em quantidade e qualidade suficiente para sua identificação e utilização em diversas áreas quando da auditoria em documentos digitais, além da entrega de Repositório, no formato JSON, planejado como um artefato tecnológico incorporado à interdisciplinariedade da Ciência da Informação.

Palavras chaves: arquitetura da informação; sistema de organização do conhecimento; ontologia; mineração de texto.

ABSTRACT

This work, whose general objective is to assist the audit process in Electronic Invoice Documents through the development of an ontology model based on Information Architecture and Text Mining to validate the information of a product's description and sale, is configured as research in the field of Information Science. Regarding the specific objectives, it was defined as: Identify possible requirements of an Information Architecture for Text Mining in Consumer Electronic Invoices and Electronic Invoices in the ontology model for the product beer; Define the main information extracted from the product when Text Mining is applied in the Invoices; Describe the relevance of knowledge organization systems, specifically, of ontology for the processes of organization and retrieval of information. It presents as theoretical basis the focus on the concepts and categories of: Knowledge Organization Systems; Information Architecture; and Ontology accompanied by applied techniques from the field of Computer Science, such as: Natural Language Processing; Machine Learning; Data Mining and Text Mining; and Metadata. The research procedure method is Structuralism; regarding its nature, it is classified as applied research; considering the purpose of the research, it fits into descriptive and exploratory research; regarding the approach to the problem, the research is presented as quantitative research; in relation to technical procedures, it is bibliographic and documentary research, whose data comes from files made available by the Department of Finance of the State of Amazonas through a sample of data in .csv file, text type, from the period of 02/01/2023 to 05/31/2023, containing transactions of Consumer Electronic Invoices and Electronic Invoices, selected from the Common Nomenclature of the Southern Common Market for the product beer. For the elaboration of the ontology, the knowledge will be a pattern or a set of patterns whose formulation may involve and relate data and information, whose logic allows the production of logical rules from inferences for the creation of models, representation of information, and extraction of knowledge. Information modeling, information flow, retrieval mapping, and the presentation of results from the perspective of business processes and user needs are some of the themes explored in this work. The results were presented in the form of answers to Competency Questions that revisit queries in DLQueries format and correctly identify the product by interpreting and composing elements in sufficient quantity and quality for its identification and use in various areas when auditing digital documents, in addition to

the delivery of a Repository, in JSON format, planned as a technological artifact incorporated into the interdisciplinarity of Information Science.

Keywords: information architecture; knowledge organization system; ontology; text mining.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

ArS	Arquitetura de Sistemas
BI	*Bussines Intelligence*
CAC	Central de Atendimento ao Contribuinte
CEET	Centro de Estudos Econômico-Tributários
CEP	Código de Endereçamento Postal
CNPJ	Cadastro Nacional de Pessoa Jurídica
CPF	Cadastro de Pessoa Física
CRF	*Conditional Random Fields*
CRISP-DM	*CRoss-Industry Standard Process for Data Mining*
DEARC	Departamentos de Arrecadação
DEARF	Departamentos de Análise e Revisão da Ação Fiscal
DEFIS	Departamentos de Fiscalização
DEINF	Departamentos de de Informações Econômico-Fiscais
DETRI	Departamentos de Tributação
DFE	Documentos Fiscais Eletrônicos
GTIN	*Global Trade Item Number*
HMMs	*Hidden Markov Models*
ICMS	Imposto sobre a Circulação de Mercadorias e Serviços
ISO	*International Organization for Standardization*
KDD	*Knowledge Discovery in Data Bases*
KOS	*Knowledge Organization System*
LATCH	*Localization, alphabet, time, category and hierarchy*
LGPD	Lei Geral de Proteção de Dados
LLM	*Large Language Models*
MIA	Arquitetura da Informação Multimodal
MAIA	Método de Arquitetura da Informação Aplicada
NEF	Núcleo de Educação Fiscal
NCM	Nomenclatura Comum do Mercosul
NED	*Named Entity Desambiguation*
NER	*Named Entity Recognition*
NFC-*e*	Nota Fiscal de Consumidor Eletrônica

NF-*e*	Nota Fiscal Eletrônica
ODS	Objetivos de Desenvolvimento Sustentável
OE	Objetivo Específico
OG	Objetivo Geral
ON-ODM	*Ontology Development Methodology*
ONU	Organização das Nações Unidas
PNL	Processamento da Linguagem Natural
PMPF	Preço Médio Ponderado ao Consumidor Final
POS	*Parts-Of-Speech*
RWR	*Relevant Words Recognition*
SEFAZ/AM	Secretaria de Fazenda do Estado do Amazonas
SER	Secretaria Executiva da Receita
SH	Sistema Harmonizado de Designação e de Codificação de Mercadorias
SINTEGRA	Sistema Integrado de Informações sobre Operações Interestaduais com Mercadorias e Serviços
SOC	Sistema de Organização do Conhecimento
SQL	*Structured Query Language*
TIC	Tecnologia da Informação e Comunicação
TBT	*Transformation-Based Tagging*
UDC	Design Centrado no Usuário
UNIF	Unidade de Inteligência Fiscal
UF	Unidade Federativa
XML	*Extended Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Os avanços tecnológicos em profusão na Era da Informação surgem caracterizados pela automatização exponencial das empresas, com a implantação de sistemas ciberfísicos, de nanotecnologia, de neurotecnologia, de inteligência artificial, de robôs, de impressão 3D e de biotecnologia. Além disso, os meios de comunicação em massa contam com interação eletrônica que implementam uma interface em linguagem natural (por texto, imagem e voz) e que colaboram com práticas de gestão moderna, eficiente, transparente, inclusiva, comprometida com a sustentabilidade e com vistas ao atendimento das prementes necessidades sociais. Assim sendo, a revolução técnica quebra paradigmas no campo do conhecimento e traz grandes impactos no trabalho dos setores da saúde, da segurança, da educação e de órgãos governamentais, em especial, alterando os mecanismos e estruturas da aquisição, organização e comunicação da informação.

Nesse contexto de uma transformação econômico-social, 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). Diante de tal mundo, há de se questionar: 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, producentes, principalmente, que sirvam para melhorar a vida e o trabalho das pessoas. Dessa forma, é preciso entender que a sociedade em rede, a sociedade da informação e da comunicação, não é passageira, e deve colaborar para evolução 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, pois 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).

As ontologias se mostram como instrumento de compartilhamento e facilitam a interoperabilidade entre sistemas. 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).

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 a lógica permite a produção de regras a partir das inferências para criação de modelos e representação da informação e da 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 Ciência da Informação conforme demonstram as pesquisas voltadas para essa temática, que foram exploradas neste trabalho.

Assim, como finalidade da Ciência da Informação são utilizados argumentos, conteúdos, ferramentas, técnicas, métodos e metodologias que se classificam como conceitos, relacionamentos e propriedades para estruturação de espaços informacionais na Arquitetura da Informação (Mori, 2009), produzido pela ontologia que, 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 Ciência da Informação são explorados em técnicas da Ciência da Computação, tendo em vista o problema da recuperação da informação, o vocabulário controlado e as técnicas automatizadas do processamento da informação. O modelo produzido é uma representação do domínio que permite 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; base para classificação e *clustering*[1]; requisito funcional para algoritmos; suporte para integração de bases diversas de informação (interoperabilidade sintática e semântica), promovendo acesso aos itens informacionais.

Considerando o cenário acima, a pesquisa se completa direcionada à construção de um modelo de ontologia abrangente, pois se trata de informações para usuários finais, apresentando 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).

Esta necessidade, no campo da auditoria, foi identificada a partir do estudo atento dos processos de controle e fiscalização da nota fiscal no Estado do Amazonas e que acompanham o fato gerador de imposto, ou seja, a arrecadação do Imposto sobre operações relativas à Circulação de Mercadorias e prestação de Serviços de Transportes Interestaduais e Intermunicipais e Comunicações (ICMS) e sua arrecadação. A prerrogativa de cobrança está sob a esfera dos Estados e Distrito

[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.

Federal e sob a responsabilidade, no caso do Amazonas, da Secretaria de Fazenda do Estado do Amazonas (SEFAZ/AM).

O entendimento da complexa legislação do ICMS passa por, de forma geral, reconhecer que os Documentos Fiscais Eletrônicos (DFE) são os principais instrumentos para acompanhar a incidência do imposto sobre a prestação de serviços de comunicação e transporte interestaduais e intermunicipais e compreender dois princípios constitucionais importantes: **princípio da não comutatividade** e o **princípio da seletividade**.

No primeiro ponto, DFE é o termo genérico para designar instrumentos que acompanham as obrigações principais e acessórias de contribuição ao tesouro. Estão incluídas nesta categoria, as notas fiscais eletrônicas adotadas pelo Estado do Amazonas desde 2009, com a obrigatoriedade de preenchimento do contribuinte para fins de interação entre a 'empresa' e o 'fisco', resultando no acompanhamento do ICMS.

No segundo ponto, os princípios constitucionais tratam da possibilidade de os Estados e Distrito Federal legislarem sobre o ICMS, definindo, por exemplo: o valor de alíquota e se um produto é mais ou menos tributado de acordo com sua essencialidade; a forma de tributação aplicada na apuração: a substituição tributária, a compensação do imposto apurado, a isenção ou não incidência do imposto na operação ou prestação, desde a origem até o seu destino. Também podem legislar sobre os demais campos relativos às políticas tributárias dos Entes Federativos que podem ser controladas por DFE e que são tratadas, em geral, nos campos lógicos dos documentos digitais Nota Fiscal Eletrônica (NF-*e*), e Nota Fiscal de Consumidor Eletrônica (NFC-*e*) por meio, principalmente, dos campos de identificação da mercadoria comercializada e do seu campo correspondente para o código estabelecido na Nomenclatura Comum do Mercosul (NCM).

Destaca-se que a identificação da mercadoria comercializada é descrita nos documentos digitais em um campo denominado "Descrição do Item", livre de formatos lógicos ou padrões textuais. Isso dificulta a prerrogativa do órgão fiscalizador na cobrança do ICMS, pois, para haver a fiscalização, deve haver, no documento digital, o registro da transação comercial e, obrigatoriamente, constar o preenchimento correto e adequado dos campos descritivos da mercadoria.

Evidencia-se, ainda, que a ausência e pouca eficiência no preenchimento e na descrição da mercadoria implica uma estagnação no processo de auditoria que

inviabilizam o seu alcance e a abrangência de diversos produtos e setores de serviços, baixando a credibilidade das ações fiscais e dificultando constantemente a atualização no tema da responsabilidade de arrecadação do imposto pelo órgão do fisco estadual. O acontecimento se dá pelo fato de os controles manuais da auditoria não acompanharem o fluxo informacional dos documentos digitais.

Outro fator diz respeito ao ICMS, que representa mais de 90% da receita total dos Estados (Frossard, 2011). Pela importância, complexidade e interferência na economia dos Estados, é o imposto que a Constituição Federal mais normatiza, entretanto, dependendo dos interesses estaduais, das peculiaridades regionais ou das circunstâncias econômicas, um produto essencial pode ser mais gravado pelo imposto que outro não essencial. Assim como o processo de controle, na forma de arrecadação, pode ser mais enfático para um produto que para outro porque traz mais dinheiro para o tesouro estadual ou porque os mecanismos, **em geral manuais**, de fiscalização são mais eficazes para um produto que para outro.

Às vésperas da implantação do “Imposto Seletivo” conhecido como “imposto do pecado”, que sobretaxa itens considerados nocivos à saúde e ao meio ambiente, é certo que as bebidas alcoólicas serão tributadas de forma diferenciada. Neste cenário, é perceptível a importância que os Estados atribuem para fins de auditoria e fiscalização do objeto deste estudo: **cerveja,** código correspondente ao **NCM: 2203.xxxx,** e o controle sobre as transações comerciais deste produto, fazendo inclusive a manutenção de legislações específicas que buscam alcançar todas as atuais marcas comercializadas no Estado.

Este esforço no Amazonas está baseado no volume financeiro da receita do ICMS gerado pela comercialização da cerveja e que entra no cofre do tesouro estadual, tanto pelo volume comercializado quanto pelo alto valor taxado na alíquota do produto. A concatenação desses dois fatores, volume comercializado e valor da alíquota, resulta, muitas vezes, em tentativas, pelo contribuinte, de ações para sonegação fiscal e desvio de receita. Casos como esses reforçam a necessidade de aprimoramento da auditoria a ser tratada por meio de requisitos da Arquitetura da Informação e a proposta de um modelo de ontologia para validar o fluxo informacional necessário à fiscalização.

Considera-se, então, que a Arquitetura da Informação, enquanto forma de organizar espaços físicos informacionais, permite uma ampla compreensão do significado da informação e da sua capacidade de comunicar, transmitir

conhecimentos e meios de usá-los. Seus requisitos, quando aplicados, ampliam a compreensão e o interesse pela informação projetada à medida que é permitida a aproximação à esta informação, pois passa pela construção de estruturas que permitem que outros a reconheçam e a compreendam.

Assim, para Wurman (1997), aplicar os requisitos da Arquitetura da Informação é importante, quando a escolha principal de como se organiza algo é feita decidindo como se quer que esse algo seja encontrado, dito de outra forma, “a organização da informação passa pela forma como ela será recuperada” (Wurman, 1997, p. 17). A proposta e construção de estruturas para a informação devem ser aplicadas desde as técnicas de recuperação de dados, quando a Mineração de Texto é ferramenta automatizada utilizada para extrair dados e torná-los acessíveis e compreensíveis à organização. Para os autores Afonso e Duque (2020):

> [...] a extração do conhecimento partindo-se de grandes quantidades de dados textuais, tem sido realizada através da aplicação de algoritmos de coleta, filtragem, mineração, análise de dados, com a meta da percepção de padrões no mar de informação (Afonso; Duque, 2020, e5325).

Continuam os autores: “Mineração de Texto consiste em extrair regularidades, padrões ou tendências de grandes volumes de texto em linguagem natural, normalmente para um objetivo específico, tal como tomada de decisão” (Afonso; Duque, 2020, e5325). Pode-se afirmar que ferramentas de mineração trabalham para recuperar informação diluída em textos livres, permitindo que seja organizada dentro da capacidade de processamento e uso do usuário final.

Em contrapartida, os requisitos de construção da informação propostos pela Arquitetura da Informação partem do pressuposto de que o objetivo da recuperação da informação seja a necessidade de uso do usuário final e, a partir de então, organiza o ambiente informacional que poderá apresentar como resultado o conhecimento específico para melhoria de processos organizacionais. Como é o entendimento dos autores Afonso e Duque (2020), “[...] extraindo a informações úteis e não evidentes para depois organizá-las de acordo com a necessidade de processamento e uso”.
A figura 01 descreve como os temas se completam em uma relação retroalimentada por intermédio de necessidade e capacidade de uso do conhecimento.

Figura 1 - Ciclo para construção do conhecimento a partir da Arquitetura da Informação e da Mineração de Texto

Fonte: Dados da pesquisa (2025).

A figura 1 demonstra o fluxo da **Arquitetura da Informação** com foco na recuperação da informação e a partir da <<necessidade de uso>> organiza os dados; o fluxo inverso caracteriza **a Mineração de Texto** com foco na organização da informação, a partir dos dados recuperados nas fases de coleta, filtragem e análise. O efeito de suas respostas intensifica e incrementa essa relação, podendo representar um ciclo infinito do ponto de vista informacional.

Por fim, os temas desta pesquisa designam o esforço à construção de um sistema informativo mais acessível, por meio de interação eletrônica, ou seja, propõe-se um modelo de ontologia para tratar as informações a partir de um ciclo de organização e de recuperação da informação com vistas ao atendimento das prementes necessidades no âmbito de uma Administração Pública eficiente, moderna e transparente.

1.1 Elementos da pesquisa

Visando esclarecer de maneira clara e precisa o problema de pesquisa identificado, os objetivos do trabalho e as justificativas para a sua elaboração, esta seção vai descrever esses contextos de modo que sirva como base para a compreensão da sequência da tese.

1.1.1 Problema

O ICMS, como principal tributo, é auditado por meio da fiscalização das notas fiscais emitidas no Estado. A NF-*e* e a NFC-*e*, são modelos nacionais de DFE com finalidade de registrar a circulação de mercadorias e serviços, 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 identificação do produto, a correspondente 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 e embalagens, além de campos de unidades e quantidades presentes nas NF-e/NFC-e não são totalmente confiáveis para uma simples avaliação. Em geral, são insuficientes para a identificação do seu conteúdo, misturados em um único campo ou ausentes na sua totalidade. 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 possui 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. Mesmo após essa investigação, caso o contribuinte informe na descrição uma nova maneira de representar o produto, o procedimento manual ou pouco inteligente não identificará a mudança rapidamente.

Especificamente, o produto cerveja, NCM 2203.xxxx, foi escolhido como modelo para esta pesquisa, a sua justificativa baseia-se em fatores como a diversidade na descrição do produto, a falta e, por vezes, ausência de padrão textual ou terminológico para descrevê-lo, concatenado ao impacto e à importância que o produto cerveja representa para a arrecadação no tesouro estadual, com o peso do “imposto do pecado”.

Como problema de pesquisa neste trabalho apresenta: de que forma o processo de auditoria em Documentos Fiscais Eletrônicos poderá ser mais produtivo com o acesso às informações sobre a descrição dos produtos das notas fiscais?

Como **hipótese**, explora a elaboração de um modelo de ontologia a partir de termos que identificam e descrevem de forma inequívoca o produto de interesse “cerveja” e suas relações mais comuns e mais importantes dispostas no campo do cupom fiscal de “descrição de produtos”. A partir dessa ontologia, espera-se caracterizar o produto como um conjunto de “atributos de qualidade” que, associados entre si e com outros campos estruturados, identificam as transações de venda em NFC-e com interesse informacional para a fiscalização e que poderão auxiliar o processo de auditoria dos Documentos Fiscais Eletrônicos, validando a descrição do produto auditado e as outras diversas variáveis da problemática das notas fiscais que serão evidenciadas.

1.1.2 Objetivo

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

Auxiliar o processo de auditoria em Documentos Fiscais Eletrônicos utilizando as Notas Fiscais de Consumidor Eletrônica (NFC-*e*) por meio da elaboração de um modelo de ontologia a partir da Arquitetura da Informação e da Mineração de Texto **para validar a informação de descrição e venda do produto.**

Os Objetivos Específicos necessários para se chegar ao objetivo são:

a) Identificar possíveis requisitos de uma Arquitetura da Informação para a Mineração de Texto em Notas Fiscais de Consumidor Eletrônicas e no modelo de ontologia para o produto cerveja (OE01);
b) Definir as principais informações extraídas de um produto quando aplicada à Mineração de Texto nas Notas Fiscais de Consumidor Eletrônica (OE02);
c) Descrever a relevância dos SOCs, especificamente, da ontologia para os processos de organização e recuperação da informação (OE03).

1.1.3 Justificativa

As justificativas estão interligadas entre si e destacam a **informação** e o processo de melhoria na **transparência da informação** como elemento fundamental para o desenvolvimento do trabalho no contexto da auditoria e da fiscalização.

Existe, à disposição das instituições públicas como a SEFAZ/AM, grande quantidade de dados sobre diversos 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 percepção do produto **e do seu significado em determinado contexto** para a auditoria nas notas fiscais. 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. Neste contexto requisitos da Arquitetura da Informação e ontologia podem servir como um modelo para incrementar qualitativamente o processo de auditoria, melhorando primeiramente a identificação do produto e, a seguir, validando a sua forma textual de descrição, permitindo a interpretação dos dados e classificando o conhecimento gerado em repositórios o que, consequentemente, amplia o volume de documentos auditados, bem como de produtos e contribuintes.

Além do aprimoramento na apresentação da informação a partir das relações existentes entre produto e atributos usando um modelo de ontologia, a pesquisa possibilita o incremento do escopo dos produtos auditados pelo setor de fiscalização porque torna acessível a informação por meio digital com vocabulário complexo e inteligível. O aumento dos produtos fiscalizados tende a propiciar o aumento da identificação de fraude e tentativas de sonegação fiscal e seus combates de forma rápida e ampla, gerando aumento da arrecadação para o tesouro estadual, boas práticas fiscais e fortalecimento da instituição.

A aplicação de temas estudados na Ciência da Informação e de ferramentas da Ciência da Computação para suporte das ações de auditoria na Administração Pública poderão estimular o desenvolvimento de políticas fiscais mais produtivas voltadas à análise de dados com a construção de padrões informacionais que melhorem a percepção do produto pela automatização das práticas de auditoria. As ações de destaque da auditoria na área de tributos estaduais da SESFAZ/AM incluem:

- Disposição de dados precisos e reais para a construção da alíquota de PMPF; para substituição tributária em produtos essenciais ao tesouro estadual;
- Definição da alíquota para tributação por Margem de Valor Agregado;

- Acompanhamento da comercialização dos produtos: Controle de Estoque (Fiscalização);
- Acompanhamento do trânsito dos produtos desde a origem até o destino: desembaraço fiscal;
- Análise fiscal para ações de verificação da regularidade fiscal, monitoramento dos setores econômicos e atualização e normatização tributária; e
- Pesquisa de preços de produtos para acompanhamento de ações tributárias da Administração Pública.

Do ponto de vista científico, a formulação de uma proposta metodológica, utilizando requisitos da Arquitetura da Informação na Mineração de Texto de termos relevantes, para construção de um modelo de ontologia como artefato para repositórios de informações de qualidade propicia:

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

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

- 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;
- Aumentar a base de consulta e a formulação do PMPF;
- Aumentar o número de contribuintes sujeitos envolvidos nos processos de fiscalização, reduzindo a desigualdade no sistema tributário;
- Incentivar uma política de padrão mais adequado à descrição do produto de acordo com a produção de descrições completas mais precisas, incluindo metadados para organização e representação da informação;
- Melhorar o uso de recursos públicos destinados à fiscalização, implementando técnicas assertivas para otimizar o uso desses recursos, maximizar os resultados e fortalecer a confiança;
- Fortalecer a Confiança Pública com transparência adequada à sociedade que busca informações de um produto específico.

Do ponto de vista social - da governança pública, é recomendado fortalecer a eficiência administrativa e transparência fiscal promovendo maior inclusão social por meio do acesso mais ágil e completo à informação. Ao encontro destes fatores estão os objetivos da Agenda 2030 para o Desenvolvimento Sustentável, adotada pela Organização das Nações Unidas (ONU), que oferece dentre suas prioridades, uma estrutura para estimular pesquisas voltadas à modernização e inovação no setor público. Dentre os 17 Objetivos de Desenvolvimento Sustentável (ODS), esta pesquisa se justifica destacando os objetivos seguintes:

- ODS 9 – Construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável e fomentar a inovação: incentiva o desenvolvimento de tecnologias resilientes e sistemas de informação modernos e sustentáveis. O objetivo prevê o desenvolvimento de infraestrutura de qualidade, confiável, sustentável e resiliente, incluindo infraestrutura regional e transfronteiriça, para apoiar o desenvolvimento econômico e o bem-estar humano;
- ODS 10 – Reduzir a desigualdade dentro dos países e entre eles: promove o acesso igualitário a recursos e oportunidades adotando políticas, especialmente fiscal, salarial e de proteção social. Melhora a regulamentação e monitoramento dos mercados assegurando mecanismos em tomadas de decisão responsiva, inclusiva, participativa e representativa em todos os níveis nas instituições econômicas, financeiras e públicas a fim de produzir instituições mais eficazes, críveis, responsáveis e legítimas;
- ODS 16 – Promover sociedades pacíficas e inclusivas para o desenvolvimento sustentável, proporcionar o acesso à justiça para todos e construir instituições eficazes, responsáveis e inclusivas: paz, justiça e instituições eficazes, responsáveis e transparentes.

Ao propor modelos com inovações tecnológicas no campo da Ciência da Informação, aplicados as necessidades e compromissos das instituições governamentais, esta pesquisa visa não apenas melhorar a auditoria com gestão da informação pública, mas também ampliar o acesso da população a serviços, direitos e informações essenciais. Isso se dá, por exemplo, pela criação sistemas de informação inclusivos e interoperáveis e mecanismos digitais que facilitem a participação social e o controle cidadão sobre as políticas públicas. Ao dialogar com as metas da Agenda 2030, esta tese reforça seu compromisso científico e social com a construção de um setor público mais inovador, inclusivo e justo, promovendo não

apenas a modernização institucional, mas também o enfrentamento das desigualdades estruturais por meio do uso estratégico da informação e da tecnologia.

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:

Seção **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 da Organização do Conhecimento (SOC); a seguir, 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 os referencias teóricos das teóricas que suportam aplicações técnicas da Ciência da Computação como a **Linguística Computacional** tratada pelas técnicas de **Processamento da Linguagem Natural (PLN)** e algoritmos de **Aprendizado de Máquina**. A seguir, apresenta tarefas de processamento da **Mineração de Dados e Mineração de Texto** incluindo, ao final, a definição e aplicação de **Metadados**.

Seção **3 Metodologia:** aborda os elementos teóricos representativos para o entendimento do processo metodológico e científico adotado e inclui detalhamento justificado sobre o contexto e o 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) processos de auditoria que envolvem notas fiscais eletrônicas; d) auditoria e aspectos teóricos da organização da informação disponível nas bases de dados; e) as

ferramentas que são atualmente utilizadas para alcançar os objetivos do tesouro estadual.

Seção 4 **Desenvolvimento do Modelo de Ontologia**: Apresenta a aplicação dos temas estudados no percurso metodológico estabelecido, de forma a construir o modelo de ontologia, extrair e validar a descrição do produto cerveja auditado.

Seção **5 Resultados**: Apresenta os resultados obtidos durante a pesquisa e a discussão sobre as relações do projeto e alternativas já discutidas na Ciência da Informação. Também demonstra como o modelo de ontologia e o repositório satisfazem o critério de qualidade do processo de auditoria discutindo algumas implicações práticas do projeto, ao final.

Por fim, a Seção **6 Considerações Finais** apresenta algumas considerações, extraindo os pontos de destaque da pesquisa incluindo propostas dos trabalhos futuros de aprimoramento e expansão do conhecimento.

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

Realizar uma pesquisa em Ciência da Informação 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 Ciência da Informação 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” 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 Ciência da Informação] 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).

Nota-se que iniciativas modernas e automáticas tratam grandes volumes de informação, 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). Uma das formas de tratar volumes de informação se dá em decorrência da Inteligência Artificial, por meio das ontologias, que se apresentam como representação do conhecimento, artefatos relativos a determinados contextos organizacionais que podem ser compartilhados, independentemente da aplicação, podendo, assim, serem utilizadas por diversos sistemas (Beira *et al.,* 2017).

Considera-se que esse termo “ontologia” teve seu início na Ciência da Informação 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 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 Ciência da Informação, por vocação, contribui para se chegar às 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 e cujas principais funções são o agrupamento de termos variantes e sinônimos em conceitos e definição de seus relacionamentos, assegurando um ordenamento lógico para que as relações sejam definidas e mantidas, tanto para a classificação quanto para a recuperação da informação.

Quando se considera o ponto de vista tecnológico, a ontologia se desenvolve a partir do trabalho especializado com o componente humano ou do aprendizado de máquina, permite a interoperabilidade das fontes de informação, 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, considerando, para isso, noções de representação, terminologias e vocabulários das ciências como Metafísica e Filosofia da Linguagem, da Lógica e Semântica, da Ciência da Computação e da Ciência da Informação em áreas como Organização da Informação e Arquitetura da Informação.

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 disciplina	Filosofia	Ontologia como sistema de categorias	Entender a realidade, as coisas que existem e suas características	Sistemas de Aristóteles, Kant, Husserl
Ontologia como artefato	Ciência da Computação	Ontologia como teoria (baseada em lógica)	Entender um domínio e reduzi-lo a modelos	BFO, DOLCE (genéricas)
		Ontologia como artefato de software	Criar um vocabulário para representação em sistemas e para gerar inferências	OWL (linguagem de RC)
	Ciência da Informação	Ontologia como teoria (informal)	Entender um domínio e classificar termos	Sistema de classificação de Ranganathan
		Ontologia como sistema conceitual informal	Criar vocabulários controlados para recuperar informação a partir de documentos	Catálogos, glossários e tesauros

Fonte: Almeida (2014, p. 252).

A partir do exposto acima, sabe-se que a Ciência da Informação, por seu caráter interdisciplinar, aplica fundamentação teórica, métodos e modelos advindos de outras áreas, numa abordagem integrada, para alcançar mapeamento de conceitos relevantes de um domínio como vocabulários controlados, esquemas de classificação, modelos mentais, interação homem-máquina etc., considerando o usuário e suas necessidades, na forma de artefatos que dialoguem com diversas áreas do conhecimento (Beira *et al*., 2017).

De tal forma, para que se alcancem os objetivos desta pesquisa, há de se enveredar por duas áreas de atuação: a Ciência da Informação e suas abordagens de SOCs, de Arquitetura da Informação e de Ontologia; e a Ciência da Computação e suas técnicas aprimoradas de PLN, Aprendizado de Máquina, Mineração de Dados e Mineração de Texto e Metadados.

2.1 Bases teóricas da Ciência da Informação

A Ciência da Informação, enquanto base conceitual para obtenção da informação e do conhecimento, precisa ser delineada visando estabelecer os campos conceituais para este estudo. Como ciência “[...] 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). Apresenta um vasto domínio no tocante a esta pesquisa e, continua o autor, permite 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).

Logo, a abordagem da Ciência da Informação no referente à importância de centrar na necessidade organizacional, de compreender o contexto e o domínio especializado e de buscar soluções organizadas e integradas, para gerar repositório de conhecimento validado e treinado, irá contribuir para a constituição das seções a seguir documentadas.

2.1.1 Ciência da Informação: conceitos e princípios

A Ciência da Informação 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.

Ademais, a Ciência da Informação, como objeto de estudo em Borko (1968), apresenta natureza prática, com preocupação de centrar no usuário o objeto de estudo para facilitar o acesso e o uso da informação. Para isso, o autor investigou o processo de representação da informação para transmissão de mensagens naturais e artificiais e técnicas de processamento em sistemas de informação para a coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização:

> [...] É uma ciência pura porque que investiga os fundamentos sem se preocupar com o campo de aplicação, **mas também uma ciência aplicada porque desenvolve produtos e serviços** [...] 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. 3, grifo nosso).

Estudos de Saracevic (1996) serão uma das bases no que tange à compreensão da área de CI para esta pesquisa. Para o autor, a Ciência da Informação diz respeito ao modo como a informação é manipulada na sociedade e como a tecnologia pode 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 Ciência da Informação 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" (Saracevic, 1996, p. 7), considerando as vantagens das modernas tecnologias informacionais.

Atualmente, técnicas de processamento de sistemas informacionais e uma variedade de algoritmos inteligentes buscam otimizar os resultados no campo da informação e do conhecimento, mais especificamente, a importância para inovação, tomada de decisão e gerenciamento de processos. No sentido organizacional, Zins (2007) destaca a multidisciplinaridade de Ciência da Informação a partir de três elementos chaves: fenômeno, domínio e escopo. Como fenômeno, o autor atribui dados, informação, conhecimento e mensagens; como domínio, o autor considera toda atividade criativa proveniente de ferramenta humana ou de máquina; e como escopo, o autor destaca o próprio campo de visão da Ciência da Informação, além de todos os outros campos de conhecimento humano.

Com o inter-relacionamento entre os três elementos chaves e o desenvolvimento de novas tecnologias, é possível incrementar o desempenho organizacional promovendo conhecimento valioso, significativo e competente, dentro da qualidade das informações que estarão disponíveis, provocando uma transformação na experiência do usuário.

Um fator de grande relevância que deve ser considerado nas pesquisas da Ciência da Informação é a interdisciplinaridade. Para que esse domínio fosse

comtemplado, Araújo (2003) é um autor que traz clareza a esse tema. Para ele, a interdisciplinaridade de Ciência da Informação 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 a 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 Ciência da Informação 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 Ciência da Informação afetaram em grande parte os processos e a necessidade de comunicação, de representação da realidade ou de 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 Ciência da Informação é o que se entende por informação. Nesse contexto, Belkin (1978) revela que definir a informação facilita a efetiva comunicação desejada entre emissor e receptor e estabelece uma visão particular do objeto (problema) da CI, demonstrando o amadurecimento 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 Ciência da Informação, 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. Ao longo do tempo, diversos conceitos de Informação buscam atender todo o campo que pertence a essa área e assim definir o que seria do escopo da Ciência da Informação. 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’”. Seguindo essa linha de pensamento, Buckland (1991) apresenta três usos da palavra informação:

> (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 (Buckland, 1991, p. 1-2).

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), expandindo o que se 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 para efeito e alteração do conhecimento no receptor; e como processo de comunicação para determinado fim.

Não passaram despercebidos, para os conceitos de informação destacados acima, os estudos de Brookes (1980) que 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 na expressa pela equação fundamental da Ciência da Informação. Tal autor reconhece nos estudos filosóficos dos três mundos de Popper (1978) a importância do conhecimento objetivo para a Ciência da Informação 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 suprindo momentaneamente a deficiência anteriormente existente e expressa por $K[S + \Delta S]$.

Portanto, é viável destacar 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). Ainda em relação à conceituação da informação, Capurro e Hjørland (2007) 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 (Capurro; Hjørland, 2007, p. 187).

Os autores consideram, então, 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 da 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).

Destaca-se que há aproximação da informação com a comunicação, pois são áreas que se interligam. Barreto (1994) analisa, a partir dos vários conceitos apresentados em estudos ao longo dos anos, que o início e o fim do processo de comunicação se dão entre emissor e receptor da mensagem, dentro da capacidade do indivíduo de comunicar o conhecimento e receber o conhecimento, percebendo-o a partir da informação e modificando a sua consciência e a sua organização de convivência:

> 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 no formato de enviar informação e recebê-la como conhecimento, não obstante suas especificações particulares, a relação mantém um caminho paralelo ao caminho de definição do conhecimento na Ciência da Informação. 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 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, transformado em informação que permite 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, p. 101).

Corrobora-se, assim, o papel da comunicação da informação como um conteúdo significativo e tratado, que permite alcançar o conhecimento como um processo social e cognitivo, obtido da transmissão de uma mensagem informacional (Pinheiro, 2003).

Com isso, é viável identificar, na Ciência da Informação, o fundamento científico e de aplicação prática, o caminho seguro para o desenvolvimento da solução desta pesquisa. Sabe-se que esta Ciência interdisciplinar investiga fenômenos de toda natureza de contexto e domínios e colabora para incrementar desempenho organizacional e a qualidade da informação aliados a tecnologias modernas da comunicação para promover constantes ofertas e buscas de conteúdos informacionais.

Nesta exposição objetiva estão também as justificativas para que esta pesquisa se destaque como elementos definidores a **informação** e **conhecimento** pautados

firmemente na necessidade do usuário, na recuperação da informação de valor mantida em um sistema organizado informacional e no processo comunicativo que altera a estrutura de conceitos do indivíduo e de seu grupo social.

A seguir, esta pesquisa apresenta um estudo sobre os Sistemas de Organização do Conhecimento para definir um modelo expressivo de conteúdo para fins de manutenção e comunicação da informação.

2.1.2 Sistemas de Organização do Conhecimento (SOC)

Para o entendimento do que sejam SOCs, faz-se necessário considerar, inicialmente, a **representação da informação** como a utilização de linguagens para a descrição física e de conteúdo de objetos informacionais, e da **representação do conhecimento** como instrumento de interpretação, feita por meio de sistemas conceituais de um determinado domínio e pelas especificações de algumas relações semânticas entre esses conceitos.

Em primeira instância, a Organização da Informação trata da análise da informação associada ao objeto, inclui a sua descrição física (suporte) e sua representação e organização de conteúdo e da linguagem em determinado domínio do conhecimento para fins de armazenamento e recuperação da informação (Brascher; Café, 2008). A representação, portanto, descreve e reproduz a informação de documentos manuais ou eletrônicos, por meio de linguagens naturais, artificiais, registros textuais, sonoros, imagens, suportes manuscritos, impressos e eletrônicos e “materializa-se por metadados relacionados ao objeto” (Victorino; Pinheiro; Santos, 2015, p. 235).

Além disso, o Processo de Organização da Informação possibilita o acesso ao conhecimento contido na informação (Brascher, Café, 2008), entretanto, o aumento de produção e exigência de disseminação de informação resultam na proliferação de objetos e nas publicações que demandam linguagens sofisticadas que, para armazenamento, recuperação e comunicação mais precisos, rápidos e fidedignos, “as linguagens servem à representação como mediadora entre a informação, na sua forma original, e o usuário na sua busca por informações de conteúdos específicos na sua área de interesse” (Frizon; Baptista, 2015, p.166).

A vantagem obtida por considerar o ato de organizar a informação a partir da aplicação de uma linguagem específica é que os constructos da linguística, tais como vocabulário, semântica e sintaxe podem ser utilizados para generalizar o entendimento e avaliar diferentes métodos de organização da informação. Outra vantagem é que esses constructos possibilitam a conceitualização que pode unificar métodos, antes díspares, de organização da informação: catalogação, classificação e indexação (Svenonius, 2000 *apud* Victorino; Pinheiro; Santos, 2015).

Por outro lado, considerando os tipos de estruturas que se ofertam aos usuários, a Organização 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, grifo nosso).

Quando estudos são realizados com base nos SOCs, 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 SOCs" (Carlan, 2010, p. 26). A partir de estudos nesses níveis, têm-se "[...] a sistematização de conhecimentos a partir de definições conceituais e suas relações representadas por signos linguísticos ou não linguísticos" (Carlan, 2010, p. 26), os critérios para selecionar os conceitos e determinar as relações semânticas entre os conceitos que diferenciam os tipos de SOCs (Hjørland, 2015).

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*. Os SOCs 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, consequentemente, facilitar a recuperação das informações contidas nos documentos (Carlan, 2010, p. 28-29).

A respeito dos SOCs vale considerar alguns pontos:

- A abrangência dos SOCs se estende a todos os instrumentos utilizados para gerenciar o conhecimento, os tipos mais comuns de sistemas usados para organizar conhecimento são os glossários, classificações simples, tipologias, folksonomias, taxonomias, tesauros, classificações facetadas, mapas de tópicos e ontologias (Gottschalg; Vilela, 2018). Em todos eles, a infraestrutura é elemento que dá o suporte ao desenvolvimento, fortemente baseado em inovações tecnológicas e a partir de uma análise das necessidades dos usuários para identificação do tipo de SOC apropriado e então, o aparelhamento de hardware do software serão adequados ao desenvolvimento, integração e manutenção do SOC.
- Apesar de distintos, para os autores Gottschalg e Vilela (2018), ambos (processos e sistemas organizacionais) buscam e cumprem a função de padronizar a representação da informação e do conhecimento, no que concerne à identificação do conteúdo do documento, em determinado domínio, auxiliando o usuário a recuperar, no texto, os elementos de dados que permitirão identificar o assunto e definir elementos conceituais, elementos constitutivos fundamentais para compreensão do seu significado.

Nos SOCs, os conceitos serão descritos a partir de termos com suas características, relações formais e estabelecimento de enunciados que expressam o comportamento restritivo de um domínio, expressando a abstração da realidade em determinado contexto. É o que que acontece com as ontologias, com alto grau de compreensão comum ao domínio, compartilhado com pessoas e máquinas e, dada sua importância para esta pesquisa, serão detalhadas na sessão a seguir.

2.1.3 Ontologias

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). Para além, fornece contexto e significado aos dados e é essencial na extração e na reutilização do conhecimento, pois 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).

Como este trabalho pretende apresentar um modelo de ontologia, é pertinente destacar a engenharia de ontologia para garantir um detalhamento do processo a ser seguido. Assim sendo, 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 o ciclo de vida de construção de uma ontologia compartilhada 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 *apud* Kors, 1997).
		Expressa os elementos do domínio com propriedades, relacionamentos, cardinalidade, valores, afirmações e pré-condições (Vizcaino *et al*., 2004).
Formalização e Implementação	Linguagem lógica e computacional	É a "identificação e representação formal e implementação computacional dos conceitos de uma determinada perspetiva" (Fox *et al*., 1993).
Validação	Generalidade	Que tarefas e soluções compartilham do modelo? Que conceitos abrange? (Fox *et al*., 1993)
	Competência	Até que ponto apoia a resolução dos problemas no domínio? (Fox *et al*., 1993)
	Eficiência	O modelo suporta um raciocínio eficiente? Fox *et al*., 1993)
	Perspicuidade	O modelo é facilmente compreendido pelos usuários? (Fox *et al*., 1993)
	Transformabilidade/ Extensibilidade	O modelo pode ser facilmente adaptado para responder aos novos problemas e tarefas? (Fox *et al*., 1993)
	Granularidade	O modelo suporta raciocínio em vários níveis e facetas? (Fox *et a*l., 1993)
	Escalabilidade	O modelo alcança respostas aos grandes projetos no domínio? (Fox *et al*., 1993)
	Integração	O modelo pode ser utilizado em conjunto com outros modelos publicado ou outros sistemas tecnológicos? (Fox *et al*., 1993)
Evolução	Manutenção	Trata das atualizações do modelo e/ou incluir novos conceitos e relações (Haridy *et al*., 2023).

Fonte: Adaptado de Grüninger; Fox (1995), Quine (1961 *apud* Kors, 1997), Vizcaino *et al*. (2004), Fox (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 e incluindo tarefas e ferramentas computacionais. 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 Aprendizado de Máquina: redes neurais, *clusters*, regras de associação, PLN, entre outros.

Quadro 3 – Metodologias para construção de ontologias

Metodologia	Desenvolvimento
***Methontology,* Fernández; Gómez Pérez; Juristo (1997)**	1.º Definição da proposta: definição do escopo, especificação dos requisitos, identificação dos recursos informacionais e ativos organizacionais que contém conhecimento específico; 2.º 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; 3.º Formalização do modelo para fins de implementação e manutenção; 4.º Avaliação: um exame técnico do modelo em relação à especificação dos requisitos; e 5.º Documentação: os artefactos criados devem documentar todas as atividades do desenvolvimento.
***101 Method* ou *Ontology Development 101,* Noy; McGuinness (2001)**	Apresenta uma prática iterativa incremental, com os seguintes passos para o desenvolvimento: 1.º Determinar domínio, escopo, requisitos e usuários; 2.º Verificar o reuso de outras ontologias já publicadas em repositórios; 3.º Determinar termos relevantes: coleta e análise de termos do domínio e agrupá-los de acordo com sua representação do domínio. 4.º Definir a hierarquia das classes; 5.º Definir as propriedades das classes: definir atributos e relações; 6.º Definir características das propriedades das classes: valores permitidos, tipos de dados etc.; 7.º Criar instâncias: identificar a que classe cada instância pertence preenchendo as propriedades da classe de forma distinta.
***On-to-Knowledge,* Sure; Studer (2003)**	Apresenta um ciclo de vida da fase inicial à fase de manutenção utilizando instrumentos de gestão e da engenharia da computação: 1.ª Fase: estudo da viabilidade da ontologia com a definição do domínio, escopo, forma de aquisição do conhecimento, seleção de ferramentas, identificação de pessoas envolvidas; 2.ª Fase: (*Kickoff*) buscar por de ontologias já publicadas para reuso, buscar fontes de conhecimento e definir requisitos, especificar questões de competência; 3.ª Fase: refinamento e reprodução da ontologia em linguagem formal; 4.ª Fase: avaliação e julgamento da ontologia e sua especificação (requisitos, questões de competência, documentação etc.
***OntoforInfoScience,* Almeida (2020)**	1.ª Fase: especificação com definição do domínio, escopo, propósito geral e requisitos em questões de competência;

	2.ª Fase: aquisição do conhecimento com ativos organizacionais, entrevistas com especialistas, análise textual de referências e extração terminológica; 3.ª Fase: fundamentação ontológica: busca por ontologias que possam servir como ponto de partida; 4.ª Fase: organização taxonômica para criar uma estrutura de classificação dos termos da ontologia e seus significados; 5.ª Fase: formalização em linguagem lógica, descrição formal do domínio; 6.ª Fase: avaliação a partir de critérios de validação e verificação.
***NeOn-Ontology,* Suárez-Figueroa; Gómez-Pérez; Fernandéz-López (2015)**	Prioriza reuso dos recursos e evolução contínua do produto: 1.º Cenário desenvolve a ontologia sem utilizar recursos ontológicos existentes, considerando as fases de: aquisição do conhecimento, especificação dos requisitos, conceitualização, formalização, implantação; 2.º Os demais cenários buscam, inicialmente e em diferentes níveis, identificar recursos ontológicos ou não que possam colaborar efetivamente por meio de reuso, reengenharia e restruturação para construção da ontologia.
ON-ODM – *Ontology Development Methodology,* Haridy *et al*., (2023)	Melhorias de qualidade na fase de conceitualização: 1.ª Aquisição dos Requisitos: identificação do domínio, análise e especificação em QC; 2.º Conceitualização: especificação dos conceitos, definição em diagrama de classes da UML, formalização em linguagem lógica OWL, implementação com Protégé; 3.º Enriquecimento por meio de PLN: 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. 4.º Avaliação: verificação baseada em QC e em métricas (precisão, coesão, compreensão, concisão); 5.º Publicação; 6.º Manutenção; 7.º Documentação.
AHIDO – *Atlas of Human Infectious Disease Ontology,* Ghozi *et al*., (2023)	Prioriza a interoperabilidade semântica dos dados. Os parágrafos longos são transformados em frases-chaves e os metadados são explorados na definição semiestruturada do *Atlas of Human*: nome, classificação, sinônimo, agentes, reservatório, vetor, transmissão, período de incubação, testes, terapia, prevenção entre outros. 1.º Passo: extração das frases-chaves do texto; 2.º Passo: recuperação dos termos e relações taxonômicas (subclasses e superclasses) ou não taxonômicas (relação literal); 3.º Passo: construção da ontologia; 4.º 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; Fernandéz-López (2015), Haridy *et al*., (2023) e Ghozi *et al*., (2023).

Ainda existem muitas possibilidades de métodos de construção em aberto quando se fala da colaboração entre Inteligência Artificial e ontologias. Em uma análise mais detalhada, **a colaboração compreende uma associação do raciocínio indutivo do Aprendizado de Máquina com o raciocínio dedutivo das ontologias resultando em uma abordagem distinta que minimiza as limitações e maximiza as características fortes de cada modelo**. O paradigma apresenta uma capacidade aprimorada de dedução, expressividade e decidibilidade das ontologias e da interpretabilidade da Inteligência Artificial e Aprendizado de Máquina (Haridy *et al.,* 2024). Isso facilita no mecanismo de conceitualizações compartilhadas, na interoperabilidade dos dados e na construção de um modelo formal e abrangente de regras.

Neste trabalho, uma adaptação da metodologia ON-ODM é proposta em um modelo híbrido de engenharia de ontologias que incorpora ferramentas e técnicas de **PLN e Aprendizado de Máquina** nas fases da Aquisição de Requisitos, Conceitualização e Enriquecimento. Além da fase de Implementação, o modelo irá propor formas simultâneas e automáticas na fase de Validação para apresentar seus resultados, todas elas **sem a presença de especialistas.**

Vale pontuar que a ontologia fornece uma compreensão de um domínio que pode ser comunicada através de sistemas heterogêneos, distribuídos ou semiestruturados para facilitar o compartilhamento e a reutilização de conhecimento entre esses sistemas (Choukri, 2014; Hamouda; Chourabi; Boughzala, 2016). De tal forma, Chang *et al.* (2020) afirmam 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.

Expandindo as afirmações acima, pode-se afirmar que a ontologia, enquanto instrumento para compartilhar o conhecimento e integrar bases informacionais com estruturas e tecnologias diferentes, possui também outras características como: “Validar o conhecimento entregue” e “Integrar domínios diferentes por meio de termos comuns”. Torna-se um apoio versátil, eficaz e seguro para dispor dos elementos necessários à satisfação e ao atendimento às necessidades do usuário e atendendo

as expectativas de proposta de ferramenta para fins identificação e compartilhamento de informação relevante.

Convém ainda sobrepesar que a relação da ontologia e Arquitetura da Informação provoca enriquecimento da estrutura semântica dos dados e facilita a organização, navegação e recuperação da informação de forma mais inteligente, precisa e adaptável ao usuário, tratando qual informação é ou não relevante. Esta visão, entretanto, não precisa ficar restrita aos dados entregues pela ontologia, mas pode ser analisada do ponto de vista dos dados recebidos pela ontologia e se transformar em requisito de construção dela, tarefa que será detalhada nos próximos capítulos desta pesquisa.

2.1.3.1 Propriedades ontológicas e reuso de componente

No contexto da modelagem ontológica, as relações semânticas entre os conceitos são fundamentais para representar o conhecimento de forma estruturada e inferível. Essas relações possibilitam não apenas a organização da informação, mas também a inferência automática e a interoperabilidade entre sistemas (Guarino, 1998; Gruber, 1993). São agrupadas em dois grandes conjuntos: propriedades dos objetos ou relacionamentos e axiomas ontológicos. Ambas as categorias contribuem para garantir consistência lógica, expressividade semântica e capacidade de reutilização da ontologia.

Com vistas ao entendimento de que as propriedades ontológicas representam como os conceitos interagem entre si ou com dados, é pertinente destacar entre as principais:

- *Is-a* (é-um): expressa uma relação de herança entre classes, sendo amplamente utilizada para estruturar hierarquias conceituais.
- *Part-of* (parte-de): utilizada para representar a composição de entidades.
- *Attributed* (atributo de): define características associadas a uma entidade.
- *Has-a* (tem-um): representa posse ou associação entre objetos.
- Funcional: limita a uma propriedade chave o valor único identificador da instância de domínio. Um exemplo poderia ser a propriedade de Cadastro de Pessoa Física (CPF) para uma pessoa, único, identificador e funcional de cada indivíduo.

Já os axiomas são regras lógicas que expressam restrições e características adicionais das classes e relações, sendo cruciais para garantir a coerência e a possibilidade de inferência automática dentro de sistemas semânticos:

- Restrição hierárquica (superclasse-subclasse): define a estrutura de herança entre conceitos, assegurando que subclasses herdem as propriedades das superclasses.
- Relacionamento ou propriedade: determina como entidades estão semanticamente conectadas por meio de propriedades ligadas ao objeto ou ao relacionamento.
- Cardinalidade: impõe limites quantitativos ao número de relações entre instâncias.
- Simetria: indica que, se a entidade A se relaciona com B, então B também se relaciona com A.
- Transitividade: garante que, se A se relaciona com B e B com C, então A se relaciona com C.
- Inversão: especifica que uma propriedade possui uma contraparte inversa, como “é pai de” ↔ “é filho de”.
- Equivalência: declara que duas classes ou propriedades são logicamente idênticas, mesmo que com nomes distintos.
- Disjunção: assegura que duas ou mais classes são mutuamente exclusivas.

Todas essas relações são centrais para o desenvolvimento de ontologias robustas e semanticamente ricas, especialmente em domínios complexos como os sistemas fiscais, e serão utilizadas durante as fases de Aquisição de Requisitos, Conceitualização e Implementação do modelo ontológico proposto neste trabalho.

Outras compreensões do domínio estudado podem partir de ontologias já construídas utilizadas na fase de Enriquecimento na engenharia de ontologias, caracterizando um recurso denominado de reuso de componentes e que favorece o aprimoramento das relações semânticas criadas em um novo modelo que se está propondo além de se tornar referência desejável pois demonstra através do reuso a flexibilidade dos conceitos e relações propostas em um componente com provável e desejável generalização do domínio, alcançando um alto grau de abstrações de desenho dos elementos capazes de serem replicados em outros estudos e modelos. Um exemplo desta capacidade de reuso de componentes foi visto nas ontologias

apresentadas no Quadro 4 e que serão tratadas como elementos de enriquecimento para o modelo de ontologia proposto para identificar produtos em documentos digitais:

Quadro 4 – Ontologias relacionadas com a pesquisa

Finalidade da ontologia	**Endereço eletrônico**	**Fonte**
Descrição de composição de cervejas	https://beer-advisor--rpi-ontology-engineering.netlify.app/oe2020/beer-advisor/usecase	Schulze *et al*. (2021)
Descrição de comércio eletrônico	https://purl.org/p2p-o	Standaert; Yaroslaski; Castro (2021)
Descrição de embalagens	https://rdf.ag/o/beer-en.html	Warren (2024)

Fonte: Dados da pesquisa (2024).

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) tratam de uma ontologia de compra e venda de produtos que acompanham 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 questões de competência 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.

Tendo em vista a necessidade de organizar e representar conhecimento, o desenvolvimento de um novo modelo ontológico que requer previamente identificar ontologias existentes no domínio de estudo e reconhecer nelas padrões de projeto

úteis, com classes e instâncias para serem reutilizados, que expande e complementa os anteriores, os recursos coletados nestas ontologias estudadas serão aplicados com as adaptações e extensões necessárias para enriquecer a conceitualização, a lista de termos padronizados e o 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. Logo, ao estruturar o conhecimento de maneira precisa, a ontologia poderá fornecer uma base sólida para aplicações como recuperação da informação e representação do conhecimento no campo da auditoria e fiscalização com interoperabilidade semântica, raciocínio automatizado e apoio à tomada de decisão.

2.1.3.2 Ambiente de aprendizado de ontologias

A incorporação do Aprendizado de Máquina no processo de construção da ontologia como *Ontology Learning* busca a extração do conhecimento novo, útil e relevante de um conjunto de dados e o tratamento da informação desde o início do processo de construção. O Aprendizado de Máquina permite utilizar, no início da construção da ontologia, algoritmos descritivos para definir conceitos, hierarquias e regras, e no final os algoritmos preditivos para predizer descrições e sentenças úteis do objeto de interesse do domínio.

Dessa forma, a *Ontology Learning* relaciona descoberta de regras, enriquecimento, aprendizagem aprimorada, proposta automática da taxonomia e de relações não taxonómicas (Hassan; Rashid, 2021). Segundo os autores, técnicas assistidas por Inteligência Artificial podem oferecer a classificação de padrões e mineração de conhecimento, descobrindo vários tipos de relações ocultas no conhecimento, como por exemplo, a relação “**termo principal minerado-termos atributos”** do objeto de interesse, relação “**termos-comportamento”** do objeto de interesse e a relação entre “**termo principal minerado-termos complementares”** não taxonômicos, mas ainda assim, dentro do domínio em que o objeto se encontra.

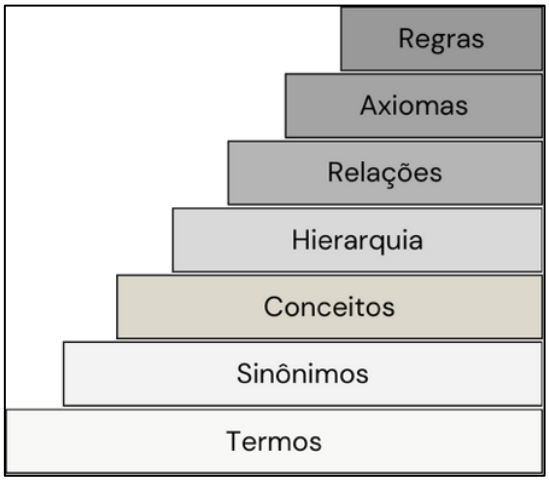
A extração e organização de conceitos e conhecimento significativo são fundamentais para a compreensão da máquina e a capacidade de raciocínio de algoritmos, de modo que a *Ontology Learning* se encarrega da extração, representação e refinamento de ontologias estruturadas que encapsulam as complexidades de vários domínios (Du *et al*., 2024). Ainda segundo os autores,

técnicas de aprendizado possibilitam descrever termos de um determinado domínio a partir da compreensão semântica e inferir relações entre entidades, tal qual a proposta descritiva e preditiva do Aprendizado de Máquina.

Sendo assim, *Ontology Learning* parte da ideia de suporte semiautomático ou automático para a construção, instanciação e evolução de uma ontologia (Hassan; Rashid, 2021) através da descoberta de conhecimento em diferentes tipos de fontes de dados brutos, (documentos de texto, imagens, dados numéricos ou mesmo outras ontologias) utilizando técnicas de mineração de dados e mineração de texto.

Para mais, o processo de *Ontology Learning* consiste em um conjunto de camadas com conceitos, relações e axiomas extraídos de texto não estruturado: mineração do termo, definição de significado, conceito e hierarquia, associação das relações, extração do conhecimento para axiomas e regras, figura 2:

Figura 2 – Modelo de *Ontology Learning Layer Cake*



Fonte: Adaptado de Buitelaar, Cimiano e Magnini (2005).

Para melhor compreensão, destaca-se cada camada em separado, visando a uma ampliação do conhecimento e a um melhor aproveitamento de cada especificação das camadas:

A camada de **Termos**, segundo os autores, seleciona termos relevantes ao domínio por definição em documentos ou entrevistas com especialistas ou por meio do processo estatístico em mineração de texto.

A camada de **Sinônimos** aborda a aquisição de variantes semânticas de termos, ou seja, termos que são semelhantes em significado, ou recíproco ao objeto de interesse.

A camada de **Conceito** parte da designação, indução ou formação da intenção de algo e descreve a realização de todas as instâncias se sobrepondo aos termos e sinônimos (Lisi, 2007). Isso envolve organizar termos relacionados em hierarquias ou categorias com base em suas similaridades, funcionalidades ou relações semânticas (Du *et al*., 2024).

A camada de **Hierarquia** de conceitos trata das relações entre palavras que têm significados mais específicos e outras com significados mais gerais (hiponímia), garantindo as futuras relações léxico-sintáticas para que as descrições do objeto de interesse em determinado domínio façam sentido e estejam bem estruturadas (Buitelaar; Cimiano; Magnini, 2005).

A etapa de **Relação** (não hierárquica) depende da Mineração de Termos e relaciona os termos essenciais com outros tantos complementares, buscando aprimorar o sentido léxico-sintático para a descrição do objeto, criando camadas de novos conceitos e hierarquias, obtidas por meio de regras de associação (Maedche; Staab, 2000).

As etapas de **Axiomas** ou Regras estão relacionadas ao trabalho de restringir as relações por meio de conhecimento especializado ou dos algoritmos de AM para extrair as regras e identificar Data Properties durante o processo de Mineração de Texto (Gorayeb; Duque, 2024). Definem dependências ou relações lógicas entre entidades ou conceitos, buscando formalizar o conhecimento do domínio e estabelecer restrições lógicas dentro da ontologia.

Em geral, a proposta do Modelo de *Ontology Learning Layer Cake* para organizar os dados utiliza Mineração de Dados e Mineração de Textos e a aplicação de técnicas de Aprendizado de Máquina como PLN (Chung; Yoo; Choe, 2020; Yang, 2020; Maedche; Staab, 2000; Haridy *et al*., 2023). Contudo, pode se expandir para muitas outras categorias de estudo para *Ontology Learning* que tratam de mapeamento de ontologias relacionadas, enriquecimento de termos da ontologia e população automática de ontologias; descoberta automática de regras da ontologia; construção automática de relações taxonômicas e não taxonômicas; e redução da granularidade dos conceitos de ontologia com aprimoramento do raciocínio, sempre com a finalidade de extrair o conhecimento para que, posteriormente, especialista e

desenvolvedor possam refinar e finalizar a construção, validando a pertinência dos dados encontrados.

Nestas categorias de estudo, existem basicamente duas formas de aprendizado para a ontologia: semiautomática e automática, com uma diversidade dos métodos de AM aplicados: raciocínio indutivo, raciocínio dedutivo e construção automática da taxonomia (Guidalia *et al.*, 2023) que sustentam a intensidade do trabalho de conceitualização, a formulação de axiomas, a escalabilidade, as ambiguidades e heterogeneidade da extração do conhecimento e a validação da ontologia (Du *et al*., 2024). Elas aceleram a extração do conhecimento em grandes bases de dados textuais e ajudam a superar dificuldades do processo de desenvolvimento. O desafio é encontrar as ferramentas de Aprendizado de Máquina adequadas, que possam ser aplicadas em grandes bases de dados textuais e que sejam capazes de encontrar o caminho para inferência de termos e relações no domínio de interesse desta pesquisa, campo de descrição do produto da NFC-*e*.

2.1.4 Arquitetura da Informação

Conceituar a Arquitetura da Informação requer, antes, posicionar seu papel e a importância desta ciência para a organização da informação. Um caminho para isto é apresentar as tarefas executadas pelo arquiteto da informação para a chamada “agregação estratégica”. Isto é, para a integração de múltiplos espaços de informação, incluindo todos os canais, modalidades e plataformas que lidam com questões relacionadas à estrutura da informação e a ferramentas de ambientes informacionais para “[...] não apenas organizar as informações, mas também simplificar para melhor compreensão com finalidade de gerir e utilizar a informação” (Wurman, 1976, p. 20).

Vale considerar que 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 caminhos para o conhecimento”,** focado na clareza e na ciência da organização da informação, em que 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” (Wurman,1997, p. 22).

Uma vez apresentada a Arquitetura da Informação sob o foco da organização da informação, pode-se compreender seu conceito pelo processo de organizar e simplificar informações; criar formas para as pessoas encontrarem e compreenderem as informações que buscam (Ding; Lin, 2010) a ciência para organizar informações desde o nível de infraestrutura até o nível da interface do usuário para que sejam localizáveis, gerenciáveis e úteis (Bailey, 2002), bem como para construir um mapa das estruturas que suportam a informação, ou seja, criar um conjunto de elementos chamados de recursos informacionais, que auxiliam no encontro da necessidade do usuário (Toms, 2002).

Pode-se destacar a Arquitetura da Informação com potencial para: "[...] encontrar problemas de projeto antecipadamente, gerenciar e aproveitar a infraestrutura de software e hardware para identificar lacunas tecnológicas e permitir o uso melhor e mais produtivo dos ativos de uma organização, incluindo informações" (Downey; Banerjee, 2010, p. 21).

Quanto ao objetivo da Arquitetura da Informação, considera-se o real potencial de moldar o contexto espacial e semântico para necessidades específicas através da criação de limites, associações, conexões e conectividade (Hinton, 2009) como um suporte informacional sob a perspectiva de gestão de arquitetura dos dados, metadados e gestão do conhecimento.

Sendo assim, a Arquitetura da Informação é a atividade de organizar a informação utilizando mecanismos tecnológicos para estruturação e categorização com a finalidade de apoiar a localização e o uso da informação e permitir compreensão e compartilhamento do conhecimento.

Para melhor entender a Arquitetura da Informação, Wurman (1997) propôs organizar informação a partir de grandes grupos representado pelos elementos: ***Location*** (localização), ***Alphabet*** (alfabeto), ***Time*** (tempo), ***Category*** (categoria) e ***Hierarchy*** (hierarquia) – (LATCH) com a finalidade de arquitetar, construir e moldar um ordenamento reconhecido como apropriado pela cognição humana.

A **localização** indica o lugar ou armazém onde os documentos com os dados necessários e desejados estão, por vezes, espalhados sem um ordenamento compreensível ao homem ou às máquinas; o **alfabeto** apresenta uma equivalência estrutural com o dicionário de dados, a manifestação organizada e sistemática de palavras construídas de símbolos de um alfabeto finito; o **tempo** indica o período inicial e final em que a pesquisa dos dados será restrita para delimitar um escopo de

busca e a avaliação dos resultados; a **categoria** manifesta a obtenção de termos e grupos de termos a partir de significados equivalentes; a **hierarquia** manifesta as relações taxonômicas ou não-taxonômicas e que podem ser tornar essenciais para a busca de mais elementos facilitadores da ontologia.

Além dos elementos indicados para organizar a informação, é importante ressaltar o papel da Arquitetura da Informação como meio de gerir e apresentar a informação de modo a torná-la útil à tomada de decisão, criando espaços para interação com o usuário: “Arquitetura da Informação cria espaços para usuários encontrarem, compreenderem e realizarem gestão da informação para tomar decisões acertadas” (Ding; Lin, 2010, p. 11, tradução nossa).

Ao refletir sobre a organização da informação, diversos autores trouxeram contribuições significativas que definiram, ampliaram o conceito e estabeleceram parâmetros para se realizar estudos neste campo do conhecimento.

Quanto à criação de 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 relacionado à natureza do conhecimento.

Siqueira (2012) é outro autor de relevância, pois explica que, em espaços ou “ambientes informacionais”, o ciclo de tratamento da informação envolve um esforço sistemático para identificação de padrões e criação de metodologias para “captura, representação, caracterização, significação, armazenamento, recuperação, comunicação, uso e descarte”, e seu desenvolvimento está ligado a Arquitetura da Informação quando se caracteriza por perceber, pensar, desenhar e habitar **espaços de informação**.

Também autores como Rosenfeld, Morville e Arango (2015) apresentam diretrizes importantes, já que consideram a formação de ambientes informacionais relacionada aos elementos **contexto**, **conteúdo** e **usuário,** relação denominada “ecologia da informação”, que pode ser alterada ao longo do tempo a partir da necessidade, da aplicação e do uso. A dependência entre eles (contexto, conteúdo e usuário) promove um fluxo de informação rico, complexo e adaptativo que atravessa a organização na forma de sistemas de informação emergente e de qualidade.

Não se pode negligenciar o papel da Tecnologia da Informação como ferramenta para Arquitetura da Informação e criação de espaços informacionais

pensando na necessidade de organização, compartilhamento e reusabilidade dos recursos representacionais na descrição, na semântica e no armazenamento de dados, pois:

> [...] permite Arquitetura da Informação 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).

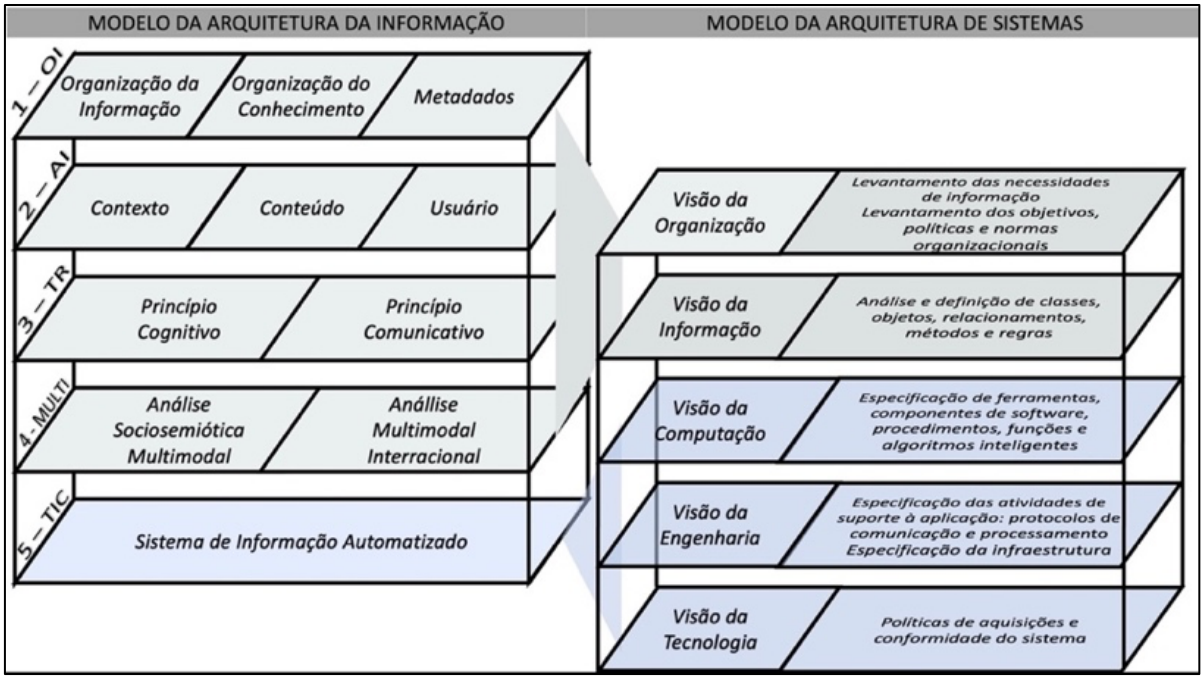
A partir dos conceitos apresentados, é possível entender que a elaboração de um artefato de natureza informacional, em que um usuário possa realizar atividades de interesse como trabalhar, se informar, se relacionar, se divertir, conviver de forma ampla requer foco na organização e na gestão da informação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

Portanto, o alcance da Arquitetura da Informação dentro da Ciência da Informação se constitui como uma disciplina dedicada a aproximar o sujeito do conteúdo, o pensamento da realidade e a necessidade à fonte informacional. A natureza abrangente dos objetivos da Arquitetura da Informação permite também que ela seja pensada e planejada no processo de desenvolvimento do sistema computacionais. Para isso, utiliza-se de modelos arquiteturais na forma de 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 contexto, conteúdo, usuário e modelos de negócios de gestão.

Dentre os inúmeros modelos já apresentados pela disciplina da Arquitetura da Informação, o Modelo de Arquitetura da Informação para sistemas automatizados apresentado por Gorayeb e Gottschalg-Duque (2022) descreve uma hierarquia composta de processos, incluindo uma etapa de Tecnologia da Informação e Comunicação (TIC) para dar suporte à infraestrutura tecnológica necessária aos

sistemas computacionais. O modelo é uma extensão do modelo original de Arquitetura de Informação apoiado pela Multimodalidade (Orlandi, 2015) apresentado em 5 etapas: 1 - Organização da Informação (associada à organização do conhecimento e metadados); 2 - Arquitetura da Informação (na percepção do contexto, conteúdo e usuário); 3 - Teoria da Relevância (a informação sob os princípios cognitivo e comunicativo); 4 - Multimodalidade (análise multimodal interacional da informação); e 5 - TIC (integração com o modelo de Arquitetura de Sistemas descrito pela norma ISO 10.746:1998). Um exemplo pode ser visualizado na figura 3:

Figura 3 – Modelo de arquitetura da informação para sistemas automatizados



Fonte: Gorayeb, Gottschalg-Duque (2022).

A finalidade de destacar o modelo arquitetural é demonstrar como a Arquitetura da Informação pode favorecer um artefato de integração entre a Ciência da Informação com a Ciência da Computação, exportando dados em formato apropriado à utilização de sistemas informacionais automatizados que apoiarão os processos de compartilhamento do conhecimento e tomada de decisão.

Outro destaque da importância dos modelos arquiteturais é tratado pelos autores Costa e Lima-Marques (2017) na evolução do modelo de representação gráfica: Método de Arquitetura da Informação Aplicada (MAIA) (Costa, 2009). Os autores destacam que os elementos Escutar, Pensar, Construir e Habitar são momentos de atuação do sujeito sobre o espaço de informação e se torna uma

ferramenta de investigação científica, gerador de ontologias de domínio. Faz-se extração de dados que caracterizam o negócio continuamente em evolução, em um conjunto de informações sistematicamente estruturadas, servindo à propósitos bem definidos.

Em resumo, a Arquitetura da Informação nesta pesquisa está no aspecto da ciência (Wurmam, 1997) como orientadora dos elementos arquiteturais LATCH que definem a organização do conteúdo e na ajuda para identificar os interesses e as necessidades informacionais dentro de um contexto (Siqueira, 2012); como princípio e procedimentos, 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 como as ontologias (Beira *et al*., 2017), reduzindo o risco de desenvolvimento desde sua estrutura até a apresentação e posterior manutenção.

2.2 Bases Teóricas da Ciência da Computação

Nesta etapa do trabalho, quer-se resguardar a importância da Ciência da Computação na construção da ontologia. Sabe-se que o termo ontologia na referida ciência pode se voltar para: (a) um vocabulário expresso em uma linguagem de Rede de Computadores (corresponde à um software, um artefato computacional), ou (b) 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

Nessa sessão, pretende-se estabelecer um cenário geral da linguística para que se adentre mais adiante nas especificações linguísticas no campo computacional com destaque para a construção do conceito e sua representação através dos termos, suas propriedades e classificações. 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 apresentam conceitos (significados) em conjunto, ou seja, inter-relacionados (Maculan; Lima, 2017).

Percebe-se como 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”.

Convém complementar que a 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 (2004a), 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, bem como 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 à realidade. É 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 em 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) 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 Gonzalez e Lima (2003, p. 7).

Conforma-se que 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 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”.

Para Novaes (2011, p. 241) “[...] 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”. Assim, o conceito aplicado a esta pesquisa seria como um conjunto de propriedades essenciais que determinam a aplicabilidade de um termo.

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.

Ademais, 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, e com isso a criação do conceito. Ainda sobre tal questão Novaes (2011) 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 caraterísticas oriundas das similaridades e diferenças existentes entre conceitos, dentro de determinado contexto (Novaes, 2011, p. 241).

Importa ponderar que notação classificatória, hierárquica; organização em categorias; relações lógicas de parte-todo e gênero-espécie são parte do processo de formação dos conceitos, haja vista que é necessária a identificação de conceitos, 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 Ciência da Informação 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).

De tal modo, a intensificação de trabalhos nestas áreas, para analisar e interpretar a capacidade cognitiva e outras atividades específicas do ser humano, resulta em estudos de formalização das línguas naturais próximas à Lógica e à Matemática em áreas como cibernética e inteligência artificial, com base no desenvolvimento de processamento automático da linguagem natural, ou seja, resulta na Linguística Computacional (Farias, 1998).

A partir dessa evolução, sistemas cada vez mais adaptados às necessidades de produção e de busca de informação, com ampliadas possibilidades de compreensão, foram criados. Tudo isso considerando que o processo de cognição passa não somente pelo contexto, mas pelo valor que a informação apresenta 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).

Observa-se, então, que a Linguística Computacional 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 na 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 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 a PLN dedica-se à 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.

Importa pontuar que o termo é uma consignação que corresponde a um conceito em uma linguagem de especialidade. Um signo linguístico que difere da palavra, unidade da língua geral, por ser qualificado no interior de um discurso de especialidade. Assim, os termos que se acomodam em determinado conceito, correspondendo a uma mesma definição, serão classificados considerando a sua equivalência e reciprocidade na mesma língua natural (ISO 25.964-1, 2011) (Barros, 2004; Lara 2004a).

Em síntese, a equivalência é estabelecida em quatro situações gerais: a) os termos são sinônimos (idênticos); b) os termos são quase-sinônimos (análogos ou similares sob determinado aspecto); c) o termo é considerado desnecessariamente específico e é representado por outro termo com escopo mais amplo; d) o termo é considerado desnecessariamente específico e é representado por uma combinação de dois ou mais termos (conhecido como “equivalência composta”) (ISO, 2011a).

A aplicação das equivalências nesta pesquisa depende inicialmente da identificação dos termos que determinam um conceito, das relações de similaridade e reciprocidade entre eles e da classificação em um elemento capaz de traduzir compreensão em linguagens informacional e computacional.

2.2.2 Metadados

Importante categoria para descrever os dados, o metadado se configura como uma metalinguagem[2]. 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

[2] 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.

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 para metadados:

> 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. 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" (Mori; Carvalho, 2004, p. 2).

Assim, o metadado é definido e qualificado no interior de um discurso, considerando sua especialidade, um termo instituído de acordo com o seu uso e emprego em contexto específico. Tanto que o metadado pode ser relacionado à documentação manual ou eletrônica de algum objeto informacional extraído por intermédio de linguagens naturais, artificiais, registros textuais, sonoros, imagens, suportes manuscritos, impressos e eletrônicos (Victorino, Pinheiro, Santos, 2015, p. 235). Esta representação pemite identificar um objeto do mundo real, suas características e relações.

Carvalho (2013) evidencia que os metadados podem servir para: criação de catálogos descritivos e operacionais, validar direitos de acesso às informações de autenticação, avaliar conteúdo, melhorar mecanismos de busca e extrair informações. Por conseguinte, os metadados podem servir para uma seleção padrão de estrutura de dados, um conjunto de termos essenciais para a consulta de produtos e serviços. De tal forma, 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 do que se precisa alcançar como um conceito estruturado do ponto de vista informacional e lógico.

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". 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 em 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 poderá ser utilizado.

Assim, na linha de registro de metadados, e considerando o objeto informacional como um **módulo de conteúdos,** Garshol (2004), Rosenfeld e Morville; Arango (2015), Victorino, Pinheiro e Santos (2015) e 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 defini-los como idênticos ou análogos (ISO, 2011a,), uma classificação apresentada na subseção 2.2.1 Linguística: base para a Ciência da Computação, e necessária para especificação, uso e compartilhamento da informação quer seja feita por um termo livre ou por uma associação de termos significativos que:

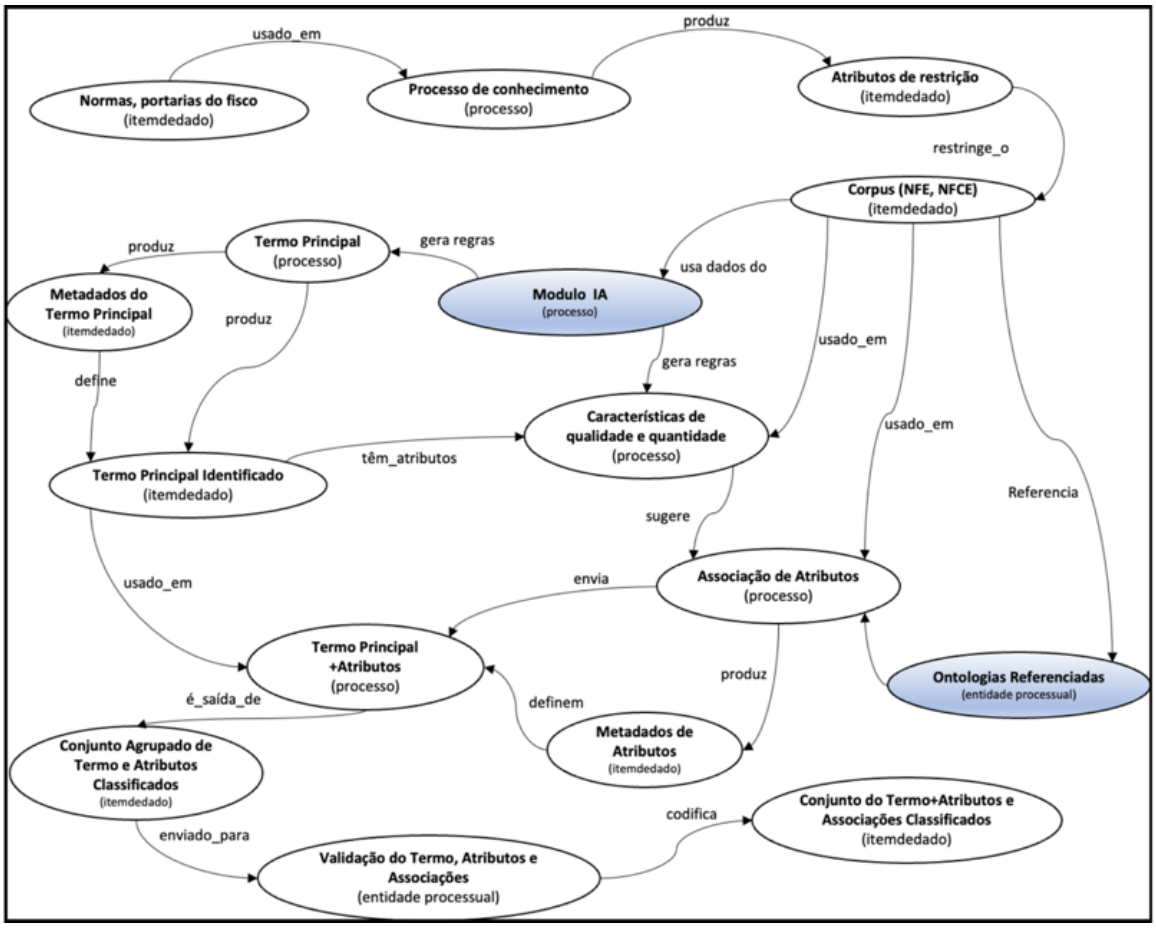
> "[...] uma vez relacionados, irão apresentar conteúdo suficiente para expressar a representação de algo que se necessita alcançar do mundo real

> na forma de um conceito estruturado do ponto de vista informacional e lógico" (Gorayeb; Duque, 2024, p. 4).

Nesta pesquisa, os metadados devem representar as características descritivas estruturais do item de interesse do usuário e receber um rótulo a partir da equivalência de termos ou de seus significados, ou seja, de características semânticas. Assim, o entendimento inferido ou deduzido do comportamento de um item de interesse define o metadado que deve ser utilizado como um conceito na ontologia, descrevendo e validando um item de interesse do domínio.

A figura 4 apresenta um planejamento para obtenção de metadados a partir de dados do tipo texto em campo não estruturado no contexto de NFC-*e* para construção de uma ontologia:

Figura 4 – Caminho para obtenção do conjunto de metadados-item de dado para ontologias



Fonte: Adaptado de Gorayeb; Gottschalg-Duque (2022).

O planejamento apresentado em Gorayeb e Gottschalg-Duque (2022) foi expandido nesta pesquisa com a inclusão das etapas **Modelo Arquitetura da**

Informação e **Ontologias Referenciadas** para extrair o conjunto dos termos, atributos e associações mais significativas em um processo semiautomatizado, para classificá-los e definir os metadados no contexto da base de dados NFC-*e* a partir de palavras que se repetem e que tem relevância para o processo de descrição do produto.

As etapas são definidas de acordo com a sua função no planejamento: **item de dado, processo ou entidade processual**. A respeito deles, pode-se entender que:

Item de dado é todo elemento que retorna uma informação categorizada (classificada) na forma de entidade ou classe.

Processo é toda ação realizada sobre um ou mais item de dado envolvendo uma lógica para interpretar dados, por exemplo um algoritmo ou uma rotina.

Entidade processual é toda ação externa que avalia um item de dado no processo.

A finalidade é obter um conjunto de termos relevantes do domínio de interesse com as características necessárias (em quantidade e qualidade), que aponte como se descreve efetivamente o objeto, tornando o processo de raciocínio automatizado mais rápido e abrangente.

Em suma, o modelo se adapta a esta pesquisa porque impulsiona a aquisição de metadados, demonstrando, em um primeiro estudo, o caminho possível para esta tarefa, ainda que sem afirmar quais as ferramentas possíveis para alcançar esta finalidade.

Algumas normas e convenções de metadados enfatizam as categorias de descritividade e padronização, com destaque para as normas ISO 23.081 que fornece diretrizes para gestão de metadados de registro e ISO 11.179 que documenta a padronização e o registro de metadados para tornar os dados compreensíveis e compartilháveis.

A norma ISO 23.081 é composta por três documentos, publicados nos anos de 2011, 2017 e 2021, e dentre as várias orientações estão os passos para identificação do formato do metadado, determinação de quais são os relevantes para apresentar um objeto informacional e a composição dos termos que descrevem os metadados. Orienta a estruturação de informações com base na identificação do formato, na determinação dos dados relevantes do objeto informacional e na composição de termos descritivos, que serão destaques nesta pesquisa.

A norma ISO 11.179 possui sete documentos que tratam sobre localizar, recuperar e transmitir dados localizados em um conjunto de dados distintos com mais facilidade. Define como os metadados são modelados conceitualmente e como são compartilhados entre as partes através do **conceito de elemento de dado**. Combina fundamentos da teoria semântica e da modelagem de dados para estruturar metadados de forma precisa e reutilizável: relação hierárquica e as equivalências semânticas entre termos diferentes para o mesmo conceito. A norma considera variações na representação de dados sem se preocupar com sua implementação física (tabelas, arquivos, colunas). O objetivo principal da ISO 11.179 é a definição e troca de metadados com significado estável, independentemente da tecnologia subjacente.

Com base nesses princípios, este trabalho irá utilizar técnicas de Mineração de Texto em textos livres e PLN para extrair automaticamente termos-chave (por exemplo: nome de entidades, valores, datas e categorias) que possam ser utilizados como metadados semânticos na representação um objeto de interesse da realidade.

2.2.3 Processamento de Linguagem Natural

PLN é um ramo da linguística que estuda a geração e a 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*[3] e utilizadas para desenvolvimento de sistemas inteligentes ou aplicação direta em massa de dados de diversos formatos (sons, imagens, palavras, sentenças, discursos etc.).

Outrossim, 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

[3] 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/

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 morfossintática e semanticamente, de modo que cada palavra é identificada separadamente e relacionamentos semânticos possíveis são identificados entre palavras e frases (Gonzalez; Lima, 2003).

Pode-se considerar que o PLN se apoia 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 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 NER é uma tarefa de PLN que consiste em 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 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).
- 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 estruturas 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 a palavras com mesmo radical a uma forma denominada stem (similar ao próprio radical), eliminado 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 à 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 $\mathbf{log}(\mathbf{f_T})\ \mathbf{x}\ \mathbf{K_t}$, 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.

Convém-se discorrer, ainda sobre a colocação ou *collocations*, que é uma sequência de palavras que aparecem juntas, utilizadas em algoritmos de análise de

termos frequentes e relações entre eles como: *if-then*[4], *is-a*[5], obtendo uma medida para a força da colocação, isto é, probabilidade de os 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).

Também, a semântica vetorial ou *vector semantics* que se apresenta como um meio algébrico que calcula o relacionamento semântico entre duas palavras. É 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).

Todas essas técnicas com uso de frequência, relação e associação de dados são relevantes na medida da necessidade de extração de metadados a partir de grandes volumes de texto e que demanda um processo sistemático capaz de identificar padrões lexicais e semânticos que representem conceitos relevantes. As técnicas combinadas partem da seleção dos termos mais representativos do corpus textual – pala com baixa frequência que representam o vocabulário mais informativo do corpus – da análise de palavras que ocorrem com muita frequência (como preposições e artigos) e que não possuem relevância informativa. A expressões também são tratadas como unidades semânticas indivisíveis, assegurando maior precisão na representação do conteúdo, indicando quais relações são fortes os suficientes para compor uma estrutura representativa da realidade e que podem representar metadados ou conceitos de dados.

---

[4] 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

[5] 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

Em conjunto, essas técnicas proporcionam uma abordagem robusta e escalável para a descoberta automática de termos e conceitos relevantes, preparando 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.

Por fim, no contexto de grandes volumes de dados textuais presentes em documentos fiscais como a NFC-e, os benefícios da PLN são: identificar de maneira mais automatizada os termos relevantes, mesmo quando descritos de forma variada, reduzindo a dependência de análise manual e promovendo padronização ao converter os textos livres em representações mais estruturadas de dados.

2.2.4 Aprendizado de Máquina

Algoritmos baseados em Aprendizado de Máquina são um ramo da Inteligência Artificial utilizados em diversas tarefas preditivas e descritivas. Tarefas 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) (Faceli *et al.,* 2021). 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).

Pondera-se que 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 Aprendizado de Máquina 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.

Fora isso, recursos não supervisionados permitem o reconhecimento e a classificação de termos, expressões linguísticas e unidades de informação relevantes no texto e são uma subtarefa 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).

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 Aprendizado de Máquina.

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:

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

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 de que padrões que ocorrem aleatoriamente não são de interesse, são eles o lift e a convicção:

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

A definição de valores para Suporte e confiança da regra de um *ItemSet* pode ser aleatória, um valor distante do ideal e que, por meio de inúmeras análises de resultados, pode ser incrementado ou decrementado até que estes e demais parâmetros, *lift* e convicção, possam ser atendidos.

Tem-se em conta, também, que em grande volume de dados e com muitas linhas de transações, é comum a utilização de valores baixos para o Suporte, variando em geral entre 0,1% e 0,5% (ou 0,001 a 0,005) para identificar padrões significativos, mesmo que menos frequentes, podem dar a certeza para eliminar aqueles padrões que, embora se apresentem na amostra, não irão interferir no comportamento dela (Han; Kamber; Pei, 2012). Isso porque, em um conjunto de milhões de registros, mesmo itens com suporte relativamente baixo podem representar milhares de ocorrências relevantes, enquanto valores muito altos de Suporte podem eliminar associações úteis, que retratam comportamentos despertos e fragmentados da amostra (Agrawal; Srikant, 1994). Suportes baixos são uma prática comum e viável computacionalmente quando há controle do número de regras geradas (Borgelt, 2005; Hahsler *et al*., 2005).

Neste trabalho, o algoritmo Apriori exerce um papel fundamental no refinamento dos dados resultantes da Mineração de Texto, ao permitir a descoberta de padrões frequentes de ocorrência entre termos ou expressões avançando na análise de extração de regras de associação que evidenciam relações sintáticas e semânticas consistentes. Essas regras indicarão a ocorrência simultânea de termos do mesmo domínio que podem, juntas, descrever um objeto de interesse como a cerveja. Esse tipo de inferência contribui significativamente para a organização conceitual do texto, favorecendo a identificação de metadados contextualmente relevantes. Além disso, o Apriori auxilia na eliminação de ruídos e associações fracas, resultando em dados mais coerentes e representativos fundamentais para aplicações como interoperabilidade entre sistemas, recuperação de informação e construção de ontologias.

2.2.5 Mineração de Dados e Mineração de Texto

Uma grande quantidade de dados organizacionais disposta em sistemas informacionais automatizados é incorporada inicialmente a partir de 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 o fluxo informacional provoca uma busca por tecnologias computacionais que dissolva a distância entre a fonte e a necessidade de informação. Por conseguinte, Técnicas de Mineração de Dados e Mineração de Texto permitem a descoberta de conhecimento em bases de dados ou *Knowledge Discovery in Databases* (KDD), isto é, possibilitam descobrir informações ocultas e associação de padrões desconhecidos ou regras de interesse em uma grande quantidade de dados. A Mineração de Dados 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 Mineração de Dados 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, e5325). 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. Envolve 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, p. 32, grifo nosso).

Assevera-se que as técnicas de Mineração de Dados 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). Tanto que as tarefas da Mineração de Dados são predição e descrição na busca por padrões de interesse e análise de dados (Fayyad; Piatetsky-Shapiro; Smyth, 1996). Portanto, 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.

Aliás, a Mineração de Dados 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:

1 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;
2 Regressão: quando se faz a descoberta de uma função que mapeie um item de dados em uma variável de predição de valor real;
3 Agrupamento: identificação de um conjunto finito de categorias (*clusters*) que descrevam os dados;
4 Sumarização: momento da busca de uma descrição compacta para um subconjunto de dados;
5 Modelagem de dependência: busca de um modelo que descreva as dependências mais significativas entre as variáveis;
6 Detecção de mudança e desvio: quando se faz a descoberta das mudanças mais significativas nos dados a partir de valores normativos ou previamente medidos.

Vale o esclarecimento de que a Mineração de Textos é considerada como variação da Mineração de Dados, realizada em documentos não estruturados ou semiestruturados com o fito de descobrir padrões e associações relevantes (Goldschimidt; Passos, 2005).

Nos estudos de Morais 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*).

Definidas 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).

Sinteticamente, 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).

Assim, a mineração associada às técnicas automatizadas capazes de processar, compreender e extrair conhecimento de grandes volumes de dados permite que os dados sejam convertidos em estruturas computacionalmente interpretáveis. Após essa conversão, os documentos podem ser representados como transações sintáticas e semânticas o que torna o conteúdo uniforme e analisável apoiando tarefas futuras como a classificação automática, criação de padrões de frequência e associação recorrentes, fundamentais para reconhecer o conhecimento diluído nos campos de textos livres das notas fiscais eletrônicas.

2.3 Bases teóricas da auditoria em Notas Fiscais Eletrônicas

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 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 (Conselho [...], 2010).

Isso posto, 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, o Fisco estadual.

Dentre outras exigências legais da NF-*e* e NFC-*e* estão: Convênio ICMS 199/10 (Conselho [...], 2010), 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 fiscal;
- Unidade Federativa (UF);
- Série e numeração sequencial única de 1 a 999.999.999 (que deve ser reinicializada quando atingido limite);
- Cadastro Nacional da Pessoa Jurídica (CNPJ)/CPF do remetente; e
- Tipo de emissão.

Isso 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 o destinatário final, garantido o registro do fato gerador do ICMS e seu futuro recolhimento como receita do tesouro estadual.

Em se tratando da NFC-*e*, seus campos de informação estão logicamente estruturados de forma a conter os seguintes dados:

- 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.

Por outro lado, a discriminação das mercadorias comercializadas com NFC-e é feita por meio do seu código correspondente, estabelecido 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).

Vale considerar que 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, que envolvem 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 onde o produto ou serviço é descrito livremente a partir do entendimento do emitente da nota fiscal. 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, Amazonas, a partir da escolha de marcas de cervejas diversas apontadas no campo “DESC” de Descrição:

Figura 5– NFC-*e* (exemplo 1)

Fonte: Acervo do autor (2023).

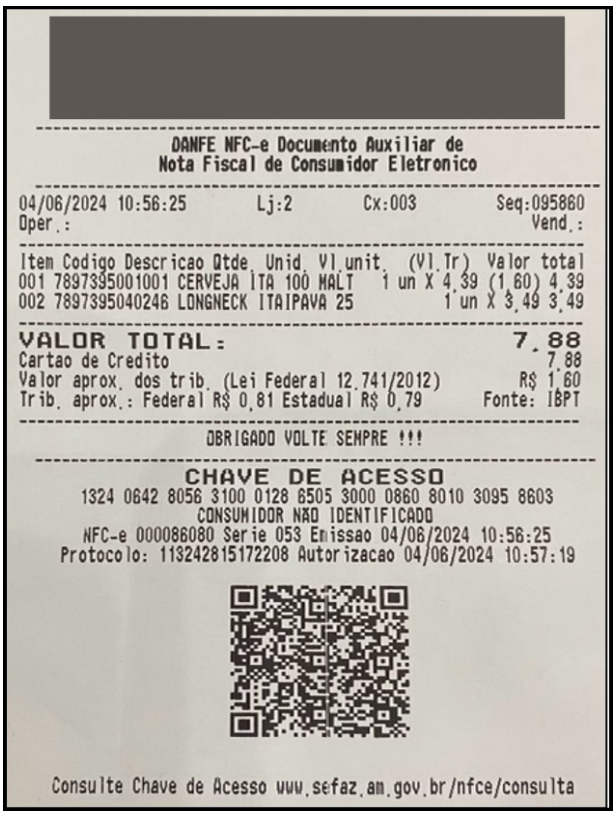
A figura 5 apresenta a compra de 3 produtos diferentes, o primeiro corresponde a “1 unidade da lata de cerveja Brahma com volume 350ML”; o segundo corresponde a “1 unidade da garrafa tipo Long Neck de cerveja Heineken com volume 330ML”; e o terceiro produto corresponde a “1 unidade da lata de cerveja 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 6 apresenta outro exemplo de NFC-*e* e a dificuldade de reconhecer a descrição do produto cerveja:

Figura 6 – NFC-*e* (exemplo 2)



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”; 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).

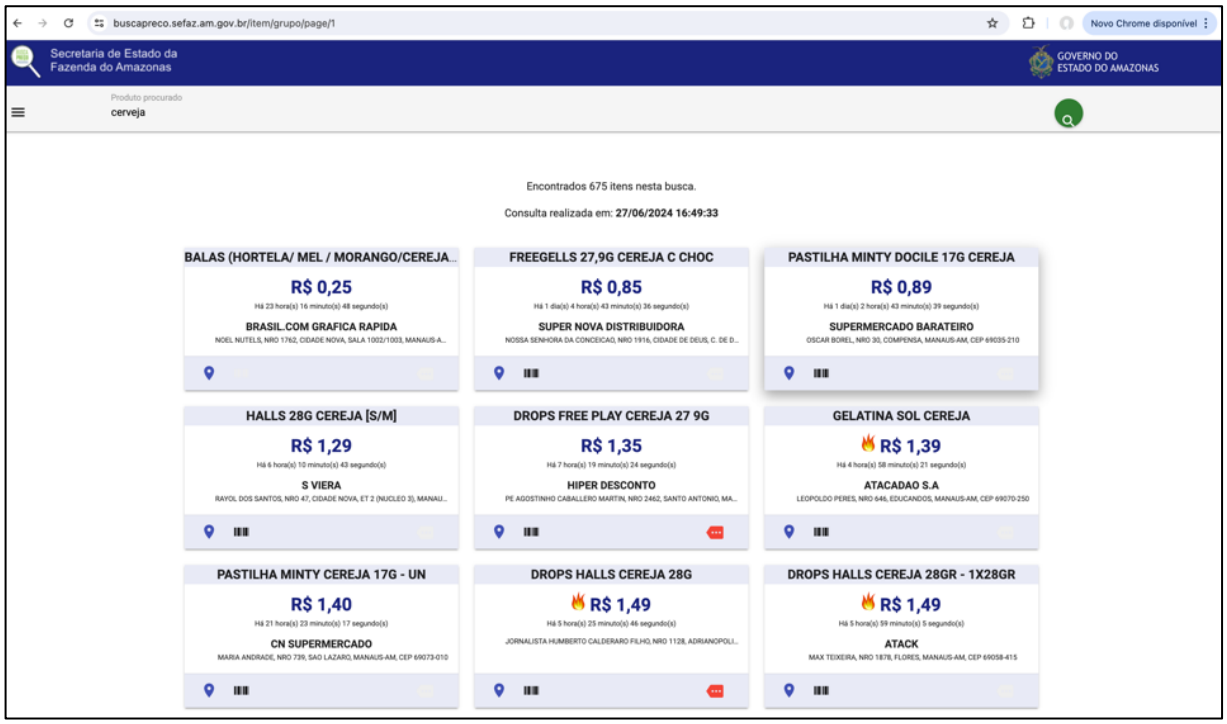
Existe à disposição da SEFAZ/AM uma grande quantidade de dados sobre produtos nas bases da NFC-e, porém, a falta de clareza na descrição, na organização da informação e de meios de recuperação confiáveis que permitam à fiscalização a percepção do produto, do seu significado em determinado contexto e uso de diferentes formas, faz com que a extração para as diversas finalidades da fiscalização seja basicamente manual quando necessário, mas somente sobre o que já é conhecido, limitando as ações de controle da arrecadação e o combate aos atos fraudulentos. Quando da 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*. Neste caso, são mais dependentes da equipe de tecnologia e que nem sempre reconhecem efetivamente o produto do qual 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, pois a SEFAZ/AM disponibiliza em sítio eletrônico[6] para consulta quando se quer 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 7 – 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 5 e 6) e de recuperar a informação (exemplo da figura 7), 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 o que for ligado às leis e às regras fiscais deve ser gerenciado e administrado pela SEFAZ/AM do ponto de vista da federação, considerando que 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/AM, há um número expressivo de funções, dentre elas, Martins (2003) destaca:

- Arrecadação de impostos: responsável por receber impostos, taxas e contribuições pagas pelos cidadãos e pelas 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 empresas e contribuintes estejam cumprindo suas obrigações tributárias e seguindo as leis fiscais. É por meio desse controle que se combate 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 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/AM é parte importante da economia do país, configurando-se como um órgão responsável por atividades inseridas nas organizações brasileiras (Martins, 2003), 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;
- 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, aos débitos automáticos ou à certidão de transferências voluntárias;
- Realização do pagamento do IPVA.

Em particular, a SEFAZ/AM dispõe de uma organização interna que contempla uma área específica para as atividades de acompanhamento fiscal e tributário denominada de Secretaria Executiva da Receita (SER), cuja finalidade é 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 N° 44.753 de 27 de outubro de 2021: "[...] I - promover a orientação normativa e a coordenação dos sistemas de arrecadação, cadastro, desembaraço de documentos, fiscalização e tributação" (Amazonas, 2011, p. 2). 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 o 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, a 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 (Instituto [...], 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, p. 10).

- Auditoria e 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 (Amazonas, 2021, p. 4).

- 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 (Amazonas, 2021, p. 4).

- 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. Existem outras ferramentas de linguagem de consulta estruturada ou *Structured Query Language* (SQL) para consultas diretas das bases de dados da NF-e e NFC-e realizadas, principalmente por técnicos da área de 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) 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, “[...] criam-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 Aprendizado de Máquina, 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; 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. Em resumo, 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 o 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 ICMS para substituição tributária; o cálculo do 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 **muitos 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, a complexidade e a existência de polissemia e a 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 heterogeneidade, ambiguidade e má qualidade dos dados 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 N*FC*-e.

3 METODOLOGIA

A Ciência da Informação apresenta uma nova demanda de cientificidade, indício das mudanças na produção e na direção do conhecimento e que 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 Ciência da Informação deve considerar seu caráter poliepistemológico, seja considerando o aspecto interdisciplinar ou multidisciplinar. Com efeito, a Ciência da Informação 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 Ciência da Informação 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 Ciência da Informação 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 Ciência da Informação, portanto, seguirá as bases metodológicas das Ciências Sociais, com combinação da Ciência da Computação.

3.1 Contextualização da pesquisa científica: configurações metodológicas

Os termos (conceitos e categorias) desta pesquisa em Ciência da Informação tem enfoque em: SOCs; Arquitetura da Informação e Ontologia. Na área da Ciência da Computação, os termos são: PLN; Aprendizado de Máquina; Mineração de Dados e Mineração de Texto 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 conseguinte 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 validez) 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 apresentam 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.

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 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 foi feito por questionário ou observação, deu-se a partir de mineração de texto, usando elementos estruturais da Arquitetura da Informação.

Também considera o caráter da **pesquisa exploratória,** quando, na fase inicial, desenvolve uma quantidade significativa de informações sobre o assunto investigado. Essa abordagem inicial facilita o delineamento do tema da pesquisa, a fixação dos objetivos e a formulação das hipóteses. Considerando o viés interdisciplinar desta pesquisa, estudou-se o tema sob diversos ângulos e aspectos (Prodanov; Freitas, 2013, p. 52), com ênfase na pesquisa bibliográfica para a exploração do tema.

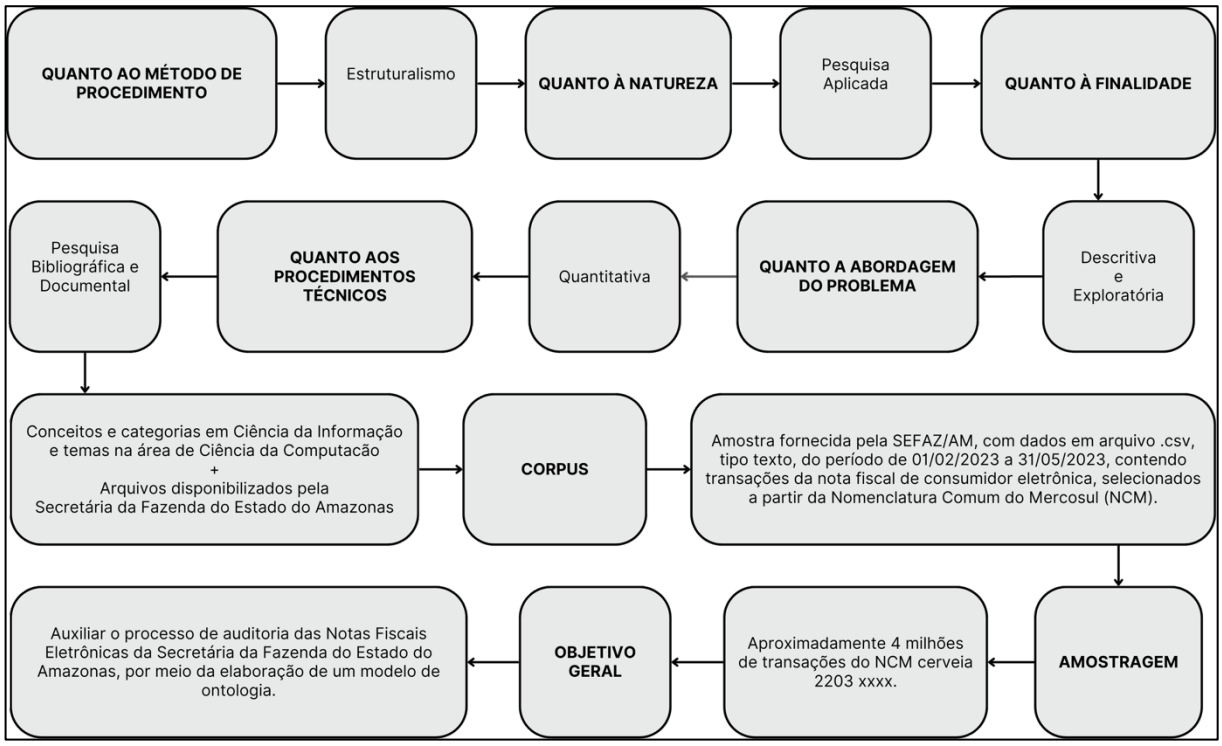
Os procedimentos técnicos, ou seja, a maneira pela qual os dados serão levantados, o que responde à questão: como os dados necessários serão obtidos? Nesta pesquisa, os procedimentos são **pesquisa bibliográfica** por fazer um levantamento teórico de estudos nas áreas desta pesquisa já realizados anteriormente 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 pode ser confundida, 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 NFC-*e*, 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 8 apresenta a trilha metodológica desta pesquisa:

Figura 8 – Percurso metodológico da pesquisa



Fonte: Dados da Pesquisa (2024).

Ao final do percurso, o modelo proposto da ontologia contribuirá para elaboração de um repositório contendo conteúdo sobre o produto escolhido na forma de informação treinada a validada por meios internos e externos ao modelo, além de código estruturado pronto para ser utilizado por outras ciências como a ciência da computação.

3.2 Configurações do procedimento metodológico da Ontologia

É importante situar a ontologia e diferenciá-la de outros tipos de SOCs. 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, tesauros e redes semânticas (Almeida; Bax, 2003).

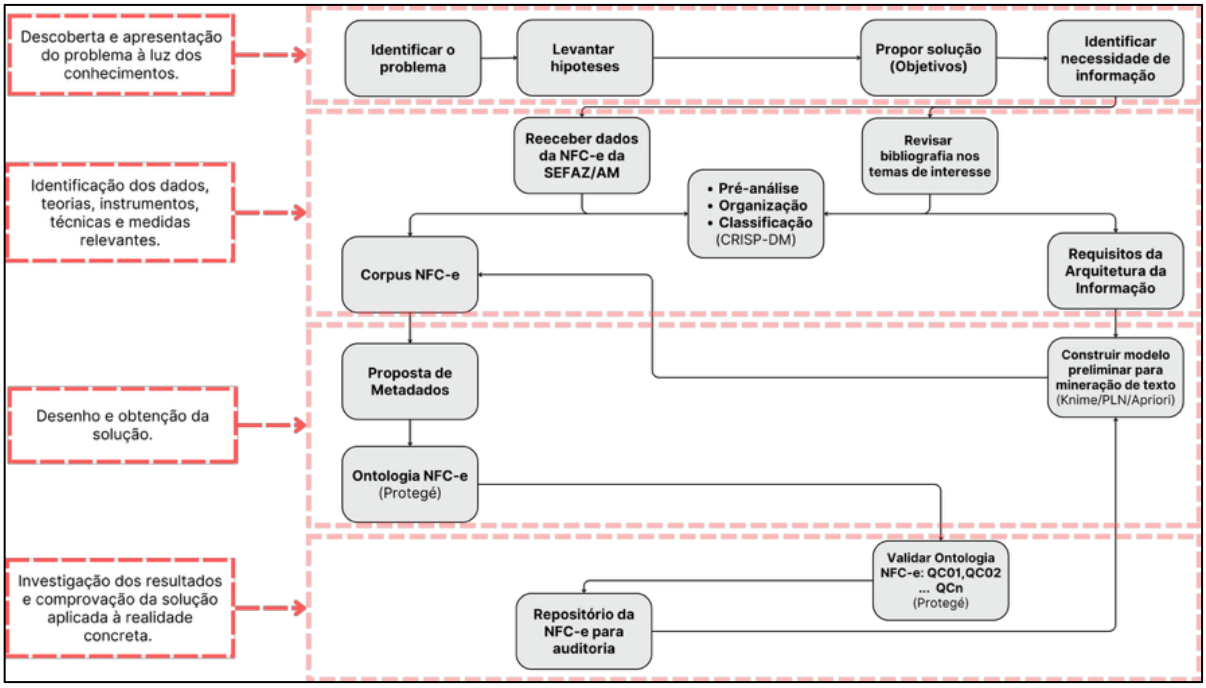
Assim, parte-se dos conceitos e daquilo que se relaciona a ele para se formular a ontologia, sendo criada por especialistas para definir as regras que regulam a combinação entre termos e relações em um domínio do conhecimento. Para fazer uso da Ontologia, os usuários usam conceitos definidos na ontologia. 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 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 Ambiente de Aprendizado de Ontologia;
- Construir proposta de requisitos da Arquitetura da Informação para aplicar com mineração de texto e alcançar a seleção de termos de interesse para o Ambiente de Aprendizado de Ontologia;
- Proposta de metadados para construção da ontologia sem especialistas;
- Construção da ontologia a partir dos resultados no Ambiente de Aprendizado de Ontologia;
- Validação por meio das Questões de Competência e outras abordagens para alcançar informação completa e útil; e
- Disposição da informação no Repositório de NFC-e para auditoria.

Para melhor visualização das etapas, a figura 9 demonstra os passos a serem seguidos:

Figura 9 – Procedimentos metodológicos 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: Filtros para segmentos mais importantes para arrecadação; Lista de produtos de substituição tributária com PMPF; Nome dos produtos que poderão enriquecer a ontologia.
Entrada dos filtros que serão aplicados na base NF-*e* e NFC-*e*.	Corpus a partir de: Número Comum do Mercosul (NCM); Indicação de termo principal e termos atributos; Períodos pré-determinados de arrecadação etc.
Definição dos metadados de descrição dos produtos por meio de MT.	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* - *Data Property Assertions* de instâncias das classes para validação dos dados.
Incorporar requisitos da AI.	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 ontologias, 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, as 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 “micromundo” (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, uma especificação explícita da conceitualização (Gruber, 1993).
> **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 que 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 10 apresenta a abordagem e a classificação da ontologia proposta nesta tese:

Figura 10 – Abordagem e classificação da ontologia

Fonte: Dados da pesquisa (2024).

A aplicação da figura 10 neste estudo descreve uma **ontologia de domínio** porque trata de um produto, especificamente cerveja de malte, no contexto da descrição do produto em um campo do cupom fiscal, com a finalidade de descrição de um produto para venda com notas fiscais eletrônicas; uma **ontologia semiformal** porque será apresentada com um pseudocódigo utilizando linguagem de fácil compreensão entendida por máquinas e humanos como OWL, Turtle e *JavaScript Object Notation* (JSON), possibilitando a interoperabilidade para integração de dados e desenvolvimento de aplicações de raciocínio automático; uma **ontologia como especificação** porque especificará os conceitos da cerveja de malte formalmente descrevendo seus significados, relações e regras; uma **ontologia de informação e modelagem do conhecimento** porque irá **modelar como o conhecimento sobre uma associação do produto cerveja de malte e a venda em notas fiscais eletrônicas está organizado** de forma compreensível e sem a presença de especialistas, utilizando o AM para alavancar a conceitualização, a inferência e o raciocínio automáticos; uma **ontologia de aplicação** porque parte de uma amostra de dados reais fornecida pela SEFAZ/AM, simulando e entregando, ao final, resultados práticos e personalizados no domínio da cerveja de malte com um repositório de informações.

Respondidos alguns questionamentos: quais procedimentos e caminhos escolhidos, 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 para a venda de cerveja de malte e denominada ontologia 2203NFCe.

4 DESENVOLVIMENTO DO MODELO DE ONTOLOGIA

Conforme apresentado no Percurso Metodológico, a construção do modelo da ontologia e a proposição de um repositório de informações para a melhoria da prática da auditoria iniciam com a revisão dos conceitos relevantes da Ciência da Informação e da Arquitetura da Informação para a pesquisa no **levantamento bibliográfico** (seção 2). A seguir, inicia uma nova etapa, com o tratamento dos dados recebidos da SEFAZ/AM, em um **levantamento operacional** para extrair informações relevantes tanto da amostra de dados (NFC-*e*) como da legislação tributária do ICMS do produto cerveja, uma etapa necessária para propor requisitos arquiteturais ao modelo de Mineração de Dados e Mineração de Texto, identificação dos **metadados** e construção da **ontologia**.

4.1 Requisitos Arquiteturais de Dados e Requisitos de Recuperação de Dados

Para o **levantamento operacional**, primeiramente, serão revisadas as legislações que regulam as aplicações de tarifas de ICMS sobre a transação comercial do produto cerveja e que impactam nas formas de auditoria em notas fiscais. Ao final da revisão, serão apontados os requisitos que impactam na recuperação da informação, na construção do conhecimento e no interesse do usuário final.

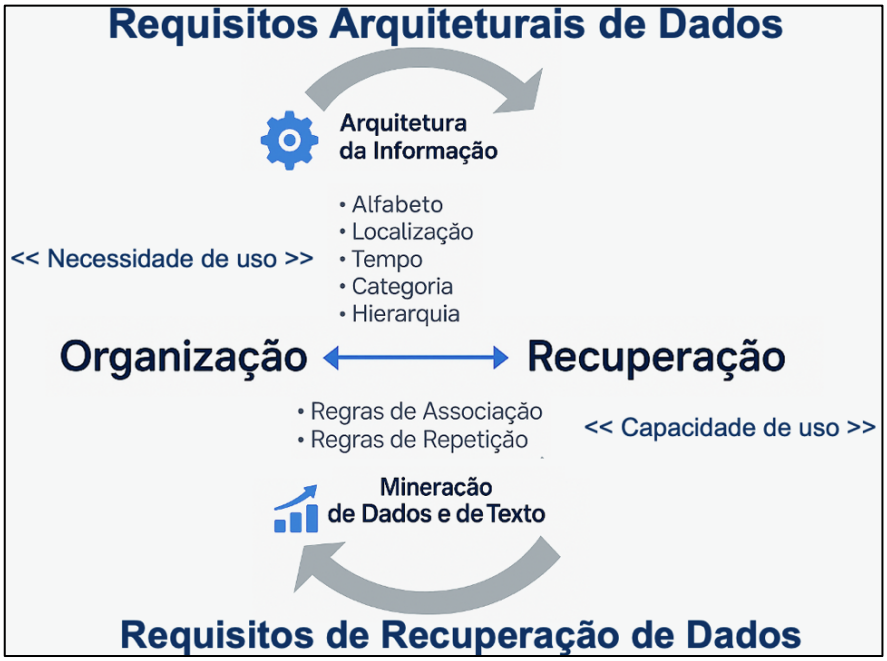
Neste sentido, a ecologia da informação (Rosenfeld; Morville; Arango, 2015) aborda as dependências entre **usuário, conteúdo e contexto** com o objetivo de compreender o negócio, os recursos disponíveis e as necessidades de uso. Esses elementos serão extraídos do levantamento operacional na forma de requisitos e, do ponto de vista da Arquitetura da Informação, chamados de “requisitos arquiteturais de dados” direcionados para <<necessidade de uso>> da informação; e, do ponto de vista da Mineração Dados e Mineração de Texto, denominados de “requisitos de recuperação de dados” direcionados para <<capacidade de uso>> da organização, figura 1 da Introdução, num ciclo contínuo para construção do conhecimento a partir da Arquitetura da Informação e Mineração de Texto e que destaca a função de recuperar a informação que a Arquitetura da Informação possui como ciência.

De início, os Requisitos Arquiteturais de Dados (figura 11) Localização, Alfabeto, Tempo, Categoria e Hierarquia, antecedem a recuperação da informação obtida com a Mineração de Texto e devem ser direcionados para conduzir a extração

de dados das bases estruturadas, semiestruturadas e não estruturadas e facilitar a compreensão rápida dos algoritmos inteligentes. No processo de mineração de texto, serão utilizados os mecanismos de organização da informação segundo Wurman (1997): LATCH. O autor destaca a função que a Arquitetura da Informação possui de **auxiliar ferramentas humanas e materiais a encontrar o que elas estão procurando**.

Além disso, os **Requisitos de Recuperação de Dados**, figura 11, **Regras de Associação e Regras de Repetição**, organizam as informações extraídas no processo de Mineração de Dados e Mineração de Texto preparando essas informações para a construção da ontologia por meio de raciocínio e aprendizado inteligente com regras estatísticas de repetição e associação no processo de treinamento denominado de "Ambiente de Aprendizado de Ontologia" (*Ontology Learning Environment*).

Figura 11 – Ciclo para construção do conhecimento a partir da Arquitetura da Informação e da Mineração de Dados e Mineração de Texto com Requisitos Arquiteturais de Dados e Requisitos de Recuperação de Dados



Fonte: Dados da pesquisa com Wurman (1997).

O processo de Organização e Recuperação, interativo e incremental (figura 11) deverá suprir as lacunas informacionais existentes, permitindo que o modelo híbrido de construção da ontologia seja alimentado constantemente na medida da sua necessidade de uso e capacidade de processamento.

4.2 Modelo híbrido para construção da ontologia 2203NFCe

A próxima etapa apresenta um paradigma híbrido entre a Engenharia de Ontologia e “Ambiente de Aprendizado de Ontologia” com Inteligência Artificial e com uma capacidade aprimorada de dedução, expressividade e decidibilidade das ontologias e da interpretabilidade da informação da Inteligência Artificial.

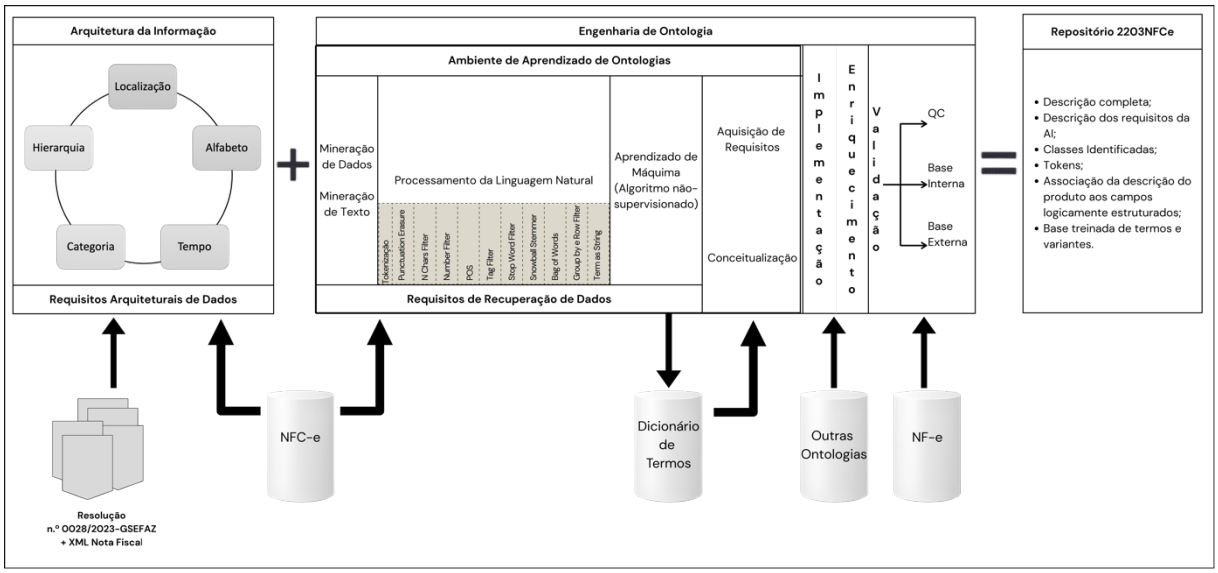
Destaca-se, de início, o aprimoramento vem da utilização de ferramentas de raciocínio indutivo da Inteligência Artificial com o raciocínio dedutivo das ontologias propostas nas metodologias estudadas, em especial da *Ontology Development Methodology* (ON-ODM) (Haridy *et al*., 2024). Uma adaptação da metodologia ON-ODM é proposta para o modelo híbrido incorporando ferramentas e técnicas de **PLN e Aprendizado de Máquina** nas fases da Aquisição de Requisitos, Conceitualização e Enriquecimento, uma referência à *Ontology Learning Environment*, além das fases de Implementação e Validação, todas elas **sem a presença de especialistas.**

Ao final, a saída do modelo apresenta as informações também na forma de um **Repositório** interoperacional com as tecnologias da ciência da computação, com informações úteis à auditoria do produto cerveja, expandindo o conhecimento de como ele é apresentado e comercializado, gerando melhorias para os diversos processos de auditoria.

Desta forma, este trabalho pretende abordar o seguinte problema de pesquisa: de que forma o processo de auditoria de Documentos Fiscais Eletrônicos poderá ser mais produtivo com o acesso às informações sobre a descrição dos produtos das notas fiscais?

Para acessar o item de descrição do produto cerveja, associá-lo aos demais campos da nota fiscal e extrair informações úteis, válidas e relevantes às áreas de auditoria, por meio de um modelo de ontologia, foi estruturado um conjunto de tarefas envolvendo Arquitetura da Informação, Mineração de Dados e Mineração de Texto, Inteligência Artificial, *Ontology Learning* e Engenharia de Ontologia, como descrito na figura 12:

Figura 12 – Proposta de construção do modelo da ontologia



Fonte: Dados da pesquisa (2025).

A saída do modelo promove uma ontologia da cerveja denominada **ontologia 2203NFCe,** cujas informações validadas serão armazenadas em um repositório no formato JSON. Será um arquivo texto amplamente utilizado para troca de informações entre sistemas por ser mais rápido de analisar, possuir arquivos mais leves e ser mais acessível à leitura que outras linguagens.

A seguir, serão detalhadas as etapas de construção do modelo da ontologia no domínio do produto: cerveja de malte.

4.3 Aplicação do modelo híbrido para ontologia do produto cerveja na NFC-*e*

Nesta seção, o modelo híbrido para construção da ontologia será aplicado na amostra de dados de NFC-*e* de venda do produto cerveja e serão definidos os Requisitos Arquiteturais de Dados e Requisitos de Recuperação de Dados a partir dos elementos LATCH da Arquitetura da Informação e do Ambiente de Aprendizado da Ontologia.

4.3.1 Levantamento operacional para definição dos Requisitos Arquiteturais de Dados

O primeiro passo é analisar os atos normativos da SEFAZ/AM para o produto cerveja de malte. Inclui a compreensão das leis e das regulamentações que abrangem

três alicerces do domínio cerveja de malte: descrição do produto, descrição da venda do produto e arrecadação do imposto ICMS. Os dados dos atos normativos do domínio NFC-e foram classificados como dados do tipo semiestruturados porque estão tabulados em estruturas flexíveis pela SEFAZ/AM:

Quadro 6 - Legislações da SEFAZ/AM para Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS) e do produto cerveja de malte

Legislação	Aplicação	Requisitos operacionais
Decreto Estadual N° 2086/1989, Artigos 110, 111ª, 233 a 236.	**APROVA** o Regulamento do Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação - ICMS e dá outras providências.	1. Definição das alíquotas de ICMS e destaque para alíquotas diferenciadas e com autorização para tributação exclusiva da cerveja de malte; 2. Descrição do produto na operação de venda compreendendo: nome, marca, tipo, modelo, série, espécie, e demais elementos, com qualidade que permita sua perfeita identificação.
LCF N° 087/1996 (Lei Kandir)	**DISPÕE** sobre o imposto dos Estados e do Distrito Federal sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual e intermunicipal e de comunicação, e dá outras providências.	1. Definição do NCM: 2203.xxxx – para cerveja de malte. 2. Identificação do *Global Trade Item Number* ou Número Global de Item Comercial (GTIN)
LCE N 019/1997	**INSTITUI** o Código Tributário do Estado do Amazonas e dá outras providências	1. Formas de auditoria para acompanhamento da tributação do ICMS: a. Pesquisa de preço para substituição tributária (PMPF); b. Controle de Estoque (Fiscalização); c. Tributação (Margem de Valor Agregado); d. Desembaraço fiscal (Fiscalização); e. Análise fiscal; e f. Preços para pesquisa de produtos para acompanhamento de ações tributárias da Administração Pública.
Convênio Estadual do ICMS N° 143/2006.	**INSTITUI** o Guia Prático da Escrituração Fiscal.	1. A discriminação do item deve indicar precisamente o mesmo, sendo vedadas discriminações diferentes para o mesmo item ou discriminações genéricas.
Convênio Estadual N° 199/2010.	**DISPÕE** sobre a emissão de documentos fiscais digitais no Estado do Amazonas.	1. Definição dos documentos fiscais digitais para o ICMS: a. NF-*e*; b. NFC-*e*.

Convênio Estadual N° 142/2018.	**DISPÕE** sobre os regimes de substituição tributária e de antecipação de recolhimento do Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviço de Transporte Interestadual e de Comunicação (ICMS) com encerramento de tributação, relativos ao imposto devido pelas operações subsequentes.	1. Definição da metodologia para tributação e apuração do ICMS: Preço Médio Ponderado a Consumidor Final (PMPF).
Resolução Estadual N° 0028/2023/ SEFAZ/AM.	**ESTABELECE** o valor do PMPF para cálculo do ICMS devido por substituição tributária nas operações com cervejas.	1. Lista as marcas de cerveja tributáveis.

Fonte: Adaptado de Amazonas (1989, 1997, 2023) Brasil (1996) e Conselho [...] (1999, 2006, 2018).

O segundo passo é analisar a amostra de dados da SEFAZ/AM. Para esta pesquisa, foram disponibilizados dois tipos de amostras, ambas as amostras em arquivo .csv, tipo texto, no período de 01/02/2023 a 31/05/2023. A primeira amostra se origina no campo “Descrição do Item” da NFC-*e*; a segunda amostra apresenta o conteúdo das linhas do campo “Descrição do Item” da NF-*e*.

Inicialmente a pesquisa foi realizada com os dados da primeira amostra recebida da SEFAZ/AM, seguindo de pré-análise, organização e classificação dos dados**.** A amostra continha três categorias de NCM: cerveja de malte, combustíveis e medicamentos, todos identificados a partir dos 4 primeiros dígitos do NCM conforme quadro abaixo:

Quadro 7 – Primeira amostra: categorias de NCM

2203.xxxx	**CERVEJA DE MALTE**
2207.xxxx	Álcool etílico não desnaturado, com um teor alcoólico, em volume, igual ou superior a 80 % vol.
2710.xxxx	Gasolina, óleo diesel e outras gasolinas.
3003.xxxx	Medicamentos constituídos por produtos misturados entre si, preparados para fins terapêuticos ou profiláticos, mas não apresentados em doses nem acondicionados para venda a retalho.
3004.xxxx	Medicamentos constituídos por produtos misturados ou não misturados, preparados para fins terapêuticos ou profiláticos, apresentados em doses (incluindo os destinados a serem administrados por via percutânea) ou acondicionados para venda a retalho

Fonte: Dados da pesquisa (2025).

As subclassificações (representadas como **xxxx** no quadro acima), correspondem aos 4 (quatro) dígitos finais do NCM e não são relevantes para fins de seleção do dado. Quando selecionado o NCM 2203.xxxx “cerveja de malte” a seleção da amostra totalizou 4.019.340 transações, isto é, 4.019.340 linhas de venda com NCM 2203.xxxx e descrição do item do produto cerveja de malte.

Quadro 8 – Detalhe da amostra de NFC-*e* recebida da SEFAZ/AM

NCM	**Descrição do NCM**	**Número de linhas de transações na amostra**	**Número de termos distintos na amostra**	**Número de descrições distintas na amostra**	**Número de palavras repetidas nas descrições**
2203.xxxx	Cerveja de malte.	4.019.340	4.585	21.539	16.781.048

Fonte: Dados da pesquisa (2025).

Importa evidenciar que o tratamento dos dados da amostra foi realizado utilizando uma adaptação prática da metodologia para Mineração de Dados denominada *CRoss-Industry Standard Process for Data Mining* (CRISP-DM), (Shearer, 2000) com pré-análise, organização, classificação e construção do Corpus, conforme abaixo descrito:

Quadro 9 – Tratamento de dados adaptado do método CRISP-DM

Etapa	**Tarefas**
Receber os dados	Coleta inicial dos dados; Compreensão dos dados; Verificação da qualidade dos dados.
Pré-análise, organização e classificação	Preparação dos dados para construir o conjunto *dataset*: Seleção dos dados: 4.019.340 linhas de venda com NCM 2203.xxxx; Limpeza dos dados; Construção de dados ausentes; Integração ou combinação de dados dispostos de tabelas, colunas, valores etc.
Construção do Corpus	Seleção de dados; Construção do conjunto *dataset; e* Avaliação final.

Fonte Adaptado de Shearer (2000).

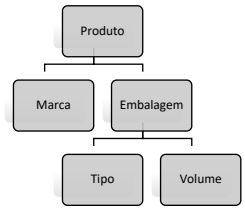
Durante a tarefa de “Avaliação final” do *dataset* foram classificados os dados do domínio de NFC-*e* como: tipos estruturados e tipos não-estruturados:

- Os tipos estruturados são aqueles que estão armazenados em tabelas ou banco de dados relacionais, em estruturas fixas, representam os campos lógicos da nota fiscal e serão facilmente recuperados pela Mineração de Dados;
- Os tipos não-estruturados não possuem organização rígida ou predefinida, como por exemplo os textos livres que descrevem o produto cerveja e serão recuperados por Mineração de Texto usando PLN.

Ao final do levantamento operacional, os Requisitos Arquiteturais dos Dados serão classificados conforme o quadro 10:

Quadro 10 - Requisitos Arquiteturais de Dados da legislação e amostra NFC-*e* do produto cerveja para construção da ontologia

#	Tipos de dados	Requisito de AI	Resultado selecionado
1	Campos lógicos e computacionalmente reconhecidos em linhas e colunas das tabelas do banco de dados da NFC-*e* para a identificação do produto:	**Localização**	NCM; cEAN (código GTIN); Número do item; Quantidade; Unidade; Valor; Descrição original (extraído do campo "descrição do produto" da nota fiscal).
2	Campos lógicos e computacionalmente reconhecidos em linhas e colunas das tabelas do banco de dados da NFC-*e* para identificação da venda:		1. Data; 2. Unidade da Federação (UF); 3. Município; 4. Emitente (Grupo C da NF-*e*); 5. CNPJ/CPF; 6. Destinatário (Grupo E da NF-*e*); 7. CNPJ/CPF; 8. Número da Nota Fiscal ou Cupom Fiscal.
3	Gramática com termos (léxico) que descrevem o produto cerveja e que são inicialmente identificados.	**Alfabeto**	Extraídos da Resolução Estadual n.° 0028/2023; Extraídos das ontologias de referência.
4	Período de interesse para auditoria das NFC-*e*.	**Tempo**	Período da amostra entregue pela SEFAZ/AM: 01/02/2023 a 31/05/2023.
5	Campos não estruturados da "descrição do produto": Classificação por equivalência ou reciprocidade de palavras relacionada aos termos que descrevem a cerveja (pela estatística).	**Categoria**	Exemplo identificado: a. CERVEJA, CERVEJ, CERV; b. LATA, LTA, LT; c. LONGNECK, LONG, LN.
6	Campos não estruturados da "descrição do produto": associações construídas entre os termos do	**Hierarquia**	Exemplo identificado: a. "CERVEJA + SKOL + LATA + 269ML"; b. "..." + "SKOL + LONGNECK + 350ML". Exemplo de Metadados:

	alfabeto (pela força das regras entre termos).		

Fonte: Dados da pesquisa (2025).

Importante considerar que, para um melhor delineamento da ontologia, fez-se necessário estabelecer alguns requisitos, destacados a seguir:

- O requisito de Localização destaca os campos estruturados de descrição do produto e descrição da venda do produto. Tem o objetivo de selecionar os campos relevantes para serem minerados no XML da nota, que irão colaborar na construção da “descrição completa” da cerveja.
- O requisito do Alfabeto aponta para todos os termos que já estão catalogados em documentos semiestruturados de qualidade e confiabilidade garantidos com o conceito e com significado assegurado.
- O requisito do Tempo evidencia o período de avaliação sobre o qual a amostra foi extraída.
- O requisito da Categoria assinala para os termos encontrados dentro do campo de descrição do produto, categorizando-os de acordo com seus conceitos ou funções.
- O requisito de Hierarquia distingue os termos já categorizados, estruturando-os em níveis de importância ou relacionando seus conteúdos.

Todos os requisitos representam os delimitadores desta pesquisa e servirão, a partir deste modelo, como os requisitos orientadores para fonte informacional organizada, estabelecendo diretrizes para entrada no ambiente de aprendizado.

4.3.2 Ambiente de Aprendizado da Ontologia para definição dos Requisitos de Recuperação de Dados

Com base nos Requisitos Arquiteturais dos Dados inicia a etapa do Ambiente de Aprendizado de Ontologia, da Engenharia de Ontologia, utilizando as técnicas computacionais de Mineração de Texto, Mineração de Dados, PLN e Aprendizado de Máquina:

a) O módulo Mineração de Texto e Mineração de Dados, concatenado ao Módulo de Aprendizado de Máquina, incluem o tratamento da informação organizada pela Arquitetura da Informação na busca de metadados para proposição de conceitos da ontologia sem o especialista. Os metadados serão extraídos do conjunto de termos e suas relações mineradas destacarão as repetições, as associações mais significativas do processo de descrição do produto e as regras para definição dos Requisitos de Recuperação de Dados. Os recursos não supervisionados com o algoritmo Apriori deve reconhecer e classificar termos, expressões linguísticas e unidades de informação relevantes no texto para:
   1 Identificar e reunir o termo e similaridades em grupos dentro de um contexto (por dedução e inferência);
   2 Estabelecer a correlação entre o número de vezes que o termo apareceu (frequência ou número de ocorrência);
   3 Estabelecer as associações entre termos e medir a dependência entre as associações: força e frequência (relevância da regra).

b) O Módulo PLN inclui o significado dos termos, classifica as descrições dos termos minerados sintática e semanticamente. Ao final, os Requisitos de Recuperação de Dados apresentarão:
   1 Lista dos termos que mais se repetem na proporção:
   - Termo que mais se repete (1.º lugar) e suas variações: **termo principal**;
   - Ranking dos termos que mais se repetem (2.º Lugar em diante): **termos atributos e termos complementares**.
   2 Lista de associações mais frequentes do tipo:
   - “**termo principal-termos atributos”** (associação preferencial e relação taxonômica);
   - **“termo atributo – termo atributo”** (associações não preferenciais);
   - “**termos-comportamento”** e;
   - “**termo principal-termos complementares”** (relação não taxonômicas).

4.3.2.1 Módulo PLN

A seguir, apresenta-se a lista de funcionalidades possíveis de PLN do campo de descrição do produto capaz de realizar tarefas automáticas que permitam desde análise até geração e interpretação da linguagem sobre o texto livre:

Quadro 11 – Funcionalidades de PNL: campos de descrição

#	ETAPA	DESCRIÇÃO
1	Tokenização	Identifica, individualmente na forma de 1 ***token***, todas as palavras que compõem a sentença que descreve o produto, transformando em um conjunto de tokens ou itens, o *itemset*. Exemplo de descrição da cerveja SKOL: “CERVEJA SKOL LATA 350ML”. Exemplo de tokenização: “CERVEJA; SKOL; LATA; 350ML”.
2	*Punctuation Erasure*	Retira todos os caracteres de pontuação.
3	*N Chars Filter*	Retira os termos com apenas 1 caractere.
4	*Number Filter*	Retira alguns caracteres numéricos.
5	*POS*	Faz a análise sintática do texto.
6	*Tag Filter*	Seleciona as palavras já classificadas pelo *POS* com exceção de determinantes, pronomes possesivos e advérbios: WDT (), WP (), WP$ (), WRB ().
7	*Stop Word Filter*	Retira palavras que não são necessárias para o significado da descrição do produto.
8	*Snowball Stemmer*	Extrai os radicais das palavras e minimiza o erro de representação da linguagem dos termos.
9	*Bag of Words*	Transforma as linhas com termos associados em linhas contendo somente 1 termo, isto é, o *itemset* representa um termo já classificado. Exemplo de descrição da cerveja SKOL com termos associados: “CERVEJA SKOL LATA 350ML” Exemplo de linhas contendo 1 termo por vez para construir o *itemset*: 1.ª linha “CERVEJA”, 2.ª Linha “SKOL”, 3.ª Linha “LATA”, 4.ª Linha “350ML”.
10	*Group by* e *ROW Filter*	Identifica e elimina termos com baixo número de frequência.
11	*Term as String*	Associa o termo à linha de transação em que ele aparece.

Fonte: Dados da pesquisa (2025).

O quadro acima descreveu a forma como cada uma das funcionalidades de PLN irão trabalhar sobre o texto livre de descrição do produto, deixando somente os termos que se associam a ele e eliminando os que provocam ruídos na descrição.

4.3.2.2 Módulo de Mineração de Texto, Mineração de Dados e Aprendizado de Máquina

Este módulo extrai conhecimento útil a partir do texto livre de descrição do produto combinando técnicas de Aprendizado de Máquina para transformar o texto em dados analisáveis e estruturáveis.

Quadro 12 - Processo de extração de termos da descrição do produto

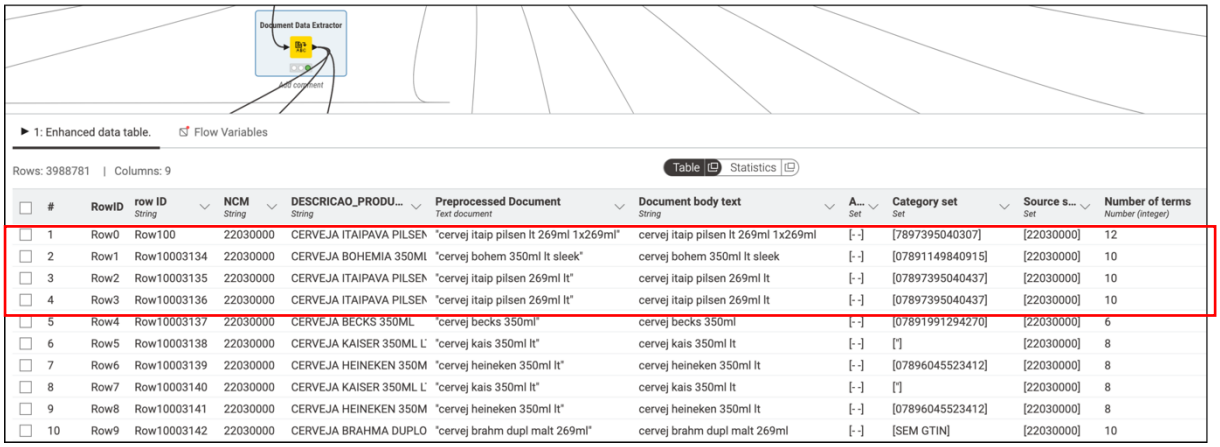
#	Descrição	Objetivo
1.ª	Extração de termos lógicos (tipos de dados estruturados) a partir das orientações dos Requisitos Arquiteturais de Dados.	Identificar com Mineração de Dados os termos lógicos que estão computacionalmente associados à venda da cerveja; Associar os termos lógicos aos metadados.
2.ª	Extração do *token* e definição do *itemset.*	Receber e identificar com Mineração de Texto o conjunto de linhas de descrição do produto cerveja. Identificar em cada linha as palavras que compõem a sentença que descreve o produto cerveja transformando-as em *Tokens.*
3.ª	Aplicação algoritmo não supervisionado Apriori para identificação dos termos (tipos de dados não estruturados).	Aplicar o suporte mínimo perto de zero, exemplo: 0,000001; Aumentar o suporte mínimo, variando (x10), descolando a vírgula para esquerda; Reconhecer o máximo de termos (e suporte) possíveis que se repetem no *itemset* de cada linha de transação de venda do produto.
4.ª	Aplicar um contador para identificar quantos termos tem em cada *itemset* (cada linha de transação de venda).	Identificar e contar quantos termos tem cada *itemset* na busca do tamanho ideal, isto é, criar um contador para somar *itemset* de 2 termos, somar *itemset* de 3 termos, somar *itemset* de 4 termos etc. Ao final, identificar a combinação de número de termos que mais se repete, coluna *Number of terms* do contador para promover o número ideal que qualifica o produto cerveja.
5.ª	Aplicação algoritmo não supervisionado Apriori para identificação dos termos.	Aplicar suporte mínimo entre $0{,}0001 < S_{min} < 0{,}001$; Identificar o conjunto de termos com suporte igual ou maior aplicado, selecionar e classificar.
6.ª	Análise das associações extraídas do Apriori para identificação dos termos antecedentes e consequentes (tipos de dados não estruturados) na melhor combinação de número de termos encontrada na etapa 4.ª	Analisar o número de regras para antecedente→consequente para os *itemset* com 2 elementos até a melhor combinação de número de termos encontrada na etapa 4.ª
7.ª	Proposta de metadados.	Classificação dos termos considerando a equivalência e a reciprocidade na mesma língua natural.

Fonte: Dados da pesquisa (2025).

No Aprendizado de Máquina utilizando algoritmos não supervisionados como Apriori, com um grande volume de dados, porém um baixo número de elementos no *itemset* (por exemplo 5 termos ou 6 termos), é recomendado que se utilize um valor de suporte baixo e variando entre: $0{,}0001 < S_{min} < 0{,}001$. Assim, neste trabalho, será aplicado o suporte mínimo de 0,0005, correspondente à média recomendada no AM. E ainda, analisando o volume das palavras repetidas na amostra (16.734.653), este suporte alcança 97,47% do total, ficando fora da análise pelo algoritmo Apriori apenas 2,528% das palavras, que, uma vez analisadas, correspondem ao ruído na descrição da cerveja. São erros como por exemplo: chocolate, água, mineral etc..

A título de exemplo, a figura 13 apresenta o contador de indicadores de termos aplicados em algumas das linhas de transação de venda. Para analisar como exemplo as linhas (#)1, 2, 3 e 4 e as palavras utilizadas para descrever a sentença de venda da cerveja, deve-se observar que nestas quatros linhas a palavra “itaip” aparece três vezes, assim como “pilsen”, já “269ml” se repete quatro vezes e “cervej” também. Embora contabilizem quatro linhas da transação de venda, juntas essas linhas apresentam vinte e uma palavras repetidas e nove termos distintos. Outra observação importante para destacar é a repetição da descrição do produto que ocorre nas linhas três e quatro:

Figura 13 – Contador dos indicadores da amostra de NFC-*e*



#	RowID	row ID	NCM	DESCRICAO_PRODU...	Preprocessed Document	Document body text	A...	Category set	Source s...	Number of terms
		String	String	String	Text document	String	Set	Set	Set	Number (integer)
1	Row0	Row100	22030000	CERVEJA ITAIPAVA PILSEN	"cervej itaip pilsen lt 269ml 1x269ml"	cervej itaip pilsen lt 269ml 1x269ml	[-]	[7897395040307]	[22030000]	12
2	Row1	Row10003134	22030000	CERVEJA BOHEMIA 350ML	"cervej bohem 350ml lt sleek"	cervej bohem 350ml lt sleek	[-]	[07891149840915]	[22030000]	10
3	Row2	Row10003135	22030000	CERVEJA ITAIPAVA PILSEN	"cervej itaip pilsen 269ml lt"	cervej itaip pilsen 269ml lt	[-]	[07897395040437]	[22030000]	10
4	Row3	Row10003136	22030000	CERVEJA ITAIPAVA PILSEN	"cervej itaip pilsen 269ml lt"	cervej itaip pilsen 269ml lt	[-]	[07897395040437]	[22030000]	10
5	Row4	Row10003137	22030000	CERVEJA BECKS 350ML	"cervej becks 350ml"	cervej becks 350ml	[-]	[07891991294270]	[22030000]	6
6	Row5	Row10003138	22030000	CERVEJA KAISER 350ML L	"cervej kais 350ml lt"	cervej kais 350ml lt	[-]	[*]	[22030000]	8
7	Row6	Row10003139	22030000	CERVEJA HEINEKEN 350M	"cervej heineken 350ml lt"	cervej heineken 350ml lt	[-]	[07896045523412]	[22030000]	8
8	Row7	Row10003140	22030000	CERVEJA KAISER 350ML L	"cervej kais 350ml lt"	cervej kais 350ml lt	[-]	[*]	[22030000]	8
9	Row8	Row10003141	22030000	CERVEJA HEINEKEN 350M	"cervej heineken 350ml lt"	cervej heineken 350ml lt	[-]	[07896045523412]	[22030000]	8
10	Row9	Row10003142	22030000	CERVEJA BRAHMA DUPLO	"cervej brahm dupl malt 269ml"	cervej brahm dupl malt 269ml	[-]	[SEM GTIN]	[22030000]	10

Fonte: Dados da pesquisa (2025).

Pode-se resumir os números das quatro primeiras linhas representados no quadro abaixo:

Quadro 13 - Indicadores do destaque da figura 13

NCM	**2203.xxxx**
Descrição do NCM	Cerveja de malte.
Número de linhas de transações na amostra	4 linhas
Número de termos distintos na amostra	9 (cervej, itaip, pilsen, lt, 269ml, 1x269ml)
Número de linhas com descrições distintas na amostra	3 (porque a linha 4 é repetição da linha 3)
Número de palavras repetidas nas descrições	21 (cervej, itaip, pilsen, lt, 269m, 1x269ml, cervej, bohem, 350ml, lt, sleek, cervej, itaip, pilsen, 269ml, lt)

Fonte: Dados da pesquisa (2025).

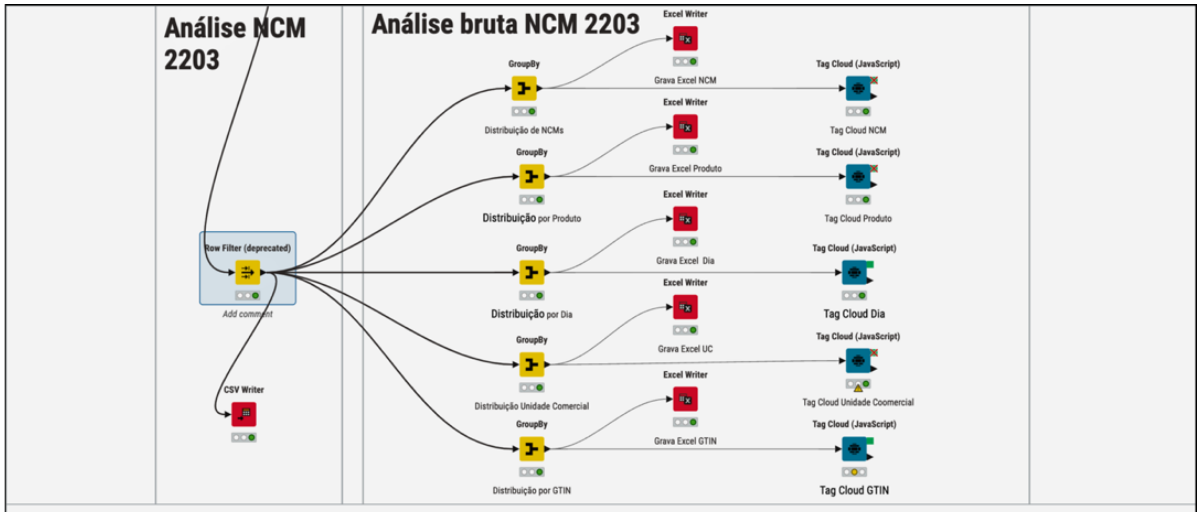
A seguir, demonstra-se a aplicação do ambiente de aprendizado da ontologia a partir dos módulos de Mineração de Texto, Mineração de Dados e Aprendizado de Máquina para a amostra da SEFAZ/AM.

De início, vale o destaque de que a ferramenta utilizada para desenvolver o Ambiente de Aprendizado de Ontologia é o software livre *Knime*[7]. Ele propõe uma análise intuitiva e confiável com uma interface visual que facilita a manipulação dos dados e apresentação dos resultados. Possui ferramentas de Inteligência Artifical com a aplicação de algoritmos de Aprendizado de Máquina de associação, agrupamento e sumarização, utilizando o algoritmo Apriori e funções de PLN.

Na sequência, foi carregada no Knime a amostra de 4.019.340 linhas de descrição do produto cerveja e iniciadas a análise e a preparação dos dados. Primeiro passo é identificar com Mineração de Dados os termos lógicos de tipos de dados estruturados e estão computacionalmente associados à venda e descrição da cerveja. Da análise bruta (figura 14), foram selecionados quais campos da nota fiscal estão disponíveis nas bases de dados e podem ser recuperados e modelados para ontologia: NCM, GTIM, DATA DE EMISSÃO, UNIDADE, QUANTIDADE e VALOR TOTAL DO ITEM:

[7] https://www.knime.com/.

Figura 14 – Análise bruta da amostra da SEFAZ/AM



▶ 1: Filtered Flow Variables

Rows: 4019340 | Columns: 7 Table Statistics

#	RowID	NCM (String)	GTIN (String)	DESCRICAO_PRODUTO (String)	DATA_EMISS... (Local Date)	UNIDADE (String)	QUANTIDADE (Number (double))	VL_TOTAL_ITEM (Number (double))
1	Row15	22030000	7897395040307	CERVEJA - ITAIPAVA LATA 269ML	2023-04-01	UND	156	468
2	Row16	22030000	7900007909708	CX CERVEJA ITAIPAVA C 12 UND [0800]	2023-04-01	UN	1	31
3	Row17	22030000	7891149103270	CERVEJA - BRAHMA LATA 269ML	2023-04-01	UND	3	9
4	Row18	22030000	7898915949162	CERVEJA IMPERIO GOLD LN 210ML - 1X210ML	2023-04-01	UN	4	18.6
5	Row19	22030000	7898904771606	CERVEJA DEVASSA LAGER LT 350ML - 1X350ML	2023-04-01	UN	8	23.12
6	Row20	22030000	78936683	CERVEJA HEINEKEN LN 330ML - 1X330ML	2023-04-01	UN	24	138.96
7	Row21	22030000	78936683	CERVEJA HEINEKEN LN 330ML - 1X330ML	2023-04-01	UN	48	277.92
8	Row22	22030000	78936683	CERVEJA HEINEKEN LN 330ML - 1X330ML	2023-04-01	UN	6	34.74
9	Row23	22030000	7891991297424	CERVEJA SPATEN SLEEK LT 350ML - 1X350ML	2023-04-01	UN	24	88.56
10	Row24	22030000	78936683	CERVEJA HEINEKEN LN 330ML - 1X330ML	2023-04-01	UN	24	138.96
11	Row25	22030000	7897395000592	CERVEJA ITAIPAVA 100%PURO MALTE LT 269ML - 1X269ML	2023-04-01	UN	12	25.08
12	Row26	22030000	7897395040307	CERVEJA ITAIPAVA PILSEN LT 269ML - 1X269ML	2023-04-01	UN	240	549.6

Fonte: Dados da pesquisa (2025).

O campo NCM, pela sua importância no domínio de NFC-*e,* foi classificado como um metadado para ser implementado na ontologia como uma classe NCM. Os valores encontrados no campo NCM se tornaram instâncias da classe NCM: 2203.0000; 2203.0022; 2203.0099; 22030300; 2203.1000; 2203.3000 e 2203.9000. Para representar os dados GTIM, DATA DE EMISSÃO, UNIDADE, QUANTIDADE e VALOR TOTAL DO ITEM foram propostos metadados do tipo: NFCe e NFCeItem que representam a NF-*e* e o item da nota fiscal.

O primeiro NFCe guarda a informação das notas, ou do cabeçalho da nota fiscal por exemplo: informações do emissor, informações do destinatário, chave da nota etc. (já descriminadas na seção 2) e será transformado em classe na ontologia. O segundo NFCeItem também será transformado em classe na ontologia, passa a representar o campo descrição do produto com todas as informações da venda e a informação da descrição da cerveja. Assim, GTIM, DATA DE EMISSÃO, UNIDADE, QUANTIDADE

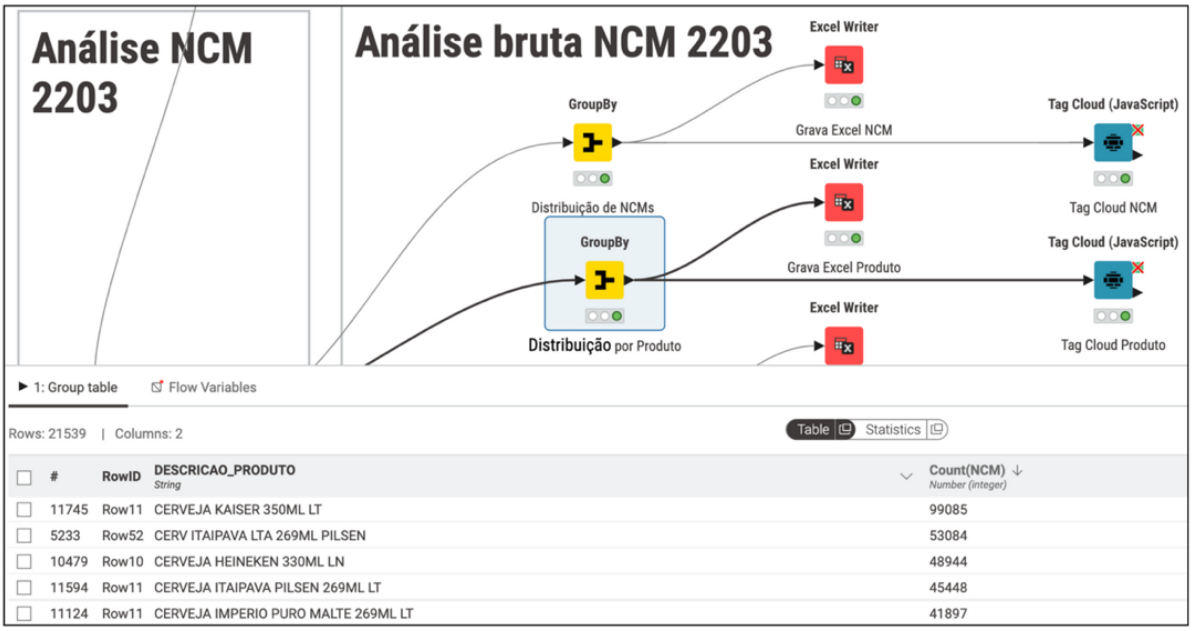
e VALOR TOTAL DO ITEM passam a ser atributos da classe NFCeItem, ou seja, *Data Properties* de NFCeItem.

A análise bruta também evidencia a necessidade de manter relações entre essas novas classes. NCM passa a ter uma relação de $1:m$ entre NFCeItem (garantido que cada item da nota fiscal possui 1 NCM, pois cada item tem um produto da nota fiscal); e NFCe tem uma relação de $1:m$ com NFCeItem (garantido que uma única nota fiscal possa ter vários itens de produto discriminados).

A seguir foram aplicadas as técnicas de PLN:

a) Eliminação das colunas de dados vazias;
b) Adequação de tipo de dado *number* (NCM) para *string* (char de 8 posições), similar ao formado da base de dados original;
c) Adequação de tipo de dado *string* (data de emissão da NFC-*e*) para formato *date&time*, similar ao formado da base de dados original;
d) Sumarização dos dados para entendimento por número de repetições da venda x descrição do produto, como demonstrado na figura 15:

Figura 15 – Exemplo do resultado preliminar de 5 produtos com descrição igual na amostra



#	RowID	DESCRICAO_PRODUTO (String)	Count(NCM) (Number (integer))
11745	Row11	CERVEJA KAISER 350ML LT	99085
5233	Row52	CERV ITAIPAVA LTA 269ML PILSEN	53084
10479	Row10	CERVEJA HEINEKEN 330ML LN	48944
11594	Row11	CERVEJA ITAIPAVA PILSEN 269ML LT	45448
11124	Row11	CERVEJA IMPERIO PURO MALTE 269ML LT	41897

Fonte: Dados da pesquisa (2025).

A imagem representa o número de vezes que uma descrição do produto cerveja aparece da mesma forma na amostra recebida da SEFAZ/AM. O contador apresenta

um total geral de linhas distintas de 21.539 descrições diferentes para esta amostra. A primeira linha de baixo para cima (#11124) apresenta a descrição do produto: “CERVEJA IMPERIO PURO MALTE 269ML LT”, totalizando 41.897 linhas repetidas, ou seja, iguais, do total da amostra. Já a primeira linha de cima para baixo (#11745) apresenta a descrição do produto “CERVEJA KAISER 350ML LT” com o total de 99.085 repetições dentro da amostra.

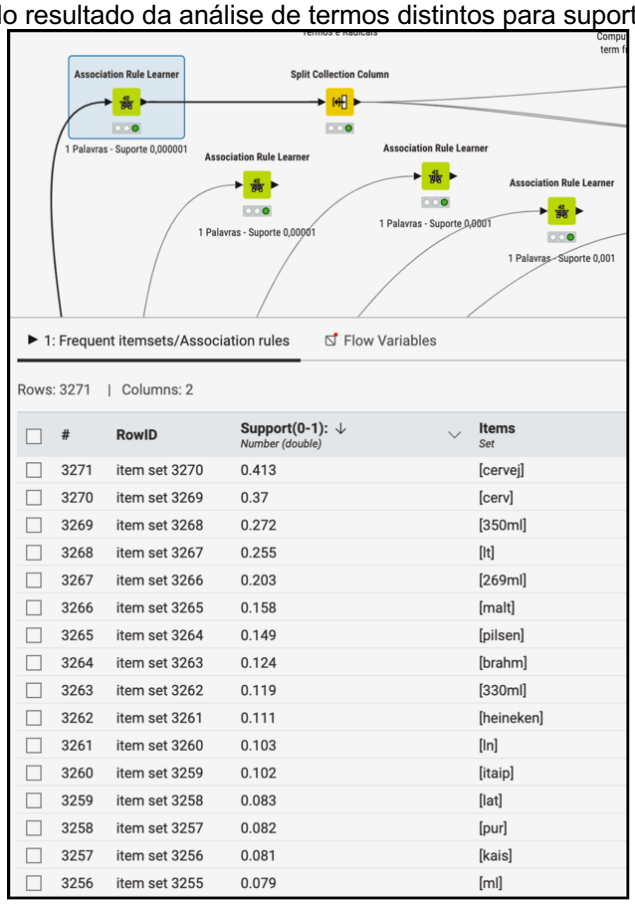
A seguir, foram realizadas as seguintes tarefas complementares no módulo PLN:

a) Remoção dos espaços no campo de descrição do produto;
b) Transformação das palavras para maiúsculas; e
c) Tokenização da sentença de descrição do produto e definição do *itemset* para ser tratado pelo algoritmo Apriori.

Ao final da aplicação do módulo de PLN, o conjunto ajustado de linhas de transações da venda da cerveja reduziu para 3.998.781 com 16.734.653 repetições de palavras.

A seguir, serão realizadas as etapas de seleção e análise dos principais termos da amostra. Foi aplicado o suporte mínimo de 0,000001 e identificados 3.271 termos distintos (figura 15). Os termos que mais se repetem são, respectivamente: CERVEJA (suporte 0,413) e CERV (suporte 0,370), termos coincidentes com a descrição do NCM 2203.xxxx: cerveja de malte. Assim o termo CERVEJA e seus equivalentes (por semelhança ou por reciprocidade) serão denominados de **termos principais** e todas as associações do termo principal com outros termos de interesse do domínio cerveja serão denominadas **associações preferenciais**. A seguir, a figura 16 apresenta o resultado parcial da análise de termos distintos para suporte mínimo de 0,000001:

Figura 16 – Exemplo do resultado da análise de termos distintos para suporte mínimo de 0,000001



#	RowID	Support(0-1): Number (double)	Items Set
3271	item set 3270	0.413	[cervej]
3270	item set 3269	0.37	[cerv]
3269	item set 3268	0.272	[350ml]
3268	item set 3267	0.255	[lt]
3267	item set 3266	0.203	[269ml]
3266	item set 3265	0.158	[malt]
3265	item set 3264	0.149	[pilsen]
3264	item set 3263	0.124	[brahm]
3263	item set 3262	0.119	[330ml]
3262	item set 3261	0.111	[heineken]
3261	item set 3260	0.103	[ln]
3260	item set 3259	0.102	[itaip]
3259	item set 3258	0.083	[lat]
3258	item set 3257	0.082	[pur]
3257	item set 3256	0.081	[kais]
3256	item set 3255	0.079	[ml]

Fonte: Dados da pesquisa (2025).

Como previsto no modelo Mineração de Texto, Mineração de Dados e Aprendizado de Máquina, o suporte mínimo foi variado em $x10$ para acompanhamento até a faixa ideal $0{,}0001 < S_{min} < 0{,}001$, com variações no número total de termos distintos. O resultado aponta uma faixa de 509 termos distintos, quando o suporte mínimo é 0,0001; até 147 termos distintos, quando o suporte é 0,001. Como modelo, propõe suporte mínimo de 0,0005 dos quais foram extraídos e identificados um conjunto de **210 termos** presentes que colaboram de alguma forma na descrição do produto. Para entender a função desses termos, o que significam e representam para descrição do produto, foram feitas as classificações e significações dos termos que poderão representar qualitativamente o produto cerveja, vide figura 17:

Figura 17 – Lista com resultado dos 210 termos distintos finais do modelo Mineração de Texto, Mineração de Dados e Aprendizado de Máquina

ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO
1	0,000501406	sixpack	43	0,000812028	12x1l	85	0,001582438	gel	127	0,003449425	und	169	0,012450170	tijuc
2	0,000503663	caipirinh	44	0,000815036	lima	86	0,001591464	trig	128	0,003524886	ipa	170	0,012762295	extra
3	0,000504665	tropical	45	0,000829075	appi	87	0,001594723	witbi	129	0,003525388	pm	171	0,014479361	teor
4	0,000513440	lour	46	0,000831332	congel	88	0,001609264	kg	130	0,003536669	210ml	172	0,017662539	cor
5	0,000523468	pi	47	0,000831332	maring	89	0,001622802	caix	131	0,003901443	cervitaip	173	0,018593651	crystal
6	0,000530989	24x330ml	48	0,000838853	weiss	90	0,001632830	c15	132	0,003940051	chop	174	0,022429910	arto
7	0,000535251	pma	49	0,000850134	unic	91	0,001645616	1x355ml	133	0,003962363	sh	175	0,022487071	355ml
8	0,000537006	prom	50	0,000850887	amber	92	0,001685477	npal	134	0,004219836	c12	176	0,022875410	300ml
9	0,000540516	350mlcr	51	0,000858408	artesanal	93	0,001710046	15x269ml	135	0,004532462	beats	177	0,025767772	stell
10	0,000547034	sub	52	0,000875205	heinek	94	0,001715311	vd	136	0,004549009	happy	178	0,028614507	600ml
11	0,000558817	35l	53	0,000895512	ice	95	0,001722331	473ml	137	0,004571321	dobr	179	0,029256056	sleek
12	0,000559068	1x300ml	54	0,000899022	atac	96	0,001723334	cristal	138	0,004721242	litra	180	0,030896156	original
13	0,000584389	12x269	55	0,000937379	golden	97	0,001741886	longneck	139	0,004745811	becks	181	0,032681413	lag
14	0,000587398	cx15	56	0,000937881	cervskol	98	0,001759685	eisen	140	0,004877931	baden	182	0,035054068	amstel
15	0,000587648	990ml	57	0,000939886	pack	99	0,001765201	brasil	141	0,004901247	1l	183	0,038580960	neck
16	0,000593164	nec	58	0,000942142	330355l	100	0,001769964	pmalt	142	0,005057435	orig	184	0,038988353	antarct
17	0,000593665	cervbrahm	59	0,000968717	negr	101	0,001781998	garraf	143	0,005083759	subzer	185	0,042131418	bohem
18	0,000597927	frut	60	0,000975235	6x330ml	102	0,001789268	cabar	144	0,005126629	lt269ml	186	0,042251254	spaten
19	0,000609710	fant	61	0,000976740	malzbi	103	0,001805313	lon	145	0,005162479	one	187	0,045696166	imperi
20	0,000615226	ext	62	0,000987770	sol	104	0,001806316	ambev	146	0,005283569	gf	188	0,047557888	dupl
21	0,000622245	unidad	63	0,000993537	gross	105	0,001809826	imper	147	0,005390118	premium	189	0,048736193	long
22	0,000634028	embtest	64	0,001000055	dup	106	0,001854451	patagon	148	0,005647089	way	190	0,050147652	budweis
23	0,000634530	escur	65	0,001001308	art	107	0,001876764	prem	149	0,005706255	proib	191	0,060514478	lta
24	0,000635783	beb	66	0,001022117	guar	108	0,001934175	sle	150	0,006030665	cerp	192	0,069857683	un
25	0,000647566	spat	67	0,001025877	12un	109	0,001936431	cx-12	151	0,006081808	12x350ml	193	0,073610459	skol
26	0,000650575	gin	68	0,001033900	camar	110	0,002064541	coronit	152	0,006269334	petr	194	0,078677671	chopp
27	0,000659349	343ml	69	0,001056714	rio	111	0,002089360	litr	153	0,006310449	lt350ml	195	0,079173061	ml
28	0,000661856	bud	70	0,001060474	ale	112	0,002143512	lneck	154	0,006656670	12x269ml	196	0,080636666	kais
29	0,000668124	pc12	71	0,001116883	unfiltered	113	0,002302959	gfa	155	0,006689011	1x330ml	197	0,081506856	pur
30	0,000669628	munichpur	72	0,001135435	350g	114	0,002333545	cer	156	0,006939965	ow	198	0,083331725	lat
31	0,000669879	cv	73	0,001141201	hour	115	0,002403993	gold	157	0,007783330	tig	199	0,102474415	itaip
32	0,000676899	5l	74	0,001203375	fi	116	0,002469175	glacial	158	0,007857037	devass	200	0,103036492	ln
33	0,000676899	1lt	75	0,001222679	pil	117	0,002523077	divers	159	0,008189971	la	201	0,110538032	heineken
34	0,000690938	gluten	76	0,001249254	ma	118	0,002527088	beer	160	0,008651265	munich	202	0,118773380	330ml
35	0,000705980	boh	77	0,001287862	caracu	119	0,002529846	pils	161	0,008775864	schin	203	0,123554289	brahm
36	0,000709991	pal	78	0,001325718	grf	120	0,002588009	350m	162	0,009254456	can	204	0,149042026	pilsen
37	0,000726287	laranj	79	0,001328225	export	121	0,002628121	pr	163	0,009357245	cx	205	0,158234809	malt
38	0,000758377	stel	80	0,001345524	ita	122	0,002660211	sens	164	0,010076512	eisenbahn	206	0,203171345	269ml
39	0,000771664	12x350	81	0,001358310	1x600ml	123	0,002789574	color	165	0,010735611	500ml	207	0,255103752	lt
40	0,000772166	dmalt	82	0,001380622	dm	124	0,002793335	250ml	166	0,010908094	1x350ml	208	0,272461186	350ml
41	0,000793476	cozumel	83	0,001462101	gt	125	0,003026489	269m	167	0,011999656	antart	209	0,369717214	cerv
42	0,000794478	american	84	0,001463354	bra	126	0,003218527	275ml	168	0,012039518	1x269ml	210	0,413370902	cervej

Fonte: Dados da pesquisa (2025).

A figura acima destaca 50 termos em amarelo que associam nomes de marcas de cerveja. A marca da cerveja e outros termos complementares foram identificados e denominados de **termos atributos** e suas relações, não incluindo o termo principal, denominadas **associações não preferenciais**. A seguir, no quadro 14 os termos foram classificados considerando seu significado, suas equivalências e reciprocidade na mesma língua natural.

Quadro 14 – Classificação dos termos: significado, equivalência e reciprocidade

Classificação	Termos equivalentes
Nome do produto CERVEJA	cervej, cerv, cervstell, cervaj, cervbohem, cervbrahm, cerevgarraf, cervpur, cervskol.
Quantidade do produto embalado (em ML)	350ml, 269ml, 330ml, ml, 300ml, 600ml, 355ml, 500ml, 350mlgross, 300mlchopp, 269mll, 300355l, 350m, 6x330ml, 210ml, 350g, 343ml, 35l, 250ml, 269m, 1l, litr, lt269ml, lt250ml, lt269m.
Unidade	un, cx, und.
Quantidade do produto vendido	1x269ml, 1x350ml, 1x600ml, 6x330ml, 12x1l, 269mlx15un, 12un, cx12und, cx15, cx-24, 12x269, pc12, 12x350, cx23, 1x355ml, c15, cx-12, c12, 12x269, 12x269ml, 1x330ml, 12x350ml, lt350ml, c23.

(quantidade embalada x unidade)	
Nome da marca (destaque em amarelo)	heineken, brahm, itaip, kais, ita, skol, budweis, bud imperi, amstel, spaten, bohem, antarct, original, stell, arto, antart, eisenbah, cabar, antarctic, caracu, eisen, antar, stel, orign, itaipav, becks, munich, tijuc, devass, schin, orig, eisenbahn, proib, petr, baden, imper, cerp, crystal, cor, coronit, glacial, color, patagon, tig, lokal.
Forma do recipiente do produto:	lt, ln, lat, long, neck, la, lata, longneck, lon, lneck, gfa, gf, cart.
Características complementares:	ale, american, appi, beats, chop, chopp, cozumel, dm, dmalt, dup, dupl, escur, export, ext, extra, gold, golden, grf, gluten, ice, lag, ma, malt, malzbi, pal, pi, pil, pils, pilsen, pm, pma, pr, prem, premium, pur, sub, subzero, teor, trig, tropical, unfiltered, weiss, witbi.
Nome do fabricante	ambev, brasil, rio, negr, heineken, cerp.

Fonte: Dados da pesquisa (2025).

Algumas análises complementares sobre o resultado de 210 termos distintos foram realizadas para ratificar a escolha antes da entrada na ontologia, buscando entender significado, completude e pertinência dos dados:

a) Dos 210 termos distintos selecionados, é possível ainda identificar alguns que não se relacionam com o domínio cerveja como: “caipirinha”, “refrig”, “kg”, “can”, “gt”, “grt”, “ar” etc, e que serão eliminados do resultado;
b) A classificação das marcas da cerveja alcançou 65% das marcas discriminadas na Resolução N° 0028/2023. Neste caso, essa Resolução é parte dos tipos de dados semiestruturados utilizados para a implementação da ontologia, enriquecendo futuramente as classes de marcas de cerveja. Entretanto, vale destacar que a amostra selecionada possuía somente 4 meses de venda de produto cerveja, podendo a ausência de termos ser originária no volume e no período da amostra dos dados selecionada pela SEFAZ/AM;
c) O número de termos distintos, quando classificados por seu significado e equivalência, representam 57,94% dos termos distintos utilizados na Resolução N° 0028/2023 para descrever os produtos relacionados à cerveja. Este valor percentual também pode ser influenciado pelo volume e período da amostra como explicado no item anterior;
d) “gt”: gin & tônica, aparece na descrição bruta de “CERV SKOL GT BEATS LATA 269ML” que não é classificada como cerveja de malte. Neste caso, para o produto “CERV SKOL GT BEATS LATA 269ML”, o NCM correto, que deveria ter sido informado na NFC-*e,* seria: 2206.xxxx ou 2208.xxxx;
e) “rio”, “negr”: nome de uma cervejaria no Amazonas, Cervejaria Rio Negro, e é classificado como um fabricante como “ambev”, produtor AMBEV. A cervejaria

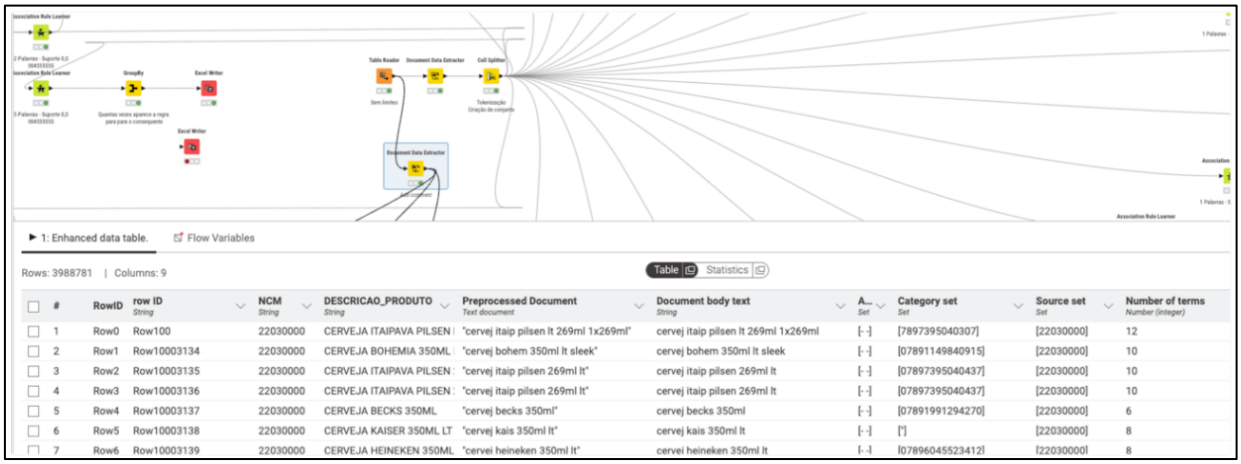
Rio Negro produz a cerveja TEMBETA, e por se tratar de cerveja artesanal, não é classificada pela SEFAZ/AM;

f) “way”, “ow”: se referem à caracterização de embalagens descartáveis, que são classificadas pela SEFAZ/AM e devem compor classes específicas dentro da forma do recipiente, se retornável ou não;
g) “sh”: abreviação de *shrink* que significa um pacote de cervejas, na forma de lata ou garrafa, envoltas por uma embalagem plástica flexível;
h) O termo “*golden*” parte da descrição “DOGMA GOLDEN VISION HAZY IPA LATA 473ML”, que possui um Suporte: $7{,}9963x10^{-5}$muito abaixo do Suporte mínimo, não sendo classificada para esta pesquisa. Inclusive, não consta na elaboração da Resolução N° 0028/2023 da SEFAZ/AM, possivelmente porque impacta muito pouco na arrecadação do imposto devido para o NCM 2203.xxxx, caracterizando um produto “*outlier*”, ou seja, um ponto fora da curva, um ponto que está distante da maioria dos dados analisados;
i) O termo “*golden*” também pode ser característica complementar de outras cervejas, “BADEN BADEN GOLDEN GRF 600ML”, “EISENBAHN STRONG GOLDEN ALE”. Estas sim, com suporte maior e mais impacto na arrecadação e que estão presentes na Resolução N° 0028/2023 da SEFAZ/AM;
j) A quantidade de produto vendido e a unidade, apesar de algumas vezes expressos no texto do campo descrição, não fazem parte das características do produto (a legislação NFC-*e* estabelece que essas informações estejam descritas em outros dois campos lógicos da nota: “quantidade” e “unidade”).

Além de conhecer termos e significados, foi preciso identificar as regras de associação para conhecer como eles se relacionam, qual o número ideal de termos do *itemset* para a representação e descrição útil do produto cerveja para a SEFAZ/AM.

Para fazer este levantamento foram agrupadas as linhas que possuíam uma associação de 2 termos, 3 termos e assim sucessivamente, ou seja, foram sumarizadas as linhas da amostra pela quantidade de termos do *itemset*, como apresentado na figura 18:

Figura 18 – Exemplo do resultado do processamento de PLN com o número de termos do *itemset*



▶ 1: Enhanced data table. | Flow Variables

Rows: 3988781 | Columns: 9 — Table | Statistics

#	RowID	row ID (String)	NCM (String)	DESCRICAO_PRODUTO (String)	Preprocessed Document (Text document)	Document body text (String)	A... (Set)	Category set (Set)	Source set (Set)	Number of terms (Number (integer))
1	Row0	Row100	22030000	CERVEJA ITAIPAVA PILSEN I	"cervej itaip pilsen lt 269ml 1x269ml"	cervej itaip pilsen lt 269ml 1x269ml	[-]	[7897395040307]	[22030000]	12
2	Row1	Row10003134	22030000	CERVEJA BOHEMIA 350ML	"cervej bohem 350ml lt sleek"	cervej bohem 350ml lt sleek	[-]	[07891149840915]	[22030000]	10
3	Row2	Row10003135	22030000	CERVEJA ITAIPAVA PILSEN :	"cervej itaip pilsen 269ml lt"	cervej itaip pilsen 269ml lt	[-]	[07897395040437]	[22030000]	10
4	Row3	Row10003136	22030000	CERVEJA ITAIPAVA PILSEN :	"cervej itaip pilsen 269ml lt"	cervej itaip pilsen 269ml lt	[-]	[07897395040437]	[22030000]	10
5	Row4	Row10003137	22030000	CERVEJA BECKS 350ML	"cervej becks 350ml"	cervej becks 350ml	[-]	[07891991294270]	[22030000]	6
6	Row5	Row10003138	22030000	CERVEJA KAISER 350ML LT	"cervej kais 350ml lt"	cervej kais 350ml lt	[-]	["]	[22030000]	8
7	Row6	Row10003139	22030000	CERVEJA HEINEKEN 350ML	"cervei heineken 350ml lt"	cervei heineken 350ml lt	[-]	[07896045523412]	[22030000]	8

Fonte: Dados da pesquisa (2025).

A figura 18 apresenta os componentes do item set, por exemplo: #1 (primeira de cima para baixo) com duas colunas descrevendo o *itemset,* são elas: *Preprocessed Document* e *Document body text*. Elas contêm a mesma informação, os termos que compõem o *itemset*, já a coluna *Number of terms* contém a quantidade de termos da linha #1, neste caso 12 termos (6 *Preprocessed Document* e 6 *Document body text*). Para fins de análise da quantidade de termos de uma linha, será utilizada a coluna *Preprocessed Document,* totalizando 6 elementos do *itemset*, neste exemplo: "cervej itaip pilsen lt 269ml 1x269ml". O resultado da sumarização do número de termos do itemset está na figura 19 abaixo:

Figura 19 – Sumarização do número de linhas x número de termos do *itemset*

▶ 1: Group table | Flow Variables

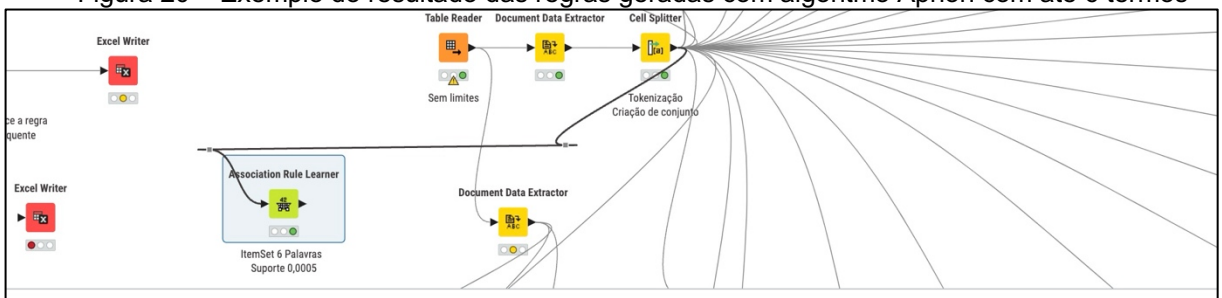
Rows: 12 | Columns: 2

#	RowID	Number of terms (Number (integer))	Count(row ID) (Number (integer))
1	Row0	2	74538
2	Row1	4	297714
3	Row2	6	794195
4	Row3	8	1117772
5	Row4	10	1127177
6	Row5	12	414914
7	Row6	14	142472
8	Row7	16	18399
9	Row8	18	1276
10	Row9	20	226
11	Row10	22	97
12	Row11	24	1

Fonte: Dados da pesquisa (2025).

Considerando que a coluna *Number of terms* na figura 19 apresenta, como já explicado, o número de termos duplicados, é possível afirmar que onde lê-se 10 termos, na realidade trata-se de 5 termos, *itemset* que mais se repete na amostra com 1.127.177 linhas. Assim, é preciso buscar na amostra o comportamento das relações com 5 termos para descrever o produto cerveja de forma completa e útil e definir a combinação desejada no Apriori. Entretanto, para assegurar uma maior possibilidade de associações e análise, o algoritmo Apriori foi executado com até 6 termos do tipo 1 consequente e combinação de até 5 antecedentes. Isso está evidenciado na figura 20:

Figura 20 – Exemplo do resultado das regras geradas com algoritmo Apriori com até 6 termos



▶ 1: Frequent itemsets/Association rules Flow Variables

Rows: 6620 | Columns: 6 Table | Statistics

#	RowID	Support *Number (double)*	Confidence ↓ *Number (double)*	Lift *Number (double)*	Consequent *String*	implies *String*	Items *Set*
3	rule2	0.001	1	22.102	ln	<---	[unfil,355ml,cerv,eisen]
4	rule3	0.001	1	24.223	cerv	<---	[ln,unfil,355ml,eisen]
5	rule4	0.001	1	40.138	355ml	<---	[ln,unfil,cerv,eisen]
6	rule5	0.001	1	316.047	eisen	<---	[ln,unfil,355ml,cerv]
10	rule9	0.001	1	24.223	cerv	<---	[itaip,lta,269ml]
14	rule13	0.001	1	45.959	stell	<---	[arto,ln,330ml]
16	rule15	0.001	1	24.223	cerv	<---	[lta,malt]
17	rule16	0.001	1	45.959	stell	<---	[arto,ln]
19	rule18	0.001	1	22.102	ln	<---	[lag,cerv,330ml]
21	rule20	0.001	1	24.223	cerv	<---	[itaip,lta]
24	rule23	0.001	1	7.969	cervej	<---	[350ml,pur,24x350ml,long,malt]
26	rule25	0.001	1	7.723	malt	<---	[350ml,pur,24x350ml,long,cervej]
27	rule26	0.001	1	22.332	pur	<---	[350ml,24x350ml,long,cervej,mal
28	rule27	0.001	1	19.364	long	<---	[350ml,pur,24x350ml,cervej,malt]
29	rule28	0.001	1	515.513	24x350ml	<---	[350ml,pur,long,cervej,malt]

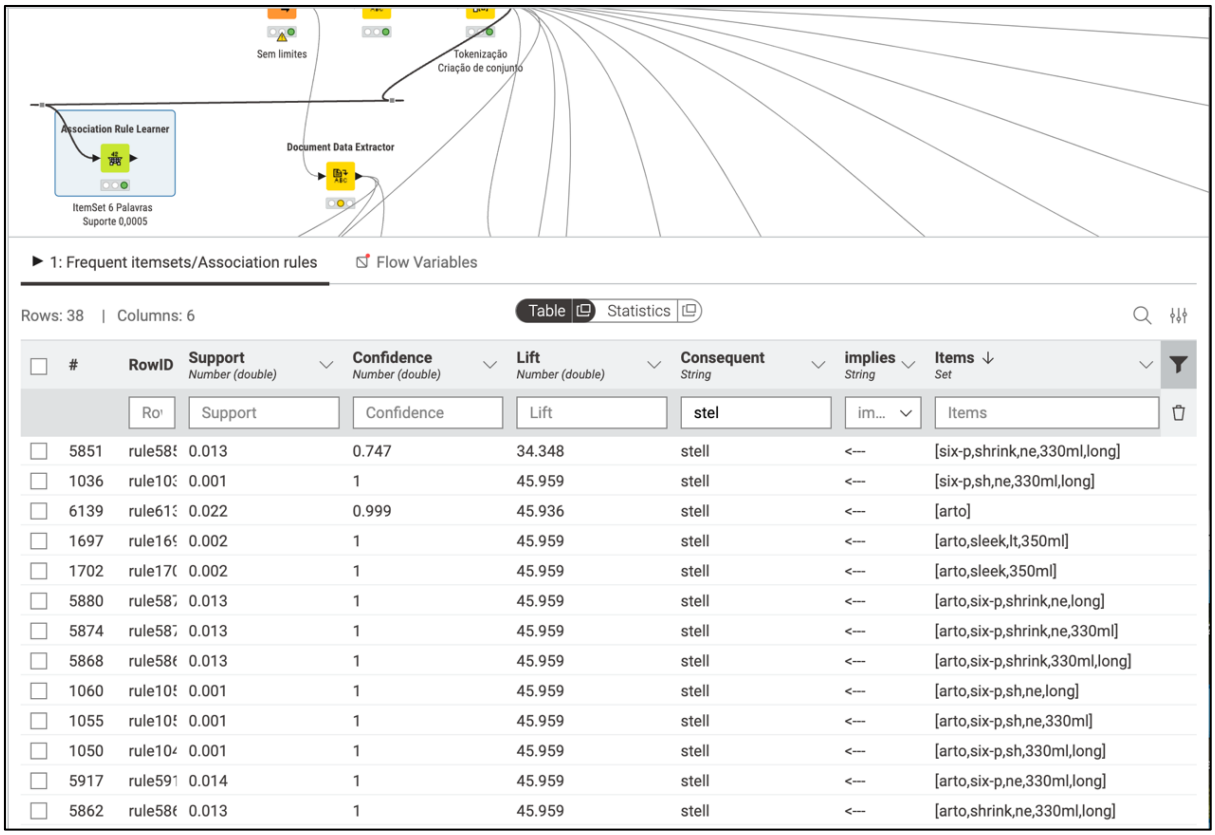
Fonte: Dados da pesquisa (2025).

A figura 20 apresenta um exemplo do resultado gerado pelo algoritmo Apriori com o suporte mínimo estabelecido para a pesquisa de 0,0005, até 6 termos e uma seleção para confiança máxima 1. As linhas destacadas #10 e #14 mostram as relações entre o termo **cerv + itaip + lta + 269ml** e **stell + arto + ln + 330ml,** respectivamente. A primeira linha refere-se à cerveja Itaipava, lata de 269ml, uma

associação preferencial, pois apresenta o nome do produto que está associado ao NCM cerveja de malte. Já a segunda linha apresenta uma associação não preferencial, pois não traz o nome do produto cerveja em nenhum dos seus formatos de escrita. Ambas as linhas trazem a confiança igual a 1, o que significa uma grande certeza que os termos aparecem juntos. Contudo, quando se analisa o *Lift*, 24,223 na linha #10 e 45,959 na linha #14, a associação não preferencial **stell + arto + ln + 330ml** é muito mais forte, apesar de não apresentar o termo “cerveja” ou “cervej” ou “cerv”. Isso acontece porque, na amostra, a associação dos termos **stella + artois** dificilmente acontece em outra relação que não seja desta forma. Outro exemplo desta alavancagem é a associação **brahma + chopp,** por isso o *Lift* mais alto. Já a relação do tipo **lata + 269ml** (linha #10) aparece em muitos tipos e marcas de cerveja, representando uma relação mais diluída na amostra.

Ainda sobre o detalhamento da linha # 14, a figura 21, confirma a força da regra para a relação **stella + artois**:

Figura 21 – Detalhamento da relação **stella + artois**



#	RowID	Support Number (double)	Confidence Number (double)	Lift Number (double)	Consequent String	implies String	Items Set
	Ro	Support	Confidence	Lift	stel	im...	Items
5851	rule58	0.013	0.747	34.348	stell	<---	[six-p,shrink,ne,330ml,long]
1036	rule10	0.001	1	45.959	stell	<---	[six-p,sh,ne,330ml,long]
6139	rule61	0.022	0.999	45.936	stell	<---	[arto]
1697	rule16	0.002	1	45.959	stell	<---	[arto,sleek,lt,350ml]
1702	rule17	0.002	1	45.959	stell	<---	[arto,sleek,350ml]
5880	rule58	0.013	1	45.959	stell	<---	[arto,six-p,shrink,ne,long]
5874	rule58	0.013	1	45.959	stell	<---	[arto,six-p,shrink,ne,330ml]
5868	rule58	0.013	1	45.959	stell	<---	[arto,six-p,shrink,330ml,long]
1060	rule10	0.001	1	45.959	stell	<---	[arto,six-p,sh,ne,long]
1055	rule10	0.001	1	45.959	stell	<---	[arto,six-p,sh,ne,330ml]
1050	rule10	0.001	1	45.959	stell	<---	[arto,six-p,sh,330ml,long]
5917	rule59	0.014	1	45.959	stell	<---	[arto,six-p,ne,330ml,long]
5862	rule58	0.013	1	45.959	stell	<---	[arto,shrink,ne,330ml,long]

Fonte: Dados da pesquisa (2025).

Nestas combinações da figura 21, a linha #5851 (primeira linha de cima para baixo) apresenta uma relação com confiança e *Lift* mais baixos que as demais

relações apresentadas nas linhas seguintes. Isto acontece porque, apesar de vários antecedentes para o consequente **stella (six-p + shrink+ ne + 330ml + long),** o termo **artois** não está presente na relação. Já na linha seguinte, # 1036 (segunda linha de cima para baixo), a confiança e o *Lift* aumentam para 1 e 45,959 respectivamente, mantendo-se nas demais linhas da figura 21. Isso porque, em todas as relações, o termo **arto (**de **artois)** demonstra a força da regra entre os termos.

Ao finalizar as análises das associações quanto ao termo e número de termos e como eles se relacionam, foi possível definir:

a) A quantidade de produto embalado está frequente na descrição do produto porque não apresenta outro campo para registro e representa a caraterística do volume do líquido embalado em um recipiente ou a capacidade, tornando-se classe da ontologia;
b) O nome da marca do produto comercializado complementa o termo “cerveja”, caracterizando o nome comercial pelo qual o produto é conhecido para venda, tornando-se classe da ontologia;
c) A forma do recipiente do produto representa a forma e o material de produção, descreve o tipo de embalagem usada para transportar o produto vendido, tornando-se classe da ontologia;
d) A associação entre volume e tipo de embalagem gerou um terceiro conceito chamado de embalagem, necessário porque cada tipo de embalagem suporta um volume ou alguns volumes específicos, por exemplo: a lata pode suportar 269ml ou 330ml, mas não suporta 1 litro. Assim, a embalagem concatena tipo de embalagem e volumes suportados e os relaciona diretamente com o produto cerveja, criando um relacionamento na forma “a cerveja é acondicionada em uma embalagem, a embalagem é de um tipo e tem um volume”, da mesma forma como a SEFAZ/AM descreve a cerveja. Esse metadado selecionado a partir da descrição do produto se tornou classe na ontologia.
e) As características complementares são termos que complementam o nome da marca do produto, ressaltando características de produção do produto, tornando-se uma classe da ontologia.
f) A partir dos resultados alcançados, pode-se propor que a descrição do produto cerveja apresenta 5 características qualitativas relevantes e frequentes, são elas: 1. nome da marca comercial do produto, 2. complemento da marca comercial, 3. volume ou capacidade, 4. tipo da embalagem e 5. embalagem.

Essas características representarão os metadados da descrição do produto, a expressão que melhor comunica algo dentro do domínio de interesse “CERVEJA”, ou “CERV” + **nome da marca** “ORIGINAL” ou “HEINEKEN” ou “SKOL” + **complemento da marca** “PURO MALTE” ou “BEATS” + **embalagem com volume ou capacidade** “269ML” ou “330ML” ou “350ML” + **tipo da embalagem** “LN” ou “LATA”.

No processo identificação das regras fortes com maior confiança e *Lift* foram identificados os termos com maior pertinência ao domínio, aprendendo como eles se relacionam entre si. Este processo se repetiu de forma iterativa (n vezes) e incremental (extraindo ruídos) até que os resultados apresentassem baixa aleatoriedade dos termos para que se tornassem as instâncias das classes na ontologia (*Individual*), como estruturas de relacionamento de superordenação e subordinação (relações hierárquicas) (Lara, 2004b). O exemplo das hierarquias criadas durante o processo iterativo incremental de superordenação e subordinação é apresentado na figura 22:

Figura 22 – Geração das instâncias (*Individual*) das subclasses da marca da cerveja

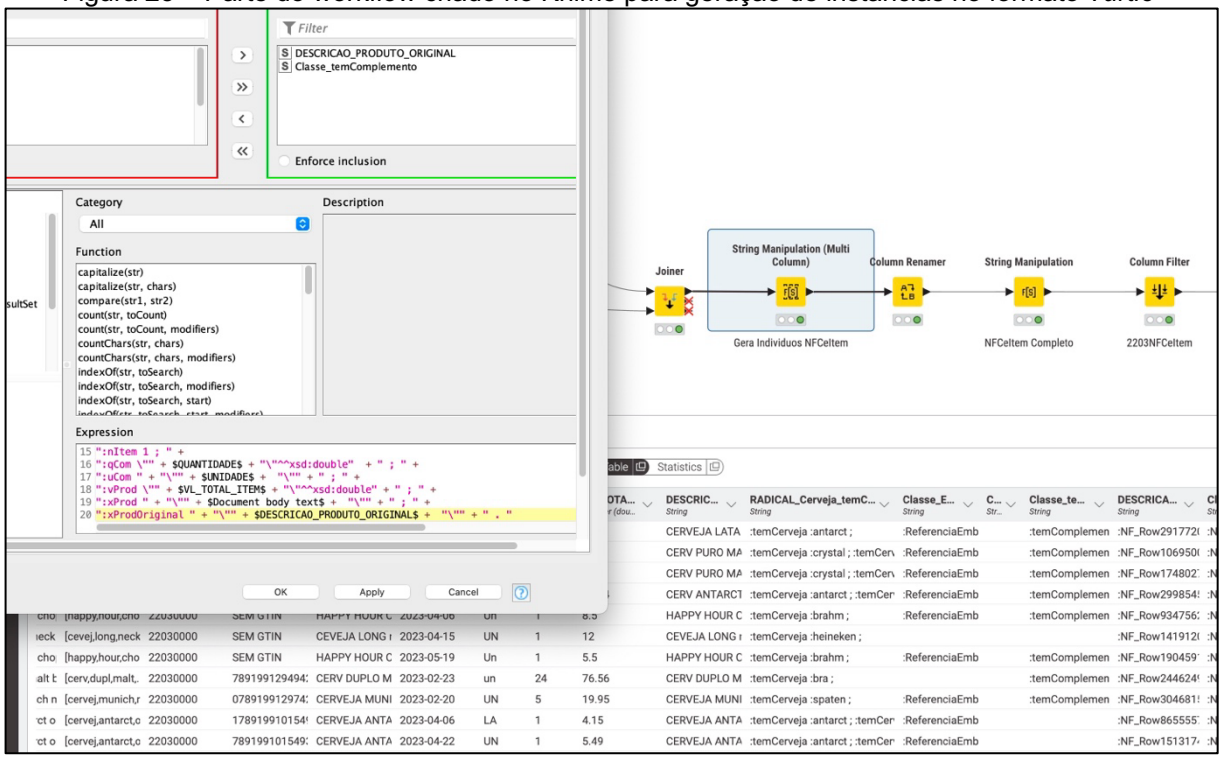
Individual	Classe	Protégé Turtle Format
amstel	Amstel	:amstel rdf:type owl:NamedIndividual , :Amstel .
ant	Antarctica	:ant rdf:type owl:NamedIndividual , :Antarctica .
anta	Antarctica	:anta rdf:type owl:NamedIndividual , :Antarctica .
antar	Antarctica	:antar rdf:type owl:NamedIndividual , :Antarctica .
antarc	Antarctica	:antarc rdf:type owl:NamedIndividual , :Antarctica .
antarct	Antarctica	:antarct rdf:type owl:NamedIndividual , :Antarctica .
antarctic	Antarctica	:antarctic rdf:type owl:NamedIndividual , :Antarctica .
antart	Antarctica	:antart rdf:type owl:NamedIndividual , :Antarctica .
bad	Baden_Baden	:bad rdf:type owl:NamedIndividual , :Baden_Baden .
baden	Baden_Baden	:baden rdf:type owl:NamedIndividual , :Baden_Baden .
badenbaden	Baden_Baden	:badenbaden rdf:type owl:NamedIndividual , :Baden_Baden .
beck	Becks	:beck rdf:type owl:NamedIndividual , :Becks .
becks	Becks	:becks rdf:type owl:NamedIndividual , :Becks .
boh	Bohemia	:boh rdf:type owl:NamedIndividual , :Bohemia .
bohem	Bohemia	:bohem rdf:type owl:NamedIndividual , :Bohemia .

Fonte: Dados da pesquisa (2025).

A figura 22 mostra a relação de hierarquia existente entre os diversos termos identificados para caracterizar uma classe da marca cerveja com os termos: ant, anta, antar, antarc, antarct, antarctic, antart; todos os indivíduos da marca de cerveja Antarctica, por exemplo. A última coluna da figura 21 apresenta a sintaxe em formato Turtle, gerada no Knime a partir de código implementado, a finalidade é importar neste formato (classificado e formatado) as milhões de linhas da amostra, não só para as

classes de marca de cerveja, mas também para as demais classes identificadas no ambiente de aprendizado da ontologia, com especial atenção para a classe NFCeItem que detém milhões de linhas na amostra, como demonstra a figura 23:

Figura 23 – Parte do workflow criado no Knime para geração de instâncias no formato *Turtle*



Fonte: Dados da pesquisa (2025).

É possível observar na figura 23 um trecho do código utilizado para gerar sintaxes de instâncias da classe NFCeItem no formato Turtle para importação no modelo de ontologia que será implementado a seguir. O detalhe de uma instância, ou seja, uma linha de NFCeItem é apresentado na figura 24:

Figura 24 – Sintaxe Turtle de uma instância da classe NFCeItem

```
:NF_Row17480272 rdf:type owl:NamedIndividual , :NFCe ; :temItem :Row17480272 ;
:é_Emitida :11111111000111 ; :Recebida_pelo
<http://www.semanticweb.org/diana_doutorado/ontologies/2024/4/2203NFCe/111111111
11> ; :dhEmi "2023-05-08T00:00:00"^^xsd:datetime . :Row17480272 rdf:type
owl:NamedIndividual , :NFCeItem ;    :temCerveja :crystal ; :temCerveja :tig ;
:temComplemento :pur ; :temComplemento :malt ;        :ReferenciaEmbalagem
:Lata_350_ml ;        :temNCM :22030000 ; :NCM "22030000" ; :cEAN "7896045506705" ;
:nItem 1 ; :qCom "1.0"^^xsd:double ; :uCom "UN" ; :vProd "3.29"^^xsd:double ; :xProd "cerv pur
malt crystal tig 350ml" ; :xProdOriginal "CERV PURO MALTE CRYSTAL TIGER 350ML" .
```

Fonte: Dados da pesquisa (2025).

A sintaxe da Figura 24 descreve uma instância da classe principal da ontologia, a NFCeItem, com seus *Data Properties* e *Object Properties*, ou seja, os relacionamentos entre classe e o conteúdo da nota fiscal, cuja importância se dá porque detecta conteúdo não explicito da descrição do produto ou descrito de forma reduzida e todas as variações das classes cerveja, volume, embalagem, tipo embalagem etc.

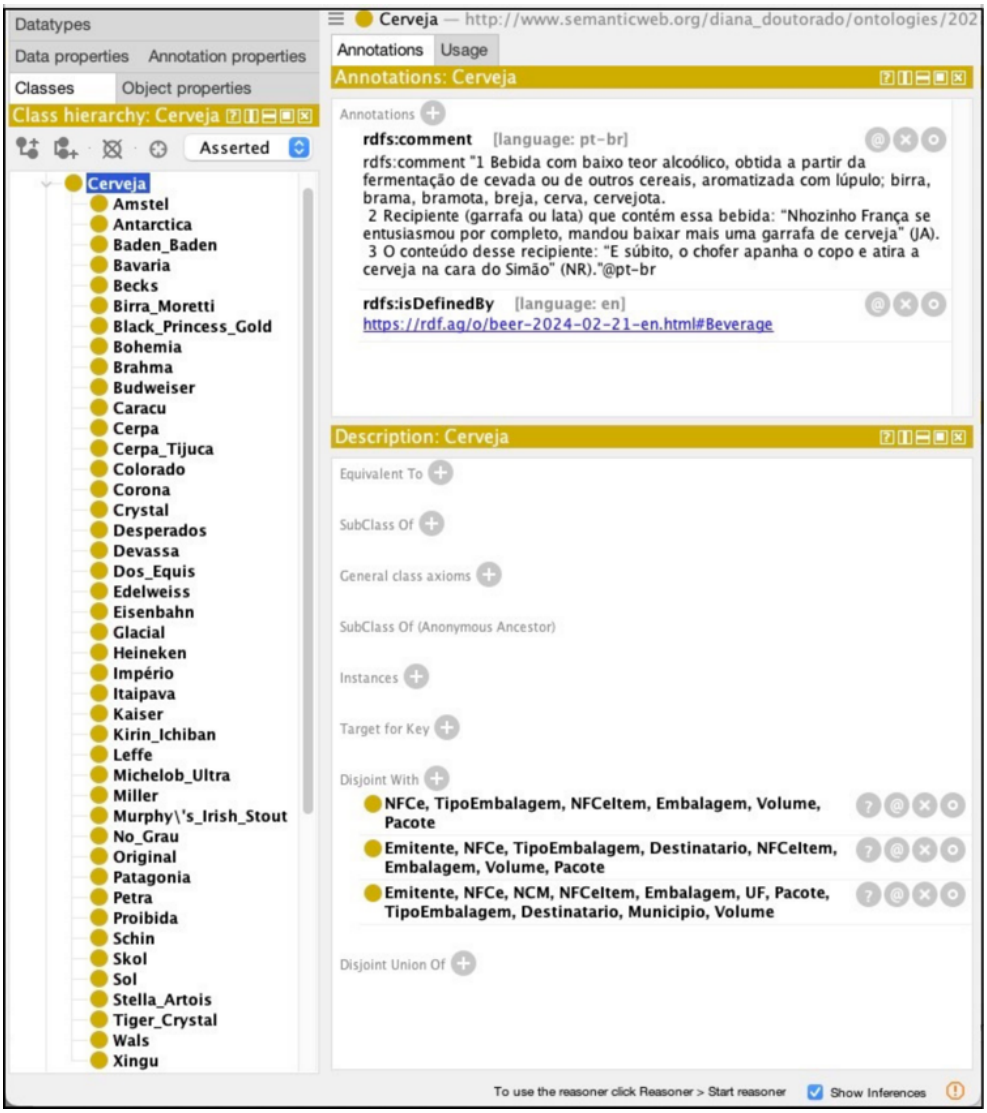
Após a geração das milhões de linhas de instâncias para o modelo de ontologia, resultado da aplicação do ambiente de aprendizado da ontologia com MT, MD e AM e requisitos da Arquitetura da Informação para indicação do termo candidato para representar o conceito considerando: contexto, fontes bibliográficas e eventuais sinônimos, como características que permitem relacionar as instâncias no formato, é possível chegar a resultados mais eficazes e partir para a implementação da ontologia.

4.3.3 Implementação e Enriquecimento da ontologia 2203NFCe

Para implementação da ontologia 2203NFCe foi utilizado o software Protégé[8], iniciando pela criação da superclasse Cerveja e das subclasses com os nomes das marcas de cerveja extraídas do Apriori na etapa de aprendizado da ontologia, validadas e enriquecidas com a Resolução Nº 0028/2023 SEFAZ/AM conforme figura 25:

[8] https://protege.stanford.edu

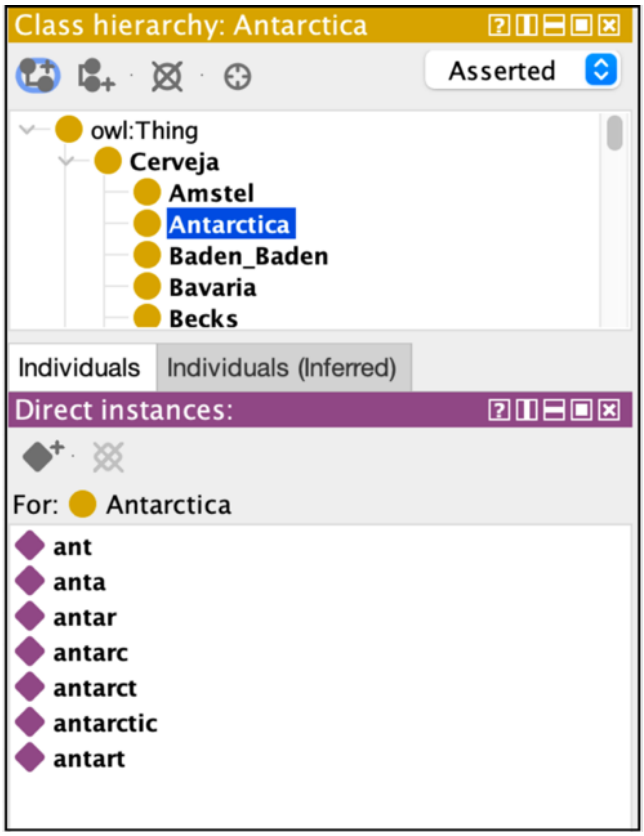
Figura 25 – Hierarquia da superclasse Cerveja



Fonte: Dados da pesquisa (2025).

Cada uma dessas classes com marcas de cerveja na Figura 25 já estão implementadas com as suas instâncias correspondentes, isto é, possíveis termos equivalentes que também representam a marca, a partir da geração da sintaxe Turtle e importadas pelo Protégé. São exemplos desta geração as instâncias da classe Antarctica na figura 26:

Figura 26 – Instâncias da classe Antarctica

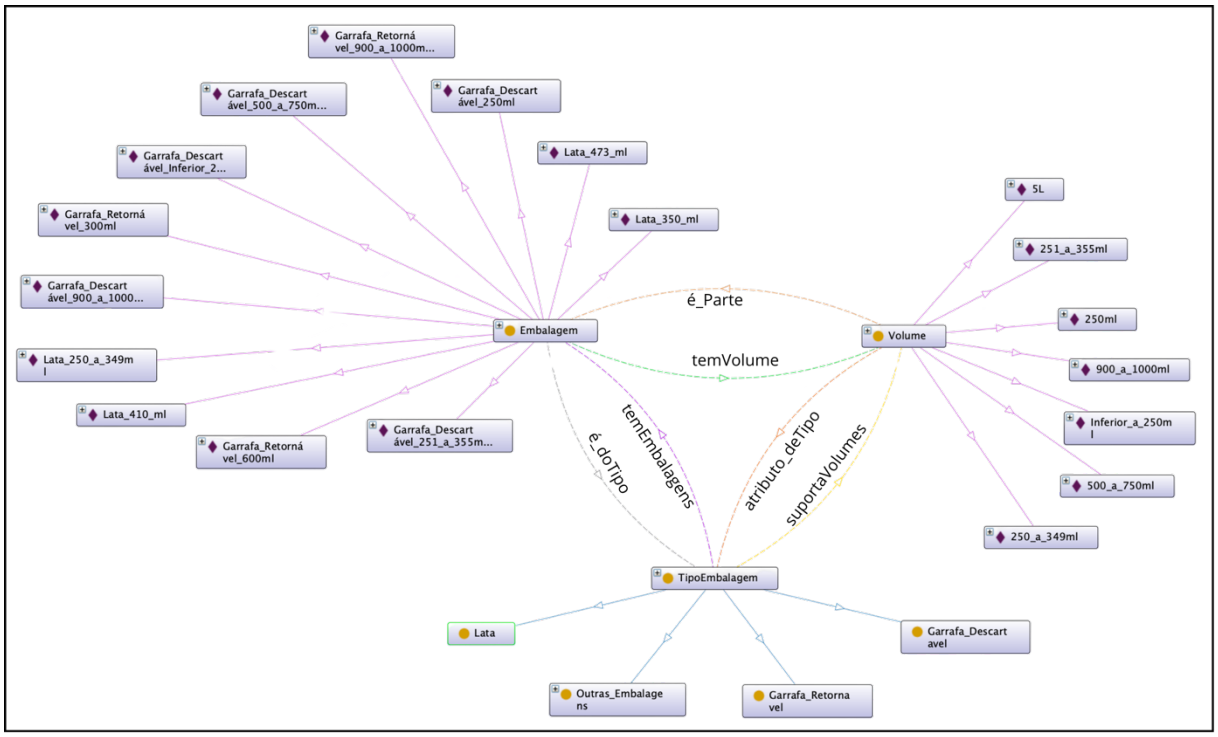


Fonte: Dados da pesquisa (2025).

Reafirma-se na figura 26 como instâncias da classe Antarctica as seguintes variações na forma de descrição do produto: ant, anta, antar, antarc, antarct, antarctic e antart. Todas elas já descritas na figura 26 e detectadas na fase de aprendizado da ontologia.

A seguir foram definidas as classes que descrevem como o produto cerveja se apresenta: a superclasse TipoEmbalagem e as classes Volume e Embalagem. Os relacionamentos ou *Object Property Assertions* entre essas classes criam a estrutura necessária para identificar qual a “embalagem” e qual o “volume” da cerveja comercializada, descrevendo as embalagens existentes no mercado e cada um dos volumes que elas aceitam, como se apresenta na figura 27:

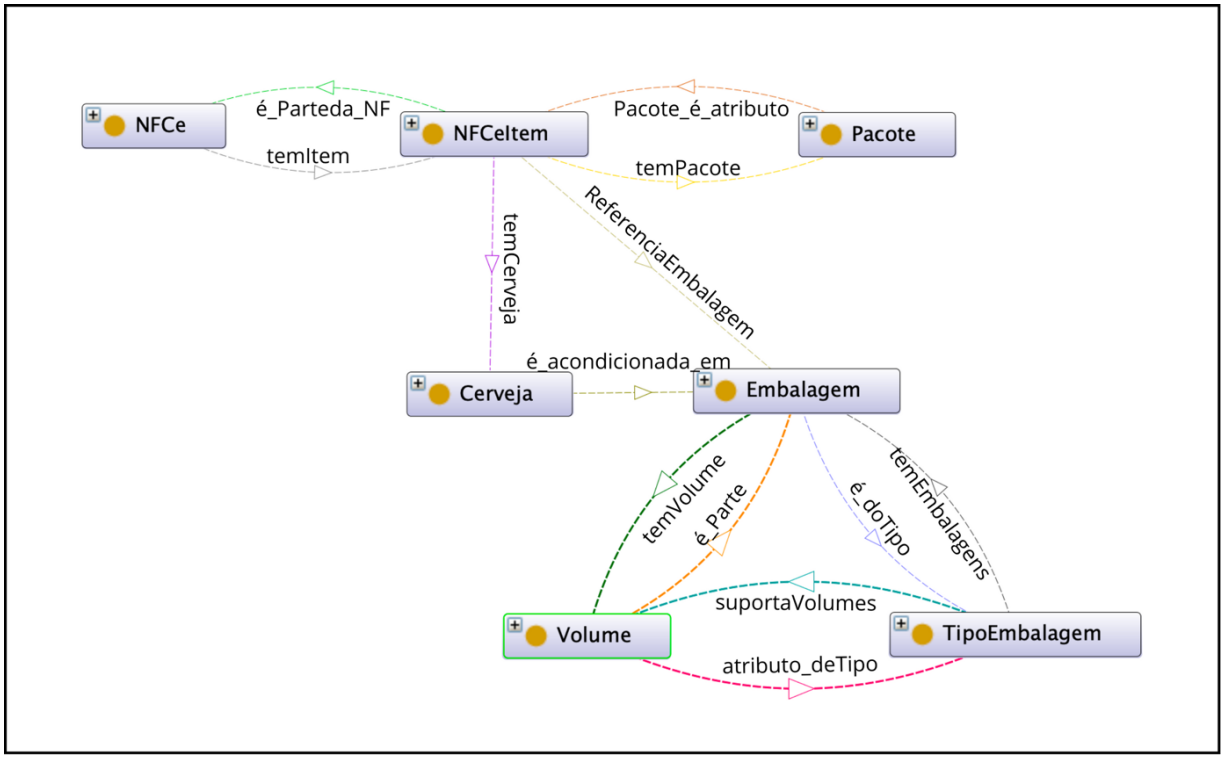
Figura 27 - Módulo de apresentação da cerveja



Fonte: Dados da pesquisa (2025).

A Figura 27 apresenta o “módulo de apresentação da cerveja” com relações hierárquicas e não hierárquicas da descrição do produto associado à superclasse cerveja. Este módulo será associado a uma classe que representa a comercialização do produto, a classe Pacote, uma classe de multiplicação para representar o número de vezes que a embalagem está sendo comercializada no item da nota fiscal. A associação entre a comercialização do produto e o módulo de apresentação da cerveja é feita na classe NFCeItem, em detalhe da figura 28:

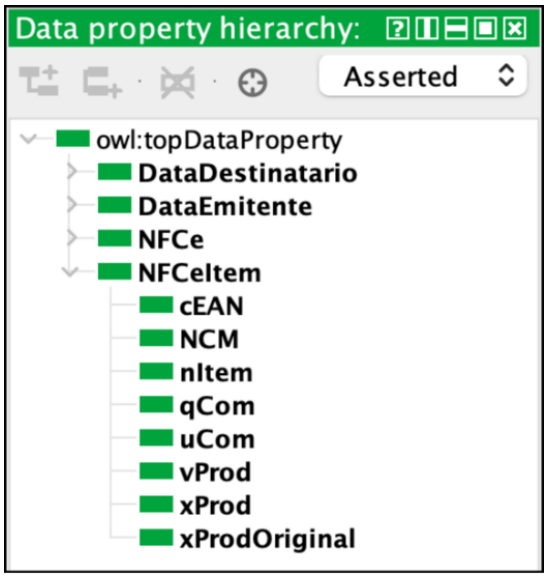
Figura 28 - Destaque da relação entre as classes TipoEmbalagem, Volume, Embalagem e Pacote e as classes NFCeItem e Cerveja



Fonte: Dados da pesquisa (2025).

A figura 26 introduz a classe NFCeItem, uma classe que representa a venda do produto, cada transação ou linha do cupom fiscal vendido. As instâncias da classe NFCeItem contém todas as informações do item cerveja vendido através de Data Properties. Eles compõem campos do .xml da nota fiscal e são conhecidos como: cEAN: código de barras do produto (código GTIN); NCM (código NCM); nItem (ordem sequencial dos itens da nota fiscal); qCom (quantidade comercial do produto); uCom (unidade comercial do produto); vProd (valor total do item da nota); xProd (descrição do produto livre tratado) e xProdOriginal (descrição do produto livre conforme nota fiscal), como se demonstra na figura 29:

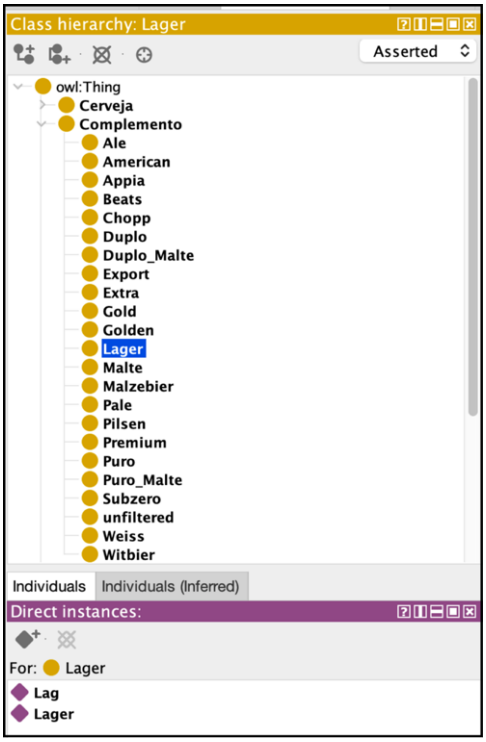
Figura 29 – *Data Properties* da classe NFCeItem



Fonte: Dados da pesquisa (2025).

A descrição da cerveja inclui também o complemento da marca representada na ontologia 2203NFCe pela superclasse Complemento, criada dentro da mesma estratégia da superclasse Cerveja, pois possui subclasses específicas e hierárquicas com as suas instâncias, isto é, variações na forma de descrição conforme exemplo da figura 30:

Figura 30 – Superclasse Complemento com detalhe para as instâncias de subclasse Lager

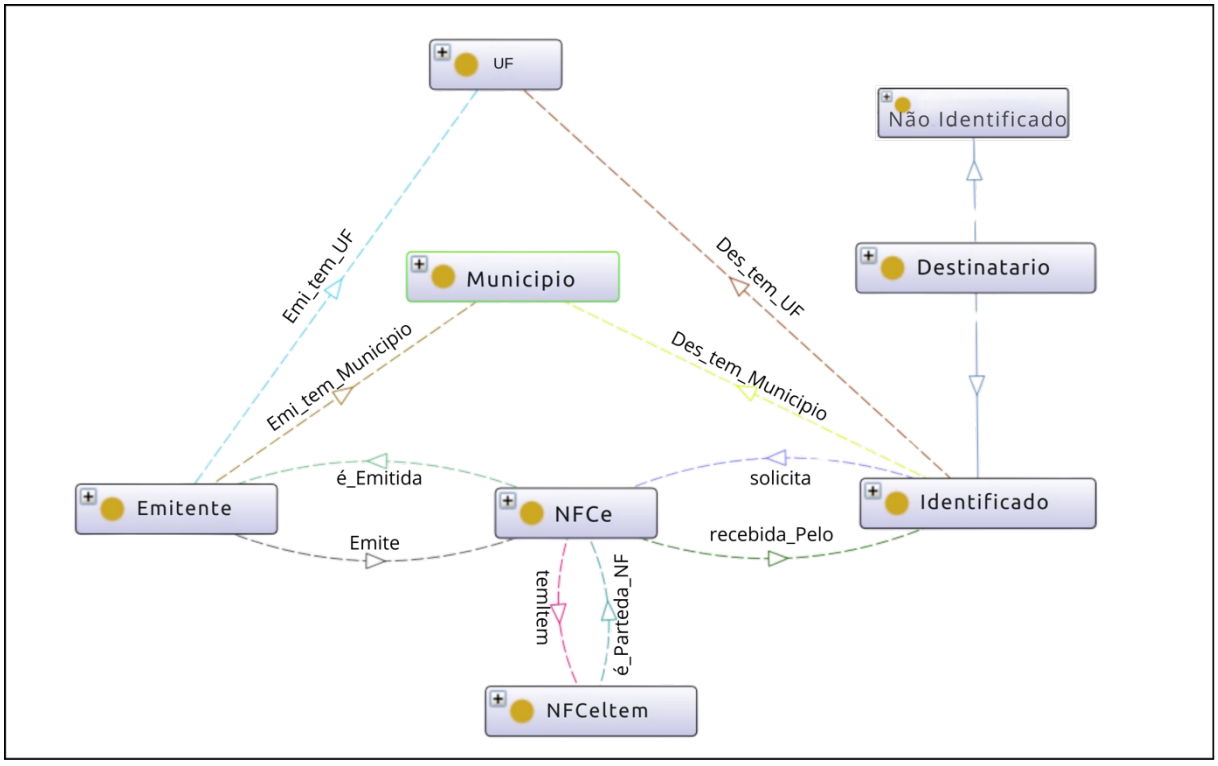


Fonte: Dados da pesquisa (2025).

A figura 30 apresenta todas as subclasses de Complemento extraídas no aprendizado da ontologia e enriquecidas pela Resolução Nº 0028/2023 SEFAZ/AM. Para a subclasse Larger, as instâncias possíveis são: Lag e Lager.

Por fim, foi criado um módulo de venda na ontologia 2203NFCe na qual estão os participantes da transação comercial, ou seja, vendedor (emitente) e comprador (destinatário) que poderá ser identificado ou não com a inclusão de CPF ou CNPJ na nota fiscal. Além dos dados temporais de data e hora e dados de localização: endereço, Código de Endereçamento Postal (CEP), cidade, unidade federativa (UF) etc., foram incluídos todos os dados conforme manual do XML da NFC-*e* (seção 2). As figuras 31 e 32 representam o modelo de classes da ontologia e os seus *Data Properties* para o módulo de venda:

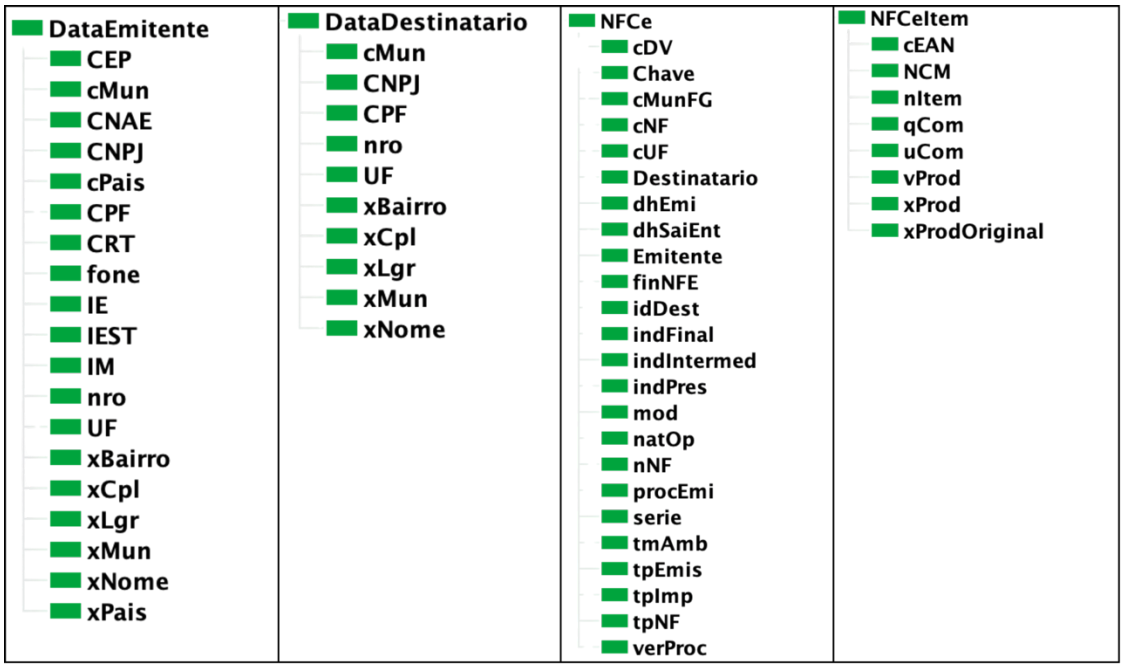
Figura 31 – Módulo de venda na ontologia 2203NFCe



Fonte: Dados da pesquisa (2025).

O módulo de venda se concatena à descrição do produto em NFCeItem através de uma nova classe da nota fiscal, classe NFCe, que representa o próprio cupom de venda e cujos Data Properties estão descritos a seguir, na figura 32:

Figura 32 – *Data Properties* do módulo de venda da ontologia 2203NFCe

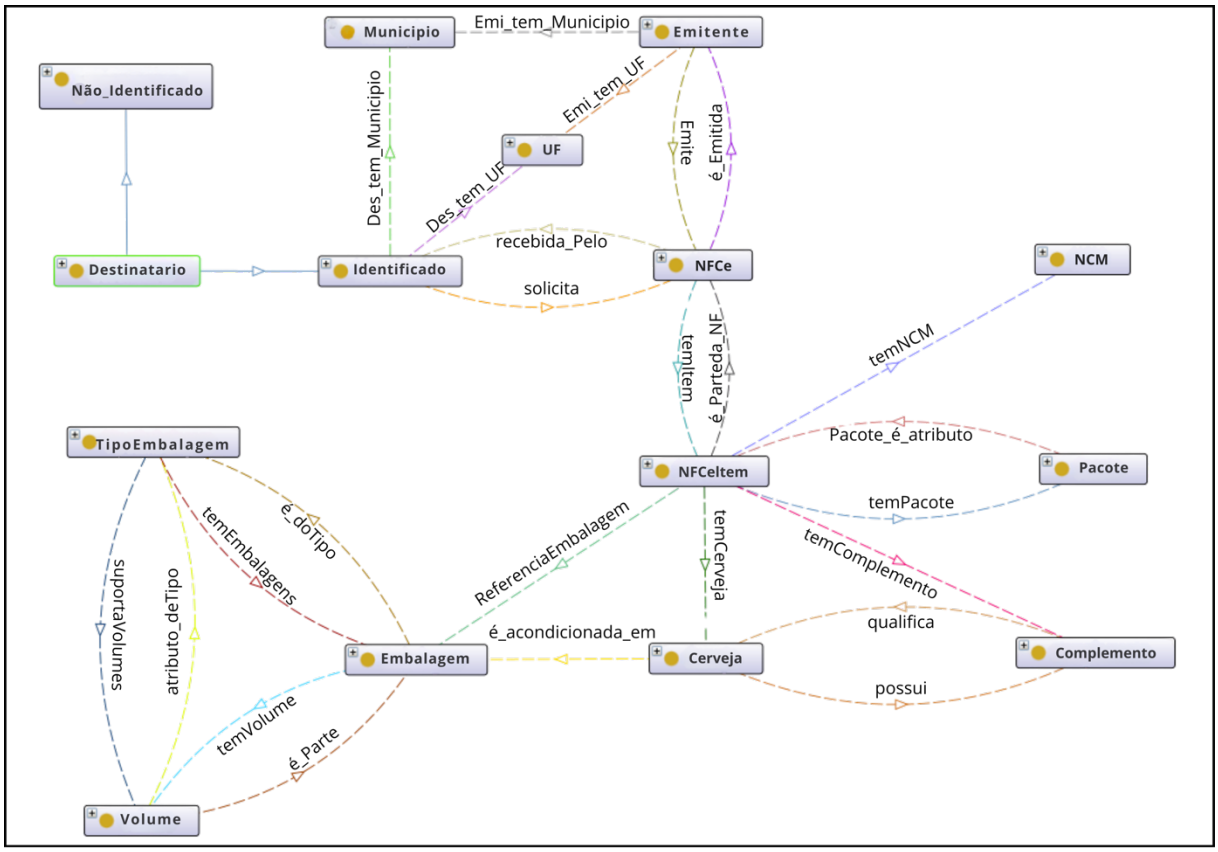


Fonte: Dados da pesquisa (2025).

A figura 32 apresenta todos os campos do XML do cupom fiscal e que estão logicamente criados no banco de dados da SEFAZ/AM. Isso é necessário para facilitar as buscas por parte da auditoria da secretaria, cruzando as informações não estruturadas da ontologia com as informações já estruturadas da SEFAZ/AM. A partir desse cruzamento, possibilitam-se diversas análises como agrupamentos por destinatários, por emissores, por localização geográfica, de preço médio de produto como por exemplo: preços médios praticados no interior do Estado do Amazonas versus preços médios praticados na Capital Manaus, ou qualquer outra necessidade dos auditores.

Ao final da construção dos módulos, definição de Object Property Assertions e importação dos *Data Property Assertions* o modelo da ontologia 2203NFCe é apresentado na figura 33:

Figura 33 – Ontologia 2203NFCe



Fonte: Dados da pesquisa (2025).

A Figura 33 é a saída da fase de implementação e enriquecimento da ontologia 2203NFCe validada pelo motor lógico de raciocínio do Protégé, o HermiT, que verifica as inconsistências entre classes e relações produzidas (linhas tracejadas), executando inferências automáticas de acordo com as restrições implementadas. Um exemplo da utilização do motor de raciocínio são as classes de marcas de cerveja, disjuntas, e que não podem aparecer juntas na mesma descrição de um produto, e as classes de complemento da marca, não disjuntas, em que mais de um complemento pode estar concatenado a uma determinada marca de cerveja.

Visando a um fácil acesso à ontologia 2203NFCe, foi criado um sítio eletrônico com toda a documentação gerada pelo Protégé[9].

9 https://nfce2203-ontologia.ddns.net

4.3.4 Validação da ontologia 2203NFCe

As formas de validação da ontologia 2203NFCe foram definidas a partir do objetivo do estudo, observando o comportamento da informação gerada pelo modelo e a forma que produz conhecimento e impacta nos processos de auditoria. Buscou-se também validar a forma como o raciocínio foi produzido no AM, visto que uma das premissas de construção do modelo é a de não ter o especialista durante as fases de extração dos termos, aquisição dos requisitos, conceitualização e implementação.

Assim, primeiro foram construídas as consultas de Questão de Competência que fazem parte da forma de validação da metodologia de construção da ontologia escolhida; a seguir a validação dos termos, conceitos e relações a partir da comparação com pesquisa em outra base interna da SEFAZ/AM, a base de notas NF-*e*; e ao final, a validação da informação do modelo com uma base externa para busca de produtos.

4.3.4.1 Questão de Competência

A validação da ontologia 2203NFCe, por meio das Questões de Competência (QC), quer demonstrar a utilidade, completude e alcance do modelo para apresentar informações corretas, relevantes e aplicáveis para a auditoria. As Questões de Competência foram elaboradas a partir das referências deste estudo, dos experimentos realizados nas amostras recebidas e utilizadas e na proposta da pesquisa, partindo da necessidade de identificar os produtos no item de descrição do cupom fiscal. Consideraram-se para isso as suas diversas formas de representação e venda da cerveja e as diversas formas de fiscalização no processo de auditoria. Sendo assim, foram elaboradas cinco Questões de Competência:

1 Quais os termos que compõem uma expressão completa para descrição do produto, ou seja, qual sequência de termos (quantidade e qualidade) forma uma sentença que pode ser utilizada para descrever corretamente a cerveja de malte em determinada área de auditoria?
2 Quais os termos que compõem uma expressão completa para descrição da venda do produto, ou seja, qual sequência de termos (quantidade e qualidade) forma uma sentença que pode ser utilizada para descrever corretamente a venda da cerveja de malte em determinada área de auditoria?

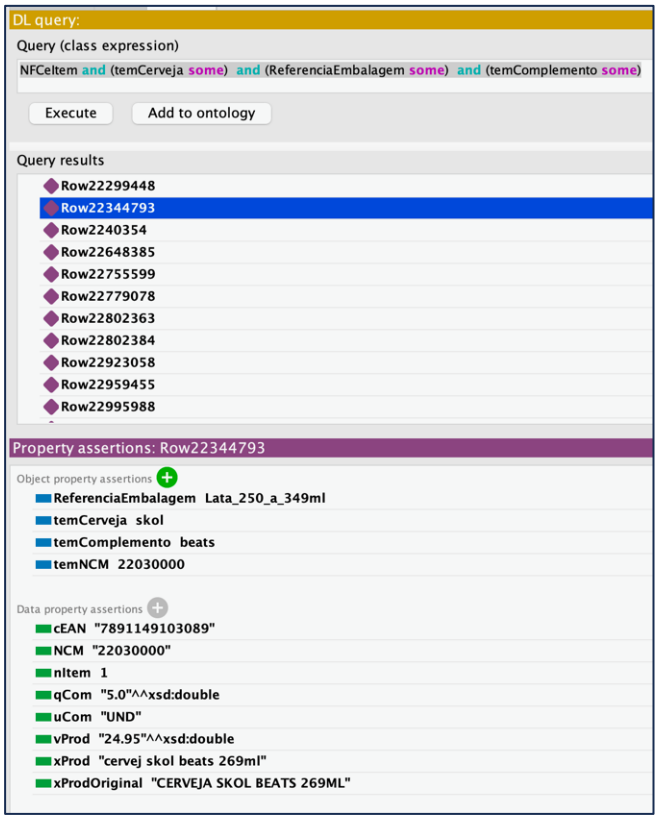
3 É possível identificar quais descrições de produto são úteis para um determinado uso na auditoria? Quais sequências de termos que, mesmo incompletas, podem ser úteis à auditoria?

4 É possível identificar descrições de produto em desacordo com o código de NCM em uma NFC-*e*?

5 É possível gerar análises sobre a identificação de produto cerveja de malte vendido por período, localidade, emissor ou destinatário?

Na resposta da Questão de Competência 1, a sentença completa corresponde à expressão correta que apresenta todos os metadados que compõem o item: “termo nome do produto + termo nome da marca + termo volume + termo tipo da embalagem + termo complemento da marca” e é obtido pela consulta (*DLQuery*): “NFCeItem and (temCerveja some) and (ReferenciaEmbalagem some) and (temComplemento some)”. Várias transações ou linhas de descrição do produto cerveja atendem à consulta conforme detalhe da figura 34:

Figura 34 – Resposta da Questão de Competência 1 através de DLQuery com destaque na linha ROW 22344793

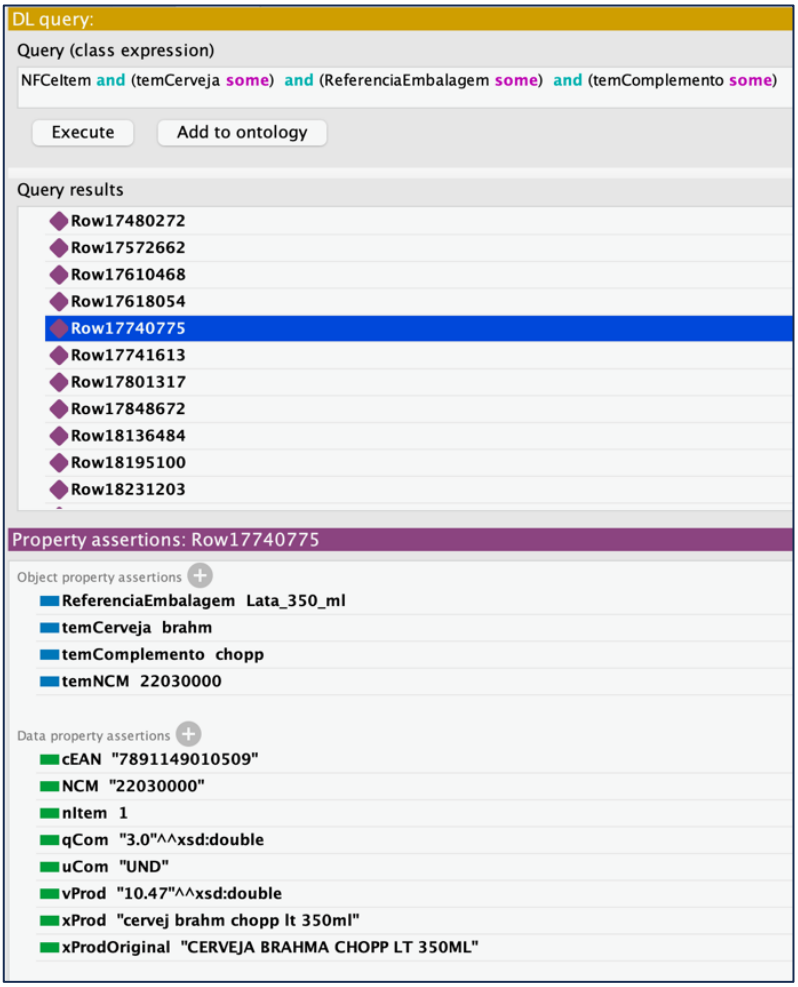


Fonte: Dados da pesquisa (2025).

O detalhamento da linha ROW 22344793, na figura 34, com os *Object Property Assertions* e *Data Property Assertions* apresenta a instância da classe NFCeItem com a seguinte descrição do produto original: CERVEJA SKOL BEATS 269ML. Neste exemplo, a classe Embalagem concatena o tipo da embalagem e o volume que ele suporta, que fica entre os limites: “Lata_250_a_349ml”; a classe Cerveja indica o nome da marca da cerveja: “skol”; a classe Complemento indica o termo ou os termos que estão completando o nome da marca, neste exemplo: “beats”. O detalhamento da sentença respondida pela Questão de Competência 1 apresenta todos os metadados selecionados para a descrição completa do produto, sendo eles: nome do termo principal + nome da marca da cerveja + complemento da marca da cerveja + embalagem (tipo da embalagem + volume da embalagem).

Outro detalhamento é apresentado pela figura 35 com a linha ROW 17740775:

Figura 35 – Resposta da Questão de Competência 1 através de DLQuery com destaque na linha ROW 17740775

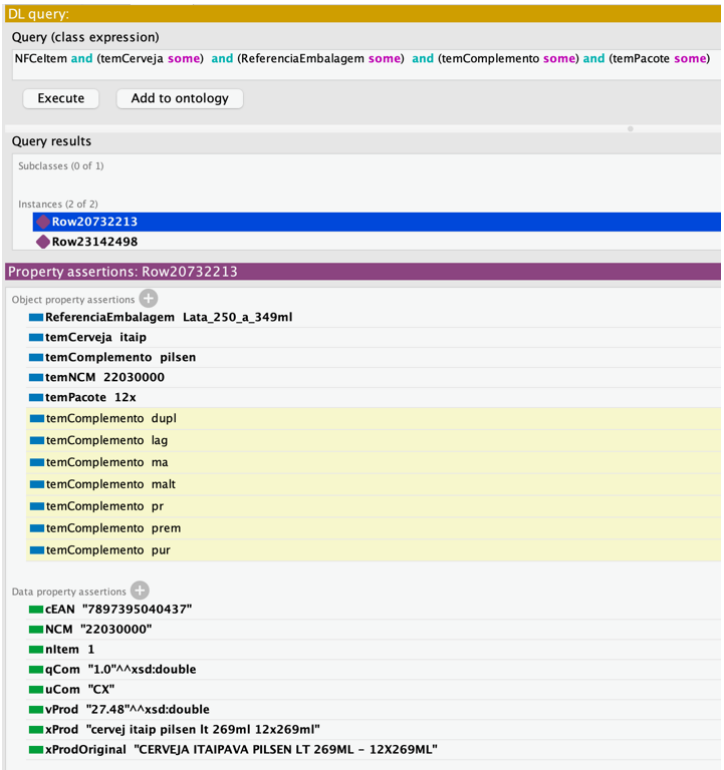


Fonte: Dados da pesquisa (2025).

Neste caso da figura 35, a linha ROW 17740775 apresenta a instância da classe NFCeItem com a descrição do produto original: CERVEJA BRAHMA CHOPP LT 350ML. Neste exemplo, a classe Embalagem está fixa no valor Lata_350_ml; a classe Cerveja indica o nome da marca da cerveja: “brahm”; e a classe Complemento indica o termo: “chopp”. O detalhamento da linha ROW 17740775, assim como o primeiro exemplo, também apresenta todos os metadados selecionados para a descrição completa do produto, assumindo outros valores para validar a linha de venda da nota fiscal.

A resposta da Questão de Competência 2 é obtida pela consulta (*DLQuery*): “NFCeItem and (temCerveja some) and (ReferenciaEmbalagem some) and (temComplemento some)and (temPacote some)”. A diferença é o acréscimo da classe Pacote porque o pacote é uma referência à parte comercial, como a cerveja é empacotada para venda. Como exemplo à resposta, o detalhe da linha ROW 20732213 na figura 36:

Figura 36 - Resposta da Questão de Competência 2 através de DLQuery com destaque na linha ROW 20732213



Fonte: Dados da pesquisa (2025).

A descrição do produto original da linha ROW 20732213 é: CEREJA ITAIPAVA PILSEN LT 269ML – 12X269ML. Neste exemplo, o termo: 12X269ML refere-se ao número de cervejas comercializadas durante a venda e descritas no item da nota fiscal: 12 latas, ou seja, um pacote com 12 latas. A inclusão do pacote na descrição do produto é um problema existente no item da nota. Neste exemplo, a unidade comercial “uCom” é identificada como caixa “CX” e a quantidade comercial “qCom” apresenta a quantidade “1”, e isso não identifica propriamente a venda, pois não há informação alguma de quantas cervejas vêm na caixa, a informação da quantidade de cervejas está na descrição do produto. O preço, por exemplo, é da caixa “vProd” é “27.48”. Entretanto, sem a identificação do pacote, não é possível identificar o preço unitário da cerveja, pois também não há qualquer informação no XML de quantas cervejas da NFC-*e* vêm na caixa vendida.

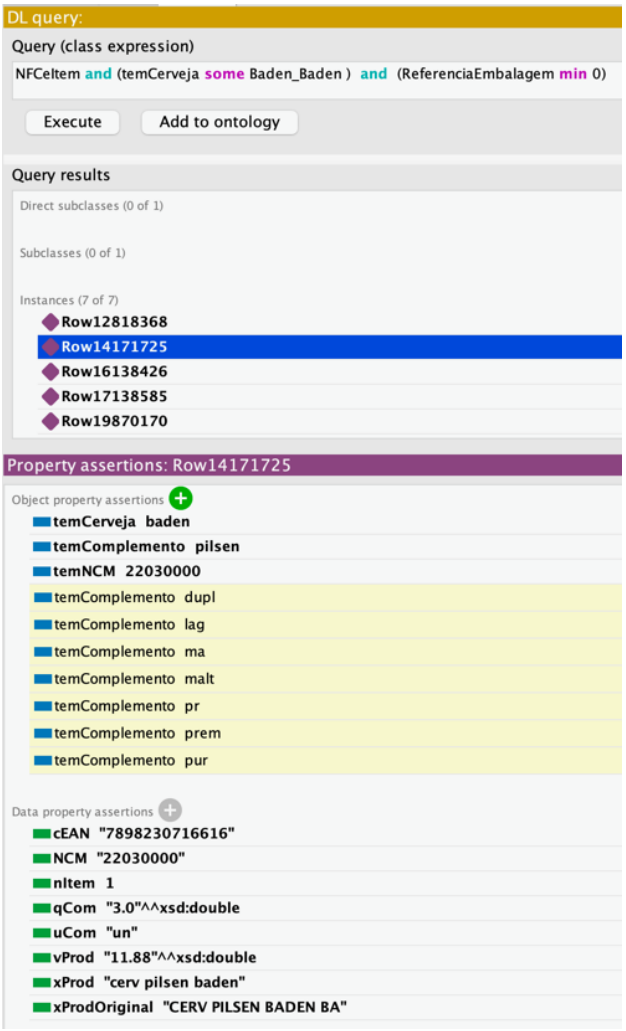
Neste caso, a ontologia resolve este problema porque identifica na descrição do produto o número de cervejas vendidas: “12X269ml” fazendo referência à 12 latas dentro de 1 caixa. Neste caso, o detalhamento da sentença respondida pela Questão de Competência 2 apresenta todos os metadados selecionados para a descrição completa da venda do produto: “nome do termo principal + nome da marca da cerveja + complemento da marca da cerveja + embalagem (tipo da embalagem + volume da embalagem)” associado aos dados estruturados da NFCeItem “a quantidade comercializada, a unidade comercializada e o valor total do produto”.

Outro destaque apresentado na figura 36 é a lista de complementos que aparecem nas 7 (sete) linhas em amarelo. A lista corresponde aos descritos que avaliam, qualificam e identificam a marca de cerveja Itaipava, além do já apresentado pelo *Object Property* “temComplemento: pilsen”. Neste caso, o modelo 2203NFCe identificou que a cerveja Itaipava poderia vir acompanhada de outros complementos como: “temComplemento: dupl”; “temComplemento: lag”; “temComplemento: ma”; “temComplemento: malt”; “temComplemento: pr”; “temComplemento: prem”; “temComplemento: pur”. Cada um desses complementos está associado à classe e às instâncias cadastradas na ontologia, portanto, prontos para serem validados a partir do reconhecimento do texto da descrição original do produto.

O uso da descrição completa do produto e da venda do produto, validadas pela ontologia 2203NFCe para construção do PMPF permitem a extração do preço unitário do produto, utilizando tanto os campos estruturados quanto os campos não estruturados da nota fiscal.

A resposta da QC3 é obtida pela consulta (*DLQuery*): “NFCeItem and (temCerveja some Baden_Baden) and (ReferenciaEmbalagem min 0)”. Nesta consulta o interesse está na descrição do nome da marca da cerveja, marca Baden Baden, entretanto não há um critério para que contenha ou não a embalagem, fazendo uma associação ao mínimo zero para ReferenciaEmbalagem por exemplo, como se identifica na figura 37:

Figura 37 – Resposta da Questão de Competência 3 através de DLQuery com destaque na linha ROW 14171725



Fonte: Dados da pesquisa (2025).

O detalhe da linha ROW 14171725 descreve metadados termo principal concatenado ao termo do nome da marca da cerveja “CERV PILSEN BADEN BA”. Como as demais consultas, a descrição do produto vem acompanhada de *Object Property Assertions* e *Data Property Assertions,* portanto, apresenta outras

informações da transação de venda como NCM, quantidade, unidade e valor total da nota.

A classificação útil para esta descrição está associada à aplicação do resultado em determinado processo de auditoria como ações de fiscalização no Desembaraço da Nota Fiscal ou ações de inteligência e análise fiscal no Acompanhamento da Comercialização do Produto. No primeiro caso, o Desembaraço da Nota Fiscal busca identificar primeiramente o NCM da nota fiscal para determinar a tributação, conferindo se a descrição do produto está de acordo com o NCM informado no campo da nota. O fato de identificar o termo cerveja e o termo do nome da marca Baden de acordo com o NCM 2203000 já caracteriza a ausência ou tentativa de fraude fiscal, liberando o agente fiscalizador para dar andamento no processo tributário do Desembaraço. No segundo caso, o Acompanhamento da Comercialização do Produto pode utilizar esta consulta, por exemplo, para analisar se regiões do Estado ou da Cidade de Manaus estão vendendo ou não a cerveja Baden para ajuste do preço médio.

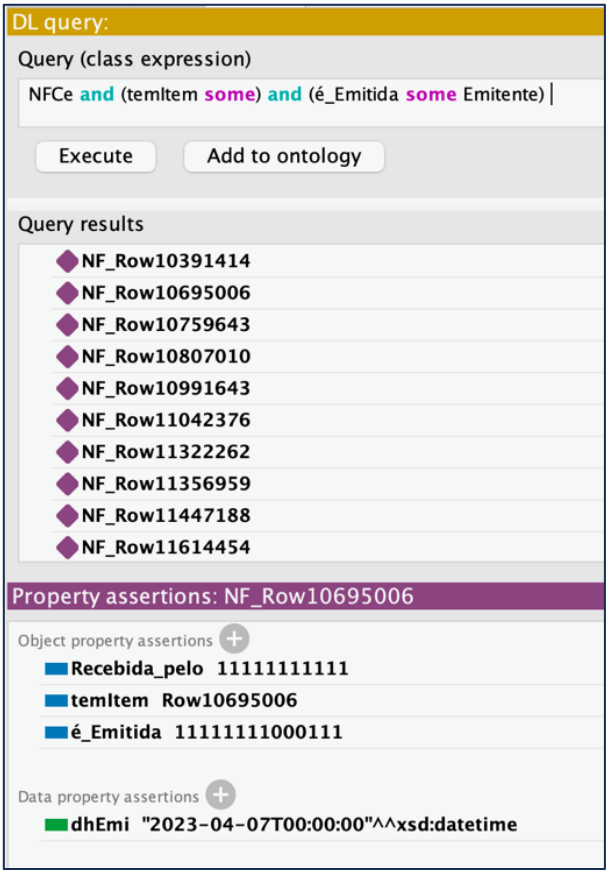
A Questão de Competência 4 é respondida na execução das *DLQueries* das Questão de Competência 1, Questão de Competência 2 e Questão de Competência 3 nas figuras anteriores (figuras 32, 33 e 34) nas quais o NCM é validado no *Object Property Assertions: "*tem NCM" com valor "22030000".

A classe NCM tem instâncias cadastradas com diversos valores como 22030000, 220300222, 22030099, 22030300, 22031000, 22032100, 22033000, 22039000 todos extraídos das análises da base de dados da amostra durante o aprendizado da ontologia. Neste caso, como a amostra possui NCM iniciados por 2203.xxxx não há como apresentar um exemplo de NCM diferente, entretanto, pode-se aplicar o mesmo modelo e comparar, primeiramente na descrição, para identificar a cerveja e, a seguir, comparar qual NCM esta descrição possui, podendo daí demonstrar se está diferente dos *Object Property Assertions* já aprendidos e absorvidos pela ontologia 2203NFCe.

Apesar do modelo da ontologia 2203NFCe conter o módulo de venda com o controle dos campos estruturados de emissor, emitente, localidade, período etc. por meio de classes, a resposta à Questão de Competência 5 fica parcialmente prejudicada quando da apresentação dos *Object Property Assertions* e *Data Property Assertions,* pois a SEFAZ/AM, ao ceder os dados, anonimizou-os em virtude do sigilo fiscal e da Lei Geral de Proteção de Dados (LGPD). Assim, não há referências à CNPJ

ou CPF de emissor e destinatário. Os dados disponíveis foram mascarados e resumem-se à sequência numérica 111.111.111-11 para CPF e 11111111/0001-11 para CNPJ. Dessa forma, o resultado das *DLQueries* para pesquisas envolvendo emissor e destinatário está apresentado na figura 38:

Figura 38 - Resposta da Questão de Competência 5 através de DLQuery com destaque na linha ROW 10695006



Fonte: Dados da pesquisa (2025).

Na figura 38 a ontologia 2203NFCe reconhece tanto o emissor quanto o destinatário quando responde à *DLQuery* se existe alguma linha de nota fiscal “emitida” e “recebida” por algum contribuinte ou pessoa física ou jurídica. Ao identificar a linha ROW 10695006 (destaque da figura 38) associando-a aos dados estruturados de emissor e destinatário para a consulta, responde à Questão de Competência 5 respeitando o sigilo fiscal da amostra, porém quando aplicada à base de produção, pressupõe-se que o modelo apresentaria os dados reais e de interesse. Já o detalhamento da linha ROW 10695006 segue o modelo das Questões de Competência 1, 2 ou 3.

4.3.4.2 Base interna Nota Fiscal Eletrônica (NF-e)

A validação da ontologia 2203NFCe utilizando os dados da base de NF-*e* a partir de uma amostra do período de 01/02/2023 a 31/05/2023, coincide com o período da primeira amostra fornecida pela SEFAZ/AM, com o total de 725.327 transações ou linhas da nota fiscal e 2.093 termos distintos ou *tokens*.

A diferença entre as bases NF-*e* e NFC-*e* se dá pois, basicamente, a primeira é derivada de distribuidores atacadistas, com um grande volume de venda de produtos em cada item da nota e uma margem maior de regularidade na descrição do produto; a segunda base, como visto durante esta pesquisa, está vinculada aos varejistas, apresenta um pequeno volume de venda de produtos em cada item da nota e uma grande variação na descrição do produto. Esta validação tem o objetivo de comparar e analisar os resultados do ambiente de aprendizado da ontologia e dos termos encontrados para construção do modelo e avaliar a sua completude e alcance, visto que, em se tratando de atacadistas, a descrição do produto cerveja tende a estar completa na NF-*e*, isto é, apresentando todos os termos em quantidade e qualidade já concatenados.

Todas as técnicas e parâmetros de preparação dos dados, Mineração de Texto e Aprendizado de Máquina utilizadas na amostra de NFC-*e* foram reproduzidas para a nova amostra NF-*e* e, ao final, foram selecionados 141 termos conforme figura 39 abaixo**:**

Figura 39 - Lista de termos selecionados na base NF-*e*

ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO
1	0,000509	24x330ml	51	0,002811	grf	101	0,021758	stell
2	0,000511	740ml	52	0,002931	pull	102	0,023406	cx12
3	0,000578	appi	53	0,002931	off	103	0,024914	355ml
4	0,000579	american	54	0,002934	cx04	104	0,028364	990ml
5	0,000594	cxpap	55	0,003098	24x600ml	105	0,031234	heineken
6	0,000597	ultra	56	0,003164	eisen	106	0,033619	ow
7	0,000604	unfil	57	0,003201	coronit	107	0,035537	way
8	0,000607	500ml	58	0,003593	pmalt	108	0,035538	one
9	0,000627	6pa	59	0,003593	eisenbahn	109	0,041283	cerv
10	0,000652	malzbi	60	0,003834	lne	110	0,043509	budweis
11	0,000669	shr	61	0,003957	barril	111	0,044778	pur
12	0,00067	cx6	62	0,004035	12x269ml	112	0,045244	ln
13	0,000673	trop	63	0,004169	keg	113	0,048608	spaten
14	0,000818	litr	64	0,004628	210ml	114	0,051115	ne
15	0,000818	car	65	0,00475	lokal	115	0,051643	long
16	0,000858	unf	66	0,004919	becks	116	0,054034	original
17	0,000873	weiss	67	0,004929	ita	117	0,056463	c15
18	0,000903	des	68	0,005824	275ml	118	0,061785	pc12
19	0,000917	descartavel	69	0,006226	cx24	119	0,061808	itaip
20	0,000942	grf300ml	70	0,006283	carta	120	0,068766	ttc
21	0,000986	baden	71	0,006407	petr	121	0,072775	chopp
22	0,001115	caix	72	0,006601	six	122	0,077054	bohem
23	0,00114	premium	73	0,007078	pil	123	0,080355	dupl
24	0,00114	retorn	74	0,007397	schin	124	0,088854	330ml
25	0,001186	30l	75	0,007846	devass	125	0,089977	600ml
26	0,001266	ale	76	0,008351	fridg	126	0,098757	sleek
27	0,001275	amb	77	0,00861	subzer	127	0,118604	antarct
28	0,001301	12x1	78	0,008837	cxpap12	128	0,12549	cervej
29	0,001348	18un	79	0,00974	tiju	129	0,129491	malt
30	0,00135	proib	80	0,010908	qtd	130	0,149927	brahm
31	0,001434	garraf	81	0,012186	c12	131	0,157051	269ml
32	0,001537	export	82	0,013157	transp	132	0,17251	lat
33	0,00154	12un	83	0,014174	amstel	133	0,219767	skol
34	0,001682	ipa	84	0,014216	300ml	134	0,220558	npal
35	0,001743	lta	85	0,015035	pa	135	0,23272	pilsen
36	0,001806	cerp	86	0,015484	lag	136	0,235348	1l
37	0,001824	tig	87	0,016033	sixp	137	0,26245	lt
38	0,001875	355ml-12	88	0,016208	cor	138	0,262738	vd
39	0,00194	24x350ml	89	0,01752	ml	139	0,267977	gfa
40	0,001949	cabar	90	0,017611	cart	140	0,279078	350ml
41	0,001983	nacional	91	0,01765	shrink	141	0,307126	sh
42	0,002023	pm	92	0,017718	crystal			
43	0,002165	gold	93	0,017795	np			
44	0,002253	24x355ml	94	0,018036	ret			
45	0,002431	473ml	95	0,018859	arte			
46	0,002471	patagon	96	0,019278	extra			
47	0,002567	color	97	0,019741	imperi			
48	0,00265	desc	98	0,020413	kais			
49	0,002682	glacial	99	0,020686	six-p			
50	0,002786	50l	100	0,021588	arto			

Fonte: Dados da pesquisa (2025).

Com 35 termos referentes às marcas das cervejas e 106 termos referentes aos atributos e complementos das marcas de cerveja, a lista apresenta uma redução de 69 termos se comparada com a lista de 210 termos da base NFC-*e*, entretanto, conta com 2 marcas de cerveja inéditas: **cabaré** e **lokal** que não estão apresentadas na Resolução Nº 0028/2023 da SEFAZ/AM. A seguir, a classificação dos termos conforme sua equivalência e qualificação, no quadro 15:

Quadro 15 - classificação dos termos selecionados na base NF-*e*

Classificação	Termos equivalentes
Nome do produto CERVEJA	cervej, cerv.
Quantidade do produto embalado (em ML)	740ml, 500ml, litr, 30ml, 473ml, 50l, 210ml, 275ml, 300ml, 355ml, 990ml, 330ml, 600ml, 269ml, 1lt, 350ml.
Unidade	ml, lt.

Quantidade do produto vendido (relação entre Quantidade embalada e Unidade)	24x330ml, 6pa, cx6, grf300ml, 12x1, 18un, 12un, 350ml-12, 24x350ml, 24x355ml, cx04, 24x600ml, 12x269ml, cx24, cxpap12, c12, cx12, c15, pc12.
Nome da marca (destaque em amarelo e azul)	car, des, baden, proib, cerp, tig, cabar, patagon, color, glacial, eisen, coronit, eisenbahn, lokal, becks, ita, petr, schin, devas, tiju, amstel, cor, crystal, imperi, kais, stell, heineken, budweis, spaten, original, itaip, bohem, antarct, brahm, skol.
Forma do recipiente do produto:	cxpap, shr, caix, garraf, lta, grf, barril, carta, six, fridg, sixp, cart, shrink, six-p, ow, one, way, ln, ne, long, ttc, sleek, lat, npal, vd, gfa, sh.
Características complementares:	appi, american, ultra, unfil, malzbi, trop, unf, weiss, descartave, premium, retorn, ale, export, ipa, nacional, pm, gold, desc, pull, off, pmalt, lne, keg, pil, subzer, qtd, transp, lag, np, ret, arte, extra, arto, pa, pur, chopp, dupl, malt, pilsen.
Nome do fabricante	amb.

Fonte: Dados da pesquisa (2025).

Uma análise mais detalhada dos termos que representam a marca da cerveja pode ser vista na figura 40 que apresenta uma comparação entre os termos da NF-*e* (coluna cinza) e NFC-*e* (coluna branca):

Figura 40 - Lista comparativa entre termos NF-*e* x NFC-*e*

1.ª amostra		2. ª amostra		3. ª amostra		4. ª amostra		5. ª amostra	
NF-*e*	NFC-*e*	NF-*e*	NFC-*e*	NF-*e*	NFC-*e*	NF-*e*	NFC-*e*	NF-*e*	NFC-*e*
amstel	amstel		bud	des		ita	ita	petr	petr
	antar	budweis	budweis	devas		itaip	itaip	proib	proib
	antart	cabar	cabar		devass		itaipav	schin	schin
antarct	antarct		caracu	eisen	eisen	kais	kais	skol	skol
	antarctic	car			eisenbah	lokal	lokal		spaten
	arto	cerp	cerp	eisenbahn	eisenbahn		munich		stel
baden	baden	color	color	glacial	glacial		orig	stell	stell
becks	becks	cor	cor	heineken	heineken		orign	tig	tig
bohem	bohem	coronit	coronit		imper	original	original	tiju	
brahm	brahm	crystal	crystal	imperi	imperi	patagon	patagon		tijuc

Fonte: Dados da pesquisa (2025).

Na comparação das amostras de NF-*e* e NFC-*e*, no mesmo período de 01/02/2023 a 31/05/2023, algumas diferenças foram encontradas no campo “nome da marca da cerveja”, considerando os itens faltantes na coluna NF-*e* (item em vermelho): na 1ª amostra, a marca “arto” de “Stella Artois; na 4ª amostra, a marca “munich” de “Munich” e na 5ª amostra, a marca “spaten” de “Spaten”, todas elas detectadas pelo radical nos termos minerados no cupom fiscal NFC-*e*. Entretanto, na amostra de NFC-*e* verificou-se que não foi possível minerar a marcas de cerveja: “des” de “Desperados” (item em azul).

Outra diferença entre as duas amostras é o número de variações na descrição de uma mesma marca de cerveja já que para a amostra de NFC-*e* é maior que para a NF-*e*. Por exemplo, a marca Antarctica na amostra NF-*e* apresenta uma única descrição “antarct”, radical para o nome da marca Antarctica, já na amostra de NFC-*e* foram detectados os seguintes radicais para descrições da marca: “antar”, “antart”, “antarct” e “antarctic”. Pressupõe-se que são erros ortográficos do cupom fiscal (NFC-*e*) quando da descrição do termo para emissão do documento, mas, para a construção do modelo e validação da marca, todas essas formas de representação são relevantes e irão ampliar o alcance da aplicação do modelo.

Além da análise quantitativa entre as duas bases, o propósito também seria a contribuição da descrição mais completa do produto na NF-*e* para que o modelo pudesse aprender os termos que compõem, em geral, a cerveja. Entretanto, é possível verificar que os requisitos arquiteturais e o ambiente de aprendizado da ontologia, utilizando o esforço e o conhecimento de Aprendizado de Máquina, proporcionaram uma extração e categorização mais abrangente de termos da amostra NFC-*e*, resultando na composição completa qualitativamente e quantitativamente de termos, com variações equivalentes em diversas formas e ainda alcançando um número maior de marcas, complementos, embalagem e pacote.

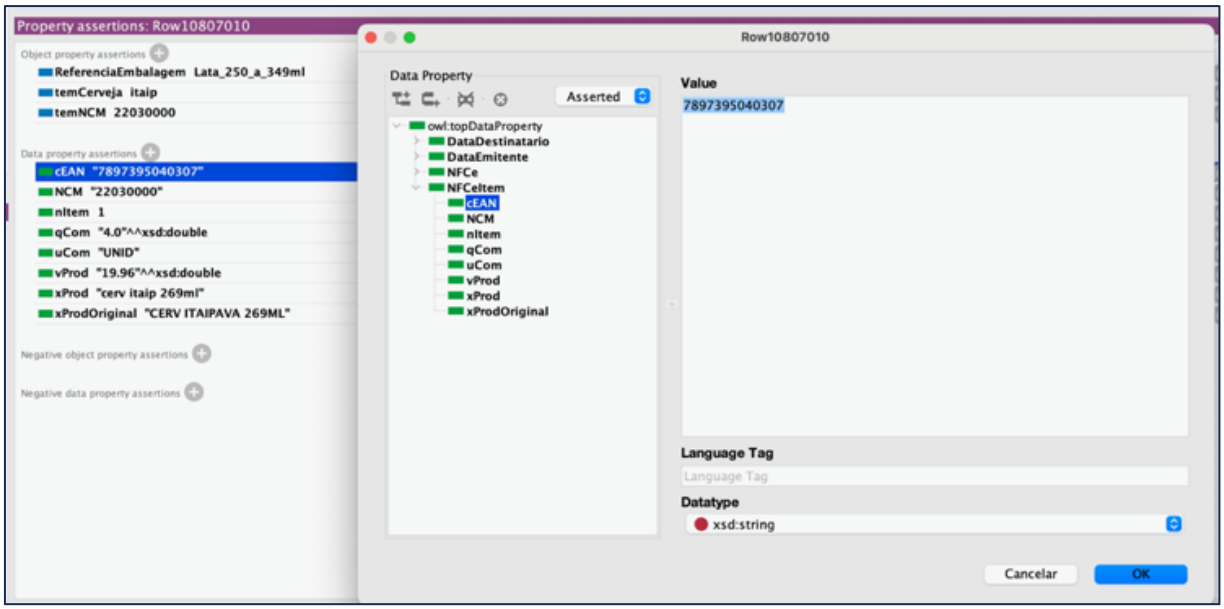
4.3.4.3 Base externa

Uma validação alternativa do produto cerveja também pode ser feita por meio do código GTIN, um campo estruturado, mas nem sempre presente no XML da nota fiscal (não obrigatório), mais popularmente conhecido como o código de barras que é usado como um identificador de produtos. Da amostra bruta recebida da SEFAZ/AM, 29,48% não estão com GTIN válido, ou seja, 13,68% das amostras apresentam no campo estruturado do GTIN a expressão “sem GTIN”, enquanto 15,80% das amostras não possuem qualquer informação (campo *null*). Mesmo sendo relevante para a identificação do produto e importante para a fiscalização, ainda não há uma legislação que proponha, exija e controle uma forma obrigatória de utilização do código GTIN para fins de auditoria.

Diante da instabilidade do campo GTIN, é perfeitamente compreensível que todos os processos de auditoria estejam baseados no campo do NCM. Entretanto, foi proposto um mecanismo de validação externa, alternativo para incrementar e ajudar

na identificação do produto, permitindo a visualização de como este produto cerveja está sendo comercializado no mercado e comparando se a forma de descrição no cupom fiscal é ou não equivalente. A figura 41 apresenta o detalhamento da linha *ROW* 10807010 com a descrição original do produto como “CERV ITAIPAVA 269ML” e destaque para o código GTIN com 13 caracteres (cEAN) com o valor 7897395040307.

Figura 41 - Detalhe da linha ROW 10807010 com destaque do código GTIN

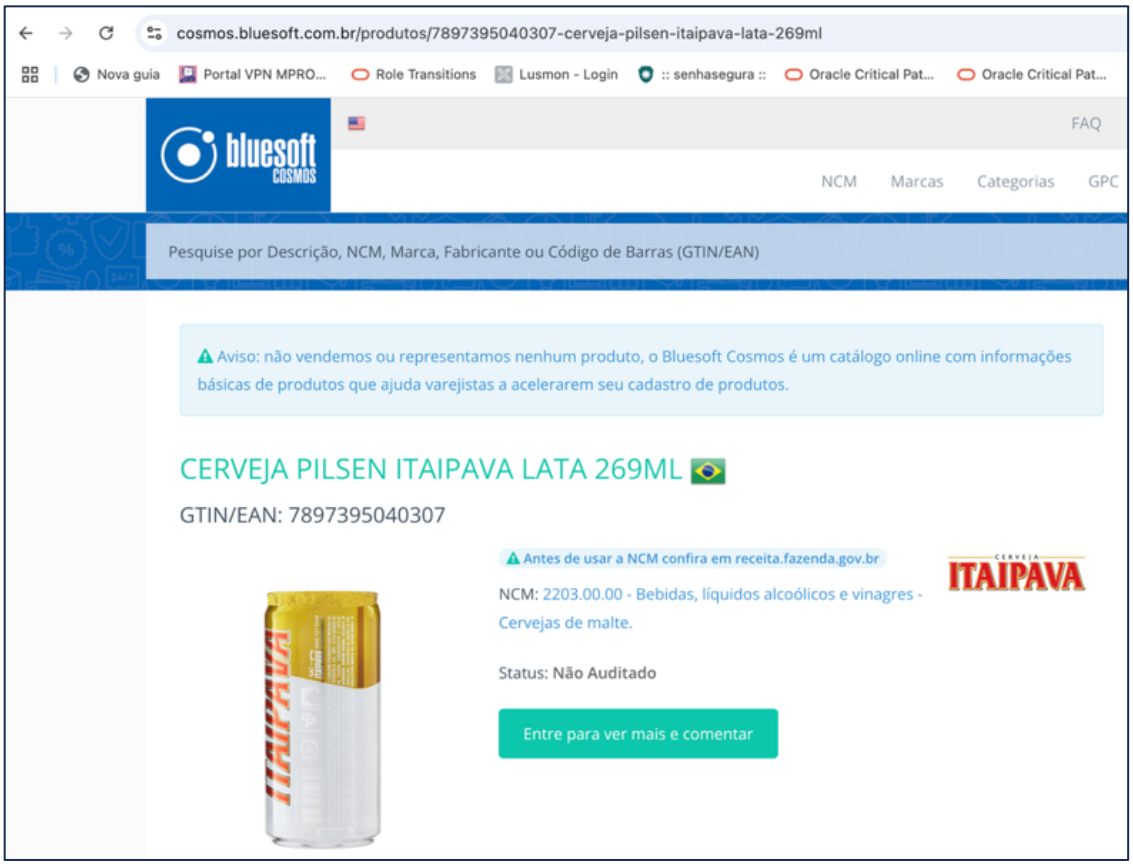


Fonte: Dados da pesquisa (2025).

A validação por meio do código GTIN permite a rastreabilidade do produto, um controle além da descrição que assegura a posição do produto na cadeia suprimento, o que seria essencial para p processo de auditoria de fiscalização e controle do estoque. Neste caso, para verificação e validação, foi selecionada uma busca no BlueSoft Cosrmos[10], que é um catálogo eletrônico de produtos usados para que fabricantes e atacadistas divulguem seus produtos e varejistas e clientes possam encontrá-los rapidamente. A busca, em geral, pode ser feita tanto por GTIN quanto por nome do produto ou até mesmo NCM. A figura 42 apresenta o resultado da pesquisa para o GTIN 7897395040307:

[10] https://cosmos.bluesoft.com.br

Figura 42 – Tela sítio eletrônico da SEFAZ/AM – consulta para o nome “cerveja”



Fonte: BlueSoft Cosmos (2025).

Na figura 42, é possível verificar que a descrição do produto é “CERVEJA PILSEN ITAIPAVA LATA 269ML”. Esta descrição, entretanto, é diferente da descrição da nota e acrescenta os termos: pilsen e lata, complemento do nome da marca e tipo de embalagem, respetivamente.

Apesar de não completar os termos na descrição original do produto, o modelo da ontologia 2203NFCe assegura, por meio das relações e associações dos *Object Property Assertions,* as classes pertencentes de cada termo da descrição original. Assegura, por exemplo, que a classificação de “269ml” é um volume associado ao tipo de embalagem “Lata”. Já no caso do complemento da marca da cerveja, “pilsen”, o modelo prevê, quando consultado, outros possíveis complementos associados a uma marca de cerveja, como já demonstrado na Questão de Competência 2.

Esta verificação do produto em um sítio eletrônico utilizado no comércio pode trazer segurança no processo de fiscalização da auditoria, inclusive transparência na definição e descrição do produto, pois é feito em uma plataforma pública e amplamente consultada, o que atende, por exemplo, aos requisitos legais exigidos nos tribunais nacionais quando da apreciação de litígios sobre processos de análises

fiscais e de auto de infração. Nestes casos, a SEFAZ/AM deve apresentar elementos concretos que possam comprovar ou contestar os fatos alegados. Assim sendo, a utilização do modelo para inovar a forma de representação e validação da informação contribui para a transparência do controle fiscal.

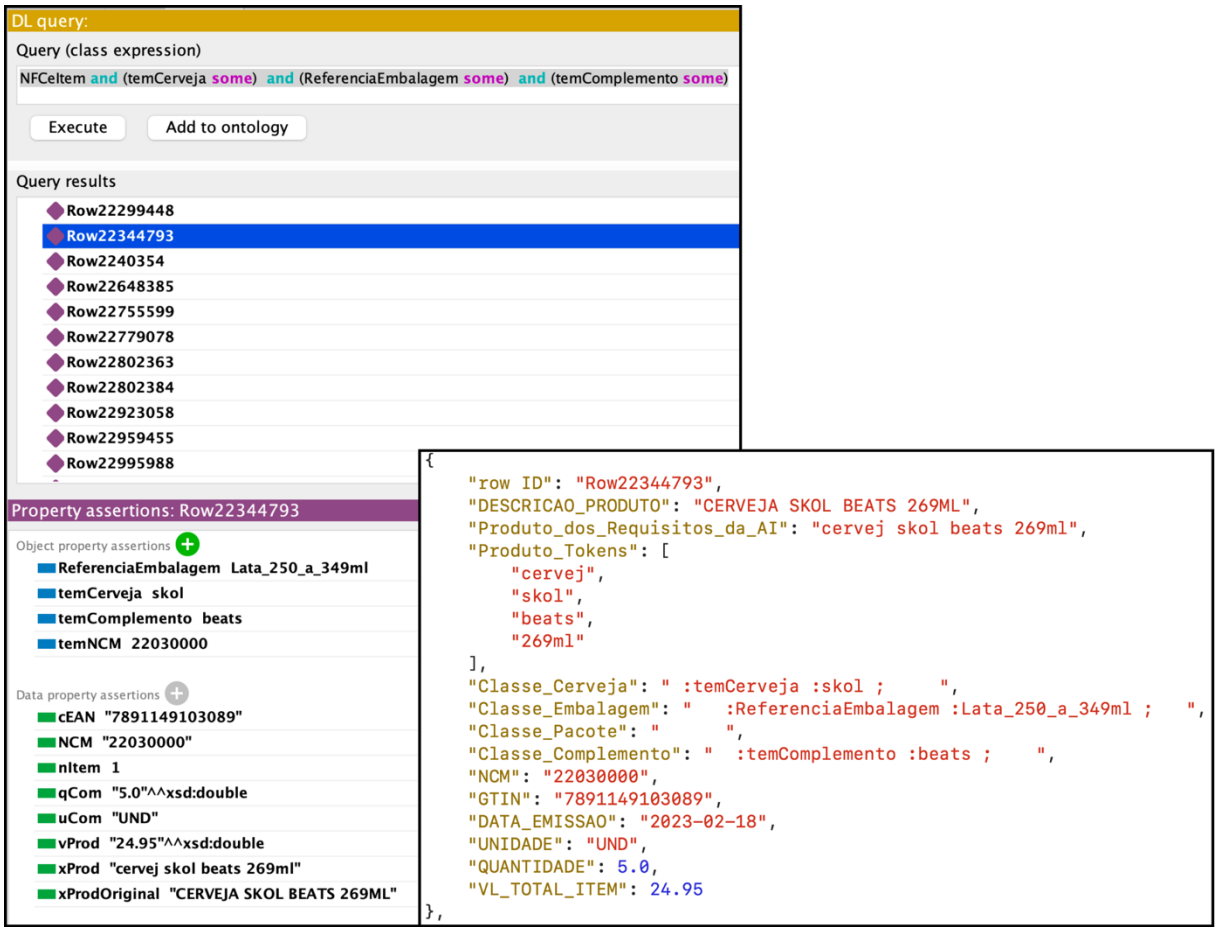
4.3.4.4 Repositório da ontologia 2203NFCe

Um dos grandes problemas de modelos de Aprendizado de Máquina é obter uma base de treinamento preparada e validada para o aprendizado. Sendo assim, o resultado da ontologia 2203NFCe também pode ser acessado e utilizado em uma base de dados consistentes, capaz de se tornar uma base de treinamento, em um formato de arquitetura aberta para aplicar com outras tecnologias.

O formato escolhido foi JSON por ser leve e ideal para dados semiestruturados, um formato bastante utilizado por ser compatível com linguagens da Ciência da Computação, produzindo um artefato com potência entre a Arquitetura da Informação e a Arquitetura de Sistemas.

A partir de classes, instâncias e relações validadas pela ontologia, foi gerada a base de treinamento em JSON com as 3.998.781 de linhas (*ROW*) e cada uma das linhas são desmembradas em: descrição original, produto dos requisitos da Arquitetura da Informação, tokens e classes e os *Data Properties* do cupom fiscal (NCM, GTIN, data de emissão, unidade, quantidade e valor total do item), resultando no produto muito maior que as 3.998.781 de linhas originais. A figura 43 apresenta como exemplo a mesma visão da instância dentro da ontologia e dentro do JSON:

Figura 43 – Detalhe da linha 22344793 na consulta no modelo de ontologia 2203NFCe e no código JSON



Fonte: Dados da pesquisa (2025).

O detalhe da linha ROW 22344793 na figura 43 é resultado da *DLQuery* “NFCeItem and (temCerveja some) and (ReferenciaEmbalagem some) and (temComplemento some)” cuja descrição original do produto é “CERVEJA SKOL BEATS 269ML”. O lado esquerdo da figura representa a informação gerada pela ontologia no Protégé e o lado direito da figura representa a mesma informação em formato JSON, um formato mais leve e acessível à outras tecnologias da Ciência da Computação e com todas as informações necessárias à compreensão e geração do conhecimento.

O Repositório dos dados[11] também acomoda outros arquivos JSON com as variações dos termos das classes, as várias formas de descrever o produto cerveja, e que vão auxiliar e incrementar aplicações com Aprendizado de Máquina. As classes escolhidas descrevem cada um dos termos equivalentes para: Marca da cerveja,

[11] O acesso aos códigos do repositório pode ser feito em: https://nfce2203-ontologia.ddns.net

Complemento da cerveja, Embalagem e Pacote. A figura 44 apresenta uma amostra dos arquivos das classes escolhidas:

Figura 44 – Amostra dos arquivos JSON para termos equivalentes das classes Marca da cerveja, Complemento da cerveja, Embalagem e Pacote

cerveja.json

```
{
    "Classe": "Antarctica",
    "Variant": [
        "ant",
        "anta",
        "antar",
        "antarc",
        "antarct",
        "antarctic",
        "antart"
    ]
},
```

complemento.json

```
{
    "Classe": "Lager",
    "Variant": [
        "lag",
        "lager"
    ]
},
{
    "Classe": "Malte",
    "Variant": [
        "ma",
        "malt"
    ]
},
```

pacote.json

```
{
    "Classe": "12x",
    "Variant": [
        "1000mlx12un",
        "12-350ml",
        "12269ml",
        "12lat",
        "12lats",
        "12pack",
        "12u",
        "12un",
        "12und",
        "12unid",
        "12x",
        "12x1l",
        "12x1lt",
        "12x250ml",
        "12x260ml",
        "12x269",
        "12x269m",
        "12x269ml",
        "12x275ml",
        "12x300ml",
        "12x310ml",
        "12x330ml",
        "12x350",
        "12x350m",
        "12x350ml",
```

embalagem.json

```
{
    "Classe": "Lata_350_ml",
    "Variant": [
        "12-350ml",
        "12x350",
        "12x350m",
        "12x350ml",
        "1x350ml",
        "1x350nl",
        "24x350ml",
        "350bohem",
        "350g",
        "350l",
        "350lat",
        "350m",
        "350ml",
        "350ml-c12lts",
        "350mlc12und",
        "350mlcr",
        "350mlcx",
```

Fonte: Dados da pesquisa (2025).

A figura 44 mostra um exemplo do detalhamento dos termos variantes das classes que ajudam no entendimento e identificação dos termos principal e atributos. Eles complementam o primeiro arquivo de detalhamento das linhas (*ROW*) de cada uma das transações apresentadas na amostra e que incluem, além das classes, as relações geradas entre elas, finalizando a entrega do produto da ontologia 2203NFCe.

5 RESULTADOS

Com base nos estudos apresentados, foi possível compreender os vários processos de auditoria, a seção **2.3 Bases teóricas da auditoria em Notas Fiscais Eletrônicas**, os apresenta destacando o esforço com que tratam a análise e fiscalização de produtos durante o seu lançamento (descrição) no corpo de documento fiscal, caracterizando o fato gerador da obrigação fiscal. Diante desse contexto da fiscalização vem a questão a ser respondida como problema de pesquisa: “de que forma o processo de auditoria em Documentos Fiscais Eletrônicos poderá ser mais produtivo com o acesso às informações sobre a descrição dos produtos das notas fiscais?”.

O estudo utilizou as amostras fornecidas pela SEFAZ/AM das bases de dados da NFC-*e* e NF-*e*, quadro 8 – Detalhe da amostra de NFC-*e* recebida da SEFAZ/AM e na seção *4.3.4.2 Base interna Nota Fiscal Eletrônica (NF-e)*, respectivamente.

As respostas ao problema de pesquisa foram construídas a partir de 3 Objetivos Específicos: OE1, OE2 e OE3, todos mencionados na seção **1.1.2 Objetivos**, e do Objetivo Geral (OG) qual seja: Auxiliar o processo de auditoria em Documentos Fiscais Eletrônicos utilizando as Notas Fiscais de Consumidor Eletrônica (NFC-*e*) por meio da elaboração de um modelo de ontologia a partir da Arquitetura da Informação e da Mineração de Texto **para validar a informação de descrição e venda do produto.**

Os objetivos específicos foram alcançados à medida que contribuíram para construção e validação do modelo de ontologia 2203NFCe. Na sequência, cada objetivo é reescrito, com os devidos indicativos sobre seu alcance:

OE01 - “Identificar possíveis requisitos de uma Arquitetura da Informação para a Mineração de Texto em Notas Fiscais de Consumidor Eletrônicas e no modelo de ontologia para cerveja”. O objetivo OE01 foi esclarecido e alcançado nas seções 2.1.3 Arquitetura da Informação; *2.1.3.1 Proposta de modelo de Arquitetura da Informação*; e **4.1 Requisitos Arquiteturais de Dados**. A revisão bibliográfica permitiu aportar ao referencial teórico os requisitos arquiteturais em Wurman (1997), que foram essenciais para a definição do trabalho de Mineração de Texto com o objetivo de “criar instruções para recuperar a informação a partir da necessidade do uso dos espaços organizados” (figura 11 - Ciclo para construção do conhecimento a partir da Arquitetura da Informação e da Mineração de Texto com Requisitos Arquiteturais de Dados e Requisitos de Recuperação de Dados) utilizando Localização, Alfabeto,

Tempo, Categoria e Hierarquia. Tudo isso para orientar a busca e a recuperação da informação. Como disciplina científica, a Arquitetura da Informação (Beira *et al.*, 2017) contribuiu para o "processo de modelagem de informações organizacionais na forma de ontologias" e permite direcionar os esforços para a estratégia de entrega dos dados para os usuários finais da SEFAZ/AM, quadro 6 – Legislações da SEFAZ/AM para Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS) e do produto cerveja de malte; e quadro 10 – Requisitos Arquiteturais de Dados da legislação e amostra NFC-*e* do produto cerveja para construção da ontologia.

Ademais, o modelo de Arquitetura da Informação (Orlandi, 2019; Gorayeb; Duque, 2022) entrega a resolução de problemas de busca e de organização da informação no âmbito da Ciência da Informação destacadas as linhas de estudo na organização da informação "aplicada a um conjunto de objetos informacionais"; de organização do conhecimento "visando à construção de modelos de mundo que se constituem em abstrações da realidade" na forma de "conceitos e suas relações semânticas"; de metadados que representam e identificam um objeto do mundo real de interesse do domínio, seção 2.1.2 Sistemas de Organização do Conhecimento e seção 2.2.5 Mineração de Dados e Mineração de Texto (Brascher; Cafe, 2008; Gottschalg; Vilela, 2018) e na ecologia da informação (Rosenfeld; Morville; Arango, 2015), expandindo entrega e disponibilidade integral da informação por meio de ferramentas da TIC da Ciência da Computação integrado à Arquitetura de Sistemas (ISO, 1998).

OE02 - "Definir as principais informações extraídas do produto quando aplicada a Mineração de Texto nas Notas Fiscais de Consumidor Eletrônicas": O alcance do objetivo encontra-se evidenciado na seção **4.1 Requisitos de Recuperação de Dados**; seção 4.3.2 Ambiente de Aprendizado da Ontologia para definição dos Requisitos de Recuperação de Dados e seção *4.3.2.2 Módulo de Mineração de Texto, Mineração de Dados e Aprendizado de Máquina*. A revisão bibliográfica, seção 2.2.4 Aprendizado de Máquina, consolida técnicas de Mineração de Dados, Mineração de Texto e raciocínio inteligente dentro do Aprendizado de Máquina para descobrir informações ocultas e associação de padrões desconhecidos ou regras de interesse para exploração de quantidades massivas de dados. O módulo de Mineração de Texto aplicado no Ambiente de Aprendizado de Ontologia "extrai conhecimento útil a partir do texto livre de descrição do produto combinando técnicas de Aprendizado de

Máquina (algoritmo Apriori, seção 2.2.3 Processamento de Linguagem Natural, figuras 12, 13, 14, 16, 17 e 18, para transformar o texto em dados analisáveis e estruturáveis (metadados, conceitos e relações semânticas), figuras 15, 19 e 20.

OE03 - "Descrever a relevância dos SOCs, especificamente, da ontologia para os processos de organização e recuperação da informação": O alcance do objetivo encontra-se evidenciado distribuído ao longo do texto. A abordagem para o alcance de OE03 descreveu as principais características intrínsecas da ontologia na seção 2.1.4 Arquitetura da Informação, como "uma representação formal e compartilhada do conhecimento em um domínio específico (Gruber, 1993), das fases do ciclo de vida compartilhada por várias metodologias largamente utilizadas em estudos e aplicações; quadro 2 – Ciclo de vida da ontologia; quadro 3 – Metodologias para construção de ontologias, utilizando um modelo híbrido que inclui Arquitetura da Informação, Ambiente de Aprendizado da Ontologia e a Engenharia de Ontologia ON-ODM (Haridy *et al.,* 2024) adaptada, seção **4.2 Modelo híbrido para construção da ontologia 2203NFCe**, figura 12 e seção **4.3 Aplicação do modelo híbrido para ontologia do produto cerveja na NFC-e**, e suas subseções.

Objetivo Geral: "Auxiliar o processo de auditoria em Documentos Fiscais Eletrônicos utilizando as Notas Fiscais de Consumidor Eletrônica (NFC-*e*) por meio da elaboração de um modelo de ontologia a partir da Arquitetura da Informação e da Mineração de Texto **para validar a informação de descrição e venda do produto**" foi atingido com 1) seção **4.2 Apresentação do modelo de ontologia 2203NFCe** (figura 12), 2) seção 4.3.4.1 **Questões de Competência** (respostas às 5 questões) e 3) seção 4.3.4 Validação da ontologia 2203NFCe, com a **produção do repositório de dados**, correspondente à uma base de dados treinada com linhas de código acessíveis e utilizáveis pela Ciência da Computação para qualquer aplicação automatizada.

Ressalta-se a apresentação **do modelo de ontologia 2203NFCe**, construído a partir da associação de módulos de venda e apresentação do produto e que permitiu a junção de dados estruturados e não estruturados, estes últimos por meio da obtenção e uso de metadados na descrição do produto cerveja, rompeu uma dificuldade de enxergar o produto e seus equivalentes dentro do campo "item" do cupom fiscal e validou as informações de produto e preço do produto entregues ao fisco estadual, mudando o foco de obtenção de dados para o resultado do Aprendizado de Máquina (classes, instâncias e relações).

Pode-se evidenciar que o modelo alcança a maturidade quando fornece dados íntegros e seguros sobre o produto cerveja na NFC*e*, destaque agora para o **preço unitário do produto cerveja** utilizado pela fiscalização para produção do PMPF. A técnica fiscal utilizada, para produção do PMPF e para projeção do imposto a ser recolhido em substituição tributária, é um processo já maduro nas Secretarias de Fazenda do Brasil. Estão regulamentadas e solidamente difundidas nos Estados, porém os insumos (obtenção, escolha e validação dos dados) para serem usados na técnica fiscal: valor unitário do produto e seu correspondente item do cupom fiscal e número da nota, emissor etc. são os grandes obstáculos para equacionar os problemas relacionados ao tamanho da amostra, tempo de produção do preço médio, transparência dos dados da amostra de segurança do processo. O preço unitário pode ser obtido pelo modelo de ontologia 2203NFCe por meio dos valores manipulados pela fórmula a seguir:

Equação do Preço Unitário Derivado do modelo de ontologia 2203NFCe:

$$vProdUnit = \frac{\left(\frac{vProd}{qCom}\right)}{qPacote}$$

Onde:

$vProd =$ valor total do item, obtido do *DataProperties* das instâncias da classe NFCeItem;
$qCom =$ quantidade comercial informada, obtida do *DataProperties* das instâncias da classe NFCeItem;
$qPacote =$ quantidade de itens no pacote, obtida nas instâncias da classe Pacote.

A classe Pacote determina a quantidade de itens no campo de descrição do produto, em cada linha ou “item” do cupom fiscal, associando-a à quantidade expressa no campo estruturado “quantidade” do cupom fiscal. Neste caso, o modelo de ontologia 2203NFCe oferece à auditoria duas quantidades diferentes e que não conseguem ser detectadas separadamente em um processo tradicional da Ciência da Computação que não contempla a categorização e os metadados da descrição do produto. Por exemplo, um *Data Mart* criado com base unicamente no campo

quantidade do XML do cupom fiscal não vai ter a acuracidade para fazer o cálculo da quantidade do produto vendido e consequentemente para o cálculo do preço unitário.

Um exemplo da aplicação da Equação do Preço Unitário Derivado pode ser extraído da figura 39 que descreve a resposta à Questão de Competência 2, cuja descrição do produto original da linha ROW 20732213 é: CEREJA ITAIPAVA PILSEN LT 269ML – 12X269ML.

Neste exemplo, os valores obtidos para os campos quantidade, valor total dos produtos e quantidade do pacote são:

$$qCom = 1.0$$

$$vProd = 27.48$$

$$qPacote = 12x$$

Assim, o valor do preço unitário da cerveja Itaipava Pilsen em Lata de 269ML é obtido por:

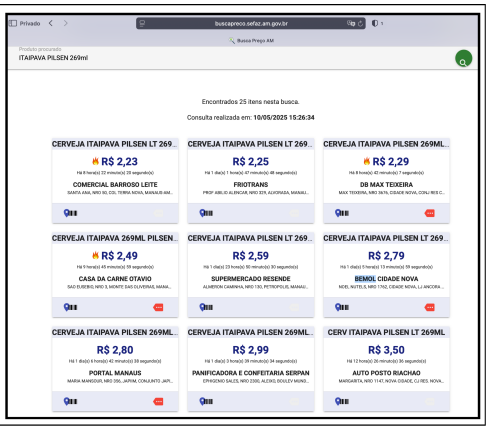
$$vProdUnit = \frac{\left(\frac{27.48}{1.0}\right)}{12} = 2.29$$

Quando comparando o valor unitário encontrado, R$ 2,29 (dois reais e vinte e nove centavos), com os valores discriminados na Resolução Nº 0028/2023 SEFAZ/AM e o sítio eletrônico da SEFAZ/AM para consulta de preços[12], é detectada uma variação nos valores identificados na figura 45:

[12] www.buscapreco.sefaz.am.gov.br

Figura 45 – Tela Busca preço SEFAZ/AM: variação de preços

4.	Lata 250 a 349ml		
4.1.	Amstel	Heineken	2,59
4.2.	Antarctica Pilsen	Ambev	2,68
4.3.	Antarctica Subzero	Ambev	2,52
4.4.	Bavaria	Heineken	1,24
4.5.	Bohemia Pilsen	Ambev	2,91
4.6.	Brahma Chopp	Ambev	2,75
4.7.	Brahma Duplo Malte	Ambev	3,10
4.8.	Budweiser	Ambev	3,22
4.9.	Cerpa Gold	Cerpa	2,46
4.10.	Cerpa Nevada	Cerpa	1,66
4.11.	Cerpa Tijuca	Cerpa	2,69
4.12.	Cerpa Tijuca Puro Malte	Cerpa	2,78
4.13.	Crystal Pilsen	Petrópolis	1,98
4.14.	Devassa	Brasil Kirin	2,36
4.15.	Heineken	Heineken	3,85
4.16.	Império Gold	Cidade Imperial	3,68
4.17.	Império Lager	Cidade Imperial	3,47
4.18.	Império Pilsen	Cidade Imperial	2,23
4.19.	Itaipava Fest	Petrópolis	2,58
	Itaipava Pilsen	Petrópolis	2,31



Fonte: Dados da pesquisa (2025).

O valor da cerveja Itaipava Pilsen Lata com volume variando entre 250 a 349ml na Resolução n.º 0028/2023 SEFAZ/AM (lado esquerdo da figura 44) é R$ 2,31 (dois reais e trinta e um centavos) e os nove primeiros valores da cerveja Itaipava Pilsen 269ml encontrados no endereço eletrônico do “busca preço da SEFAZ/AM” (lado direito da figura 44) variam entre o menor valor R$ 2,23 (dois reais e vinte e três centavos) e o maior valor R$ 3,50 (três reais e cinquenta centavos). Os valores obtidos originam-se em períodos diferentes e correspondem às informações diferentes. A amostra deste estudo corresponde aos meses de fevereiro, março, abril e maio do ano de 2023 e o valor calculado corresponde apenas a um preço unitário; a Resolução Nº 0028/2023 SEFAZ/AM é do final do ano de 2023 e corresponde ao preço médio PMPF da cerveja Itaipava Pilsen; a amostra para a consulta da busca do preço da SEFAZ/AM é atual, do ano de 2025, da base NFC-*e* da SEFAZ/AM dos últimos sete dias, classificados do menor para o maior valor unitário. Apesar das diferenças, é possível constatar que o modelo da ontologia 2203NFCe fornece insumos para a equação de preço unitário dentro de uma seleção de dados condizentes com a transparência, integridade e segurança necessárias à auditoria e produção do PMPF.

Também sobre o modelo de ontologia 2203NFCe, o seu alcance ou abrangência é consolidado ao permitir validar os mesmos dados do produto cerveja na base de dados de NF-*e*. Estes, em geral, usados no processo de comercialização da cerveja que vai do fabricante ao atacadista, ou atacadista para atacadista, ou seja, na cadeia de fornecimento de cerveja. Pelo fato da definição do XML das duas notas ser a mesma (para NFC-*e* e NF-*e*), o modelo de ontologia 2203NFCe, uma vez

aplicado, consegue extrair e validar o preço do consumidor final, também o preço de revenda na cadeia de suprimento, desde o fabricante até o consumidor final.

Quando avaliado sob o objetivo de “auxiliar o processo de auditoria em Documentos Fiscais Eletrônicos por meio da elaboração de um modelo de ontologia a partir da Arquitetura da Informação e da Mineração de Texto para validar a informação de descrição e venda do produto em notas fiscais”, o modelo apresentado pode não se limitar ao produto cerveja e ajudar na auditoria de outros produtos com habilidade de ser reaproveitado e adaptado para outras categorias de produto, por exemplo: categorias de bebidas destiladas como o vinho, categoria de construção como o cimento ou categoria de combustível. Todas essas são áreas sensíveis que influenciam na arrecadação de impostos no Estado do Amazonas e apresentadas pela SEFAZ/AM como áreas de estudo e vigilância constantes. Como exemplo de uma nova categoria, pode-se presumir que o combustível poderia se adaptar ao modelo, pois apresenta marca do combustível: diesel, gasolina, etanol e gás de cozinha (que poderia ser associado no modelo de ontologia ao nome da marca da cerveja) e tipo de combustível: aditivada, comum, S10 ou S500 etc. (que poderia ser associado no modelo de ontologia ao complemento da marca da cerveja). Os demais campos estruturados do cupom fiscal da venda de combustível poderiam ser associados facilmente ao módulo de venda do modelo de ontologia 2203NFCe, pois ele atende a qualquer produto comercializado.

As respostas às **Questões de Competência,** seção 4.3.4.1**,** retomam consultas realizadas por meio das *DLQueries* e integralizam um artefato que permite o aprimoramento das ações de promoção e controle fiscal na:

1. Construção da alíquota de PMPF para substituição tributária em produtos essenciais ao tesouro estadual que usa a descrição completa e padronizada com quantidade, unidade e valor do produto para calcular de forma precisa e confiável o ICMS por substituição tributária. As respostas às Questões de Competência 1 e 2 entregam a descrição completa do produto e a descrição completa da venda do produto (seção 4.3.4.1) e as variações das descrições que podem contribuir para ajustar o valor do imposto PMPF. Além desta, destacam-se abaixo outras implicações para a construção do PMPF:
    - Identificam corretamente o produto dentro de sua marca e categoria (ex: "cerveja skol beats lata 269ml”, ou “cerveja budweiser zero álcool, lata 350ml");

- Evitam agrupamentos incorretos que compõem a base de cálculo da substituição tributária e que podem distorcer o cálculo do preço médio.
- Diferenciam produtos com nomes semelhantes, porém apresentam marcas, embalagem e volume distintos;
- Garantem que apenas produtos equivalentes sejam usados na média ponderada, mesmo que descritos de forma abreviada ou com erros ortográficos ou por sinônimos etc.;
- Utilizam o preço por unidade apontado pelo modelo 2203NFCe e, muitas vezes, de difícil percepção na base de dados da SEFAZ/AM, visto que a quantidade real do produto vendida nem sempre está refletida no campo "quantidade do produto" no cupom fiscal, mas sim, na sua maioria, refletido na categoria Pacote no campo "descrição do produto; e
- Alcança a realidade de mercado para construção do preço médio favorecendo a transparência fiscal.

2. Acompanhamento da comercialização dos produtos: Controle de Estoque (Fiscalização) utiliza a descrição completa dos produtos com seus equivalentes (classes e instâncias cadastradas no modelo de ontologia 2203NFCe) e utiliza a quantidade de produto obtida por meio da classe Pacote. A resposta da Questão de Competência 1 consegue permitir, então, o rastreio do produto quando identifica uma entrada com uma quantidade "x" do produto na NF-*e* ou NFC-*e* (emitida para um CNPJ no papel de destinatário identificado) e identifica uma saída com quantidade "y" do produto (o mesmo CNPJ agora como emissor), conseguindo um cálculo aproximado entre entradas e saídas, independente da forma como a descrição do produto foi comercializada nas notas fiscais, enxergando potenciais distorções na composição do estoque, o que pode provocar o início do processo de fiscalização deste CNPJ.
3. Definição da alíquota para tributação por Margem de Valor Agregado: utiliza o campo estruturado do NCM para buscar os produtos na base de notas fiscais definindo um percentual fixo para alíquota de ICMS sobre o NCM do produto e, por vezes, pela descrição do produto. A Questão de Competência 4 é responsável por entregar a conformidade entre NCM da nota ou cupom fiscal e o produto descrito no campo "item". É importante destacar o papel da Questão de Competência ao responder e diferenciar a descrição do produto em conjunto

à identificação do NCM pura e simplesmente para a cobrança da alíquota Margem de Valor Agregado.

- Um exemplo hipotético seria: uma cerveja com complemento SUBZERO e outra com o complemento ZERO. A primeira com alíquota de 50%, pois refere-se à cerveja com teor alcoólico e refrigerada sob baixa temperatura; a segunda com alíquota de 20% refere-se à cerveja que não contém álcool. Para ambos os casos, o NCM é 2203.xxxx, neste caso o contribuinte destaca o valor da alíquota e cabe à Secretaria de Fazenda fiscalizar o ato gerador do imposto (descrição do produto na NFCe x NCM do cupom fiscal) e validar a arrecadação para o tesouro.
- O mesmo acontece com o Desembaraço fiscal para acompanhamento do trânsito dos produtos, que necessita do NCM para avaliar se os impostos sobre o produto estão corretamente destacados. Entretanto, o modelo de ontologia 2203NFce oferece um aprimoramento desta avaliação, pois apresenta uma dupla verificação, do NCM e do produto no campo “item”, independente da forma como o produto foi apresentado (Questão de Competência 3 e 4).

4. Análise fiscal para ações de verificação da regularidade fiscal, monitoramento dos setores econômicos e atualização e normatização tributária realizada com a ajuda da Questão de Competência 2 e 5 que oferecem monitoramento da venda do produto por diversas dimensões e permite, assim, acompanhamento de ações tributárias, por exemplo uma ação para redução de impostos por período para determinado programa social, ou para suprir um investimento do governo estadual etc.
5. Como ganhos adicionais, o modelo de ontologia 2203NFCe pode contribuir na pesquisa de preços de produtos, por exemplo, para compras da Administração pública respondida pelas Questão de Competência 1, 2 e 5. Quando a Administração Pública tem a descrição completa do produto também alcança o preço médio praticado no mercado, podendo associá-lo ao preço mínimo lançado em processos licitatórios, ao preço de referência coletado em fornecedores consultados. Quando da análise fiscal pela Questão de Competência 5 a admiração pública pode buscar os principais fornecedores que vendem em quantidade ou os que mais vendem, os que vendem com mais economia para realizar cotações oficiais e comparar com o preço médio. Como

resultado, a Administração pode trabalhar com a redução dos prazos de processos licitatórios, superfaturamento do preço ofertado etc.

Vale destacar que A produção do repositório, seção 4.3.4 Validação da ontologia 2203NFCe, é uma saída para que a Ciência da Computação utilize uma base de informação já validada e treinada. Neste caso, estes códigos disponibilizados podem auxiliar na fase computacional para incrementar os modelos e os algoritmos que se tornariam uma aplicação real sobre a base de dados da NFC-*e*.

Sabe-se que a Ciência da Computação utiliza técnicas como CRISP-DM (Shearer, 2000) para desenvolver aplicações que utilizem o AM e para preparar e treinar uma base confiável. Por isso, utilizar o repositório é reduzir a distância entre modelagem, desenvolvimento e implantação de uma aplicação que permitiria acesso direto às bases de dados de nota fiscal e cupom fiscal e que promoveria o acesso ilimitado, rápido e preciso às informações, colaborando para os processos de auditoria.

Para além disso, os tipos de códigos disponibilizados incluem não somente o detalhamento de cada transação de venda (item do cupom fiscal) como as classes envolvidas, as relações existentes entre as classes e os tokens, adicionada a todas as possibilidades identificadas pela ontologia 2203NFCe de variantes dos termos que descrevem o produto: Marca da cerveja, Complemento da cerveja, Embalagem e Pacote, figuras 39 e 40, assegurando que a descrição do produto é reflexo do praticado no mercado atual e que termos não explícitos podem ser inferidos a partir dos termos explícitos e restrições cadastradas na ontologia.

Outro destaque é conjunto de termos do repositório que pode ser reconhecido como um sistema de recomendação para a descrição do produto ou até mesmo como elemento de validação do produto com outras fontes porque trabalha com precisão do termo recomendado, em um ranking entregue pela Mineração de Texto e Aprendizado de Máquina (suporte, confiança da regra e *lift*).

Assim, o Repositório, por ser o artefato dentro de uma atividade pertinente à Ciência da Informação, integrado à Ciência da Computação por meio de Arquitetura da Informação e ontologia, é um meio robusto de organizar a informação a partir de metadados e distribuí-la para Arquitetura de Sistemas (Gorayeb; Duque, 2022) na criação de ambientes informacionais automatizados da gestão do fisco estadual. Como planejado em Gorayeb e Duque (2022), figura 3 - Modelo de arquitetura da informação para sistemas automatizados, a etapa de integração de TIC não limita o

modelo de ontologia 2203NFCe diante da interdependência entre a Ciência da Informação com a Ciência da Computação e da evolução tecnológica já incorporada à Ciência da Informação, em uma clara demonstração de que as duas ciências não convivem em modelos separados.

6 CONSIDERAÇÕES FINAIS

Esta seção apresenta a conclusão desta tese e, de modo sintético, responde à questão de pesquisa norteadora do trabalho, além de apresenta contribuições à área da Ciência da Informação e discorre sobre oportunidades futuras.

6.1 Conclusão

O papel da Auditoria, como é de conhecimento geral, é essencial para o controle da arrecadação dos tributos estaduais, para o desenvolvimento e o cumprimento da legislação tributária. Sua finalidade é garantir, na gestão fiscal, o desenvolvimento econômico da região, combatendo a sonegação, gerindo incentivos fiscais, créditos tributários, créditos presumidos e isenções fiscais no âmbito do imposto ICMS em DEF como as NF-*e* e as NFC-*e*.

Constantemente, ferramentas materiais são atualizadas como sistemas informacionais automatizados transacionais, *Bussines Intelligence* (BI) com implementação de *Data Marts* tradicionais (baseado em tabela fato e tabela dimensões, utilizando buscas com expressões “like”) e acesso à base de dados via linguagem SQL como esforço necessário ao desenvolvimento evolutivo do trabalho. Tais ferramentas oferecem um meio de acesso à informação e comunicação e facilitam a colaboração entre auditores, pois permitem a troca de conhecimento, compartilhamento de ideias e resolução de problemas, o que certamente contribui para o sucesso das ações fiscais, bem como para o aprimoramento da gestão tributária. O uso dessas ferramentas garante o acesso aos dados estruturados das notas fiscais e dos cupons fiscais, elementos centrais da verificação da conformidade tributária, proporcionando a comprovação das operações comerciais, os cálculos para tributos e o cruzamento de dados com outras bases de dados integradas à base fiscal.

Contudo, essas informações e ferramentas ainda são insuficientes para atender às necessidades dos auditores porque não alcançam todo potencial dos dados não estruturados do DEF, como os disponíveis na descrição do produto, um item essencial para a comparação das informações de escrituração e classificação tributária. O esforço utilizado com as técnicas computacionais atualmente utilizadas não oferecem aos auditores um retorno rápido e completo da informação de descrição do produto,

inviabilizando o relacionamento eficaz entre dados estruturados como NCM, GTIN, valor do produto e quantidade do produto, valor destacado do imposto etc.

Vale citar algumas dificuldades potencializam o problema de acesso à informação e o esforço para obtê-la no contexto do Amazonas, dentre elas: o pequeno contingente de auditores disponíveis no corpo de servidores das instituições, a alta carga de trabalho proveniente das ações fiscais e o custo tecnológico para movimentar o grande volume de dados. Esses são desafios que fazem com que a SEFAZ/AM, a exemplo de outras Secretarias de Fazenda dos Estados, não disponibilize especialistas tributários para trabalhar nas alternativas de soluções inovadoras e concentre as ações principalmente na regra 80/20, isto é, 80% da arrecadação dos Estados está concentrada em apenas 20% dos contribuintes (Não há uma lei específica ou um termo oficial, mas sim o Princípio de Pareto, uma teoria geral aplicada na área de finanças públicas que sugere a distribuição das ações sobre as fontes de impostos e tributos que geram a maior parte da receita estratégica). Nesta relação está a comercialização de produtos como: veículos automotores, combustíveis e bebidas, deixando à margem das ações todo o restante de produtos comercializados que ficam submetidos ao rigor da obrigação de fazer e ao controle da malha fiscal.

Soluções, ainda que tenham sido amplamente discutidas no setor público ou encontros técnicos com gestores, ou amadurecidas pelos integrantes das equipes de auditores e técnicos de computação, discutidas repetidas (duplicadas) vezes por seus interessados, não lograram êxito em avançar no campo da Ciência da Informação. Na área SOCs, a pouca disponibilidade do especialista e a baixa familiaridade dos aspectos técnicos, filosóficos e terminológicos de um artefato ontológico, frequentemente limitam a adoção de soluções que ofereçam, neste modelo de SOC, o compartilhamento de conhecimentos sobre a descrição do produto e o reuso de módulos de descrição de venda do produto como requisitos principais para seu desenvolvimento.

Dado que o problema desta pesquisa é de cunho prático da auditoria sobre a descrição de produtos em notas fiscais, utilizou-se um modelo híbrido de construção de ontologia para apoiar a produção de um artefato inovador com vistas aos problemas práticos que envolvem a mineração e classificação de termos do campo “descrição do produto”, utilizando como principal contribuição científica os requisitos da Arquitetura da Informação aplicados na Mineração de Texto para construção de

ontologia. A solução para o problema aqui abordado foi concretizada em um modelo com foco no domínio de comercialização de cerveja, que representou, de forma estruturada, os termos e atributos que a definem. Tal domínio descreve o produto com marca, complemento da marca, embalagem e pacote. Foi construída sem a presença de especialistas para auxiliar diferentes tipos processos de auditoria tributária e fiscal a partir dos dados disponibilizados da NFC-*e*.

De tal modo, o modelo de desenvolvimento se constituiu híbrido porque possui uma abordagem técnica e estruturada que permitiu entender, modelar, representar e organizar conceitos e suas relações dentro de um determinado domínio de conhecimento (comercialização do produto cerveja de malte), contando com 3 elementos oriundos da Ciência da Informação: Arquitetura da Informação; Ambiente de Aprendizado de Ontologia e uma adaptação da metodologia de construção de ontologias, a ON-ODM (Haridy *et al*., 2023).

Viabilizada a construção, foram utilizadas ferramentas de Mineração de Dados e Mineração de Texto, oriundas das áreas de recuperação de informação, e técnicas de PLN e algoritmos de aprendizado não supervisionado que forneceram os requisitos e os conceitos para as demais fases metodológicas de construção da ontologia: Análise de Requisitos, Conceitualização, Implementação, Enriquecimento e Validação.

As conclusões desta pesquisa são apresentadas a seguir, conforme os tipos de conhecimento investigados.

Quando se trata da **Arquitetura da Informação**, é possível justificar a utilidade no modelo destacando os seguintes pontos:

1. Permitiu alertar o cientista da informação sobre os pontos de vista arquiteturais para fontes informacionais, utilizando requisitos da Arquitetura da Informação de Localização, Alfabeto, Tempo, Categoria e Hierarquia, para explicar o método de aquisição e extração do conhecimento do domínio para uso no Ambiente de Aprendizado de Ontologia, preenchendo a lacuna existente quando a figura do especialista não está presente no ciclo de desenvolvimento para auxiliar os desenvolvedores de modelos de ontologia nos passos que devem ser executados para tais tarefas;

2. Permitiu orientar, com resultados promissores, as tarefas de classificação da informação obtida a partir dos requisitos de Arquitetura da Informação para a fase de Conceitualização, estabelecimento das relações, desde as mais relevantes até os

ruídos existentes (como dados que não possuem frequência ou densidade suficientes para gerar conceitos ou participar de um agrupamento), quebrando, assim, o paradigma da presença de especialista no processo de construção e tornando a utilização de ontologia, em especial no ambiente de auditoria, mais acessível e familiar aos profissionais do setor público.

3. Explicou formas de minerar os termos por classificação da frequência e pertinência ao domínio, servindo de ajuda na validação da ontologia desenvolvida;

4. Explicou formas de minerar relações entre termos e de que forma é possível caracterizar seu significado e importância para usá-los adequadamente na representação do domínio;

5. Tratou o enriquecimento dos termos e as relações através do uso de outras ontologias como um aspecto essencial, que inicia no processo de definição dos Requisitos Arquiteturais de Dados;

6. Tratou da apresentação do resultado do modelo, organizando meios de saída das informações como Repositório de códigos.

A respeito do **Ambiente de Aprendizado de Ontologia**, é possível justificar a utilidade no modelo destacando os seguintes pontos:

1. Permitiu que os Requisitos Arquiteturais de Dados fossem pré-processados com técnicas de PLN até a obtenção dos tokens para produção dos itemset utilizados no Aprendizado de Máquina;

2. Apresentou, com resultados promissores, a utilização de algoritmo aprendizado não supervisionado Apriori como minerador automático de texto para proposição dos Requisitos de Recuperação de Dados;

3. Determinou margens seguras para Suporte dos termos e Confiança, Lift e Convicção das regras para selecionar os elementos essenciais e eliminar os ruídos das amostras;

4. Apresentou o conceito de Termo principal como o termo que mais se repete (1.º lugar) e suas variações para a fase de Conceitualização na construção da ontologia;

5. Apresentou o ranking dos termos que mais se repetem, conceitualizando os termos atributos e termos complementares.

6. Apresentou a lista de associações mais frequentes, considerando as diversas formas de relações dos termos e estabelecendo as restrições para o modelo de ontologia;

7. Propôs os metadados a partir da equivalência dos termos, isto é, das variações dos termos, para a formatação de classes e instâncias do modelo de ontologia, determinantes para a descrição completa do produto cerveja, considerando as classes Embalagem e Pacote;

8. Propôs classes no módulo de venda do produto a partir da mineração de dados sobre os Requisitos Arquiteturais de Dados, permitindo o reuso de componentes para outros produtos e para outra base de dados de nota fiscal.

Quanto à adaptação das fases da **metodologia de construção da Ontologia**, tem-se:

1. Adaptou as fases tradicionais de construção da ontologia, amplamente divulgadas na literatura cientifica, em uma abordagem semiautomática, baseada em técnicas de PLN, técnicas de mineração otimizadas e algoritmo de ranque para apoiar a identificação de classes, os relacionamentos e as restrições, resultando em agilidade no processo de importação do código em formato *Turtle*, importação dos *Data Properties* e *Object Properties*;

2. Propôs módulos de descrição e venda do produto para favorecer o reuso do modelo ontológico para outros produtos ou outras bases de dados de notas fiscais como a NF-e;

3. Apresentou, na fase de Validação, respostas às Questões de Competência correspondentes às necessidades dos processos de auditoria de forma automatizada e baseada em raciocínio inteligente, com resultados potencialmente favoráveis, principalmente ao permitir a completa e importantíssima identificação do preço unitário do produto, permitindo concluir que o modelo de ontologia 2203NFCe entrega informação única suficiente para adicionar conhecimento aos usuários;

4. Permitiu a entrega dos dados validados da ontologia no Repositório em formato JSON utilizado tanto para armazenamento quanto para intercâmbio de dados. O formato de fácil interpretação por humanos e máquinas apresenta amplo suporte em várias linguagens de programação para ser utilizado em novas aplicações transacionais automáticas para auditoria.

6.1.1 Resposta da Questão de Pesquisa

Nesta seção, a questão de pesquisa desta tese será respondida. Sendo assim, tem-se:

De que forma o processo de auditoria em Documentos Fiscais Eletrônicos poderá ser mais produtivo com o acesso às informações sobre a descrição dos produtos das notas fiscais??

1. Leitura e interpretação corretas dos termos que descrevem o produto: os estudos realizados evidenciaram resultados positivos na utilização do modelo de ontologia 2203NFCe, mais especificamente com a Arquitetura da Informação, o Ambiente de Aprendizado da Ontologia, baseado na Mineração de Texto com algoritmos de Aprendizado de Máquina não supervisionado para organizar informação, tanto na entrada dos processos de construção da ontologia quanto na entrega final do modelo ontológico. No contexto desta pesquisa, os resultados apresentaram que a descrição de produto é relevante para a auditoria, quando viabiliza a extração dos termos da cerveja de forma qualificadora, classificada, completa e útil (aplicável em um processo de auditoria), permitindo a correta interpretação do produto descrito no cupom fiscal, hoje comprometida em virtude da ausência de termos ou dos erros encontrados na descrição da cerveja.

2. Associação e disponibilidade da informação: a disponibilidade dos dados validados pela ontologia vai além da entrega dos termos: nome da marca, complemento do nome da marca, embalagem e pacote, associados aos campos estruturados da venda do produto: NCM, GTIN, quantidade, unidade e valor da transação, e dos campos estruturados da identificação de emissor, destinatário e venda, ela promove também a entrega de variações desses termos, as diversas formas de apresentação e escrita do produto, associando-as a uma determinada classe como instâncias e oferecendo ao auditor uma correta e completa interpretação do produto quando a descrição original no cupom fiscal não oferece.

3. Obtenção do preço unitário: a classificação de cada um dos termos que compõem as sentenças de descrição do produto na ontologia e a introdução do conceito de Pacote e Embalagem no modelo permitiu a equivalência da ontologia com modelos de legislações utilizados, por exemplo, pela SEFAZ/AM e a extração do preço unitário e preço total do produto, uma das informações mais importantes para o campo da auditoria, quando associados ao NCM, pois parte dessas informações detalham todas as ações de acompanhamento, fiscalização e controle da arrecadação; regularidade, gestão e análise tributária; e contribuição com a administração pública.

4. Flexibilização e agilidade na construção de modelos de ontologia: os resultados também destacaram a importância dos Requisitos Arquiteturais de Dados

que permitiram evolução e agilidade na identificação, extração e organização da informação sem a figura do especialista, apontando uma forma de flexibilizar o paradigma do processo da construção e enriquecimento da ontologia. Assim, permitiu-se que o modelo híbrido diversifique e amplie o escopo dos produtos auditados, promova compartilhamento e reuso de conhecimento.

5. Continuidade: o reuso do conhecimento garante que, mesmo que novos atributos e termos qualificadores dos produtos sejam introduzidos no mercado e não identificados no modelo atual, a ontologia pode ser evoluída, bastando adicionar as classes e instâncias correspondentes aos novos termos, ou seja, o esforço inicial não é perdido, pois assegura-se a continuidade do modelo.

6. Redução de Riscos: o acesso ao conhecimento do modelo ontológico ajuda a evitar erros técnicos ou possíveis inconsistências na identificação do produto e consequentemente na fiscalização, pois assegura, na sequência de termos, o grau de pertinência de uma transação de venda à determinado produto ou marca.

7. Documentação: o conhecimento compartilhado pelo modelo ontológico documentado em código JSON no Repositório contribui diretamente para construção de novas soluções automatizadas para auditoria.

8. Novas ideias: expor algumas discussões e decisões praticadas na fase de desenvolvimento do modelo pode estimular ideias para encontrar novas formas de extrair conhecimento do domínio e explorar outras soluções eficazes.

Por outro lado, a não utilização de um modelo híbrido em projetos de desenvolvimento de ontologias, em especial para auxiliar na interpretação e validação de descrição de produtos de cupom fiscal, pode levar a alguns prejuízos no estímulo às mudanças na gestão fiscal e tributária. Primeiro, porque sem mudança na perspectiva das ferramentas utilizadas para tratar a informação de descrição dos produtos, incluindo soluções como SOCs, ocorre uma estagnação na extração do conhecimento, não há melhoria contínua ou evolução das técnicas que auxiliam o auditor, ficando a equipe limitada à informação recuperada da base transacional; segundo, porque, atualmente, não existe alta disponibilidade da equipe de auditores para trabalhar junto ao desenvolvimento tradicional de ontologia e isso impacta no tempo de desenvolvimento e na forma de acesso aos insumos utilizados na construção do modelo como as fontes de informação, transformação em conceitos, definição das regras; por último, a introdução de requisitos consolidados de Arquitetura da Informação e do Ambiente de Aprendizado de Ontologia aproxima o

modelo híbrido às novas ferramentas de interpretação de informação e tomada de decisão como *Large Language Models* (LLM) com possibilidades de produzir respostas cada vez mais especializadas sobre grandes volumes de dados de textos treinados.

Sendo assim, o modelo híbrido utilizado para construção da ontologia 2203NFCe surge como uma abordagem inovadora para auxiliar auditores na gestão do conhecimento do produto cerveja comercializados em NFC-*e*; muda o paradigma de acesso e classificação das informações ao utilizar requisitos de Arquitetura da Informação e o Ambiente de Aprendizado de Ontologia para dar agilidade e completude nos tipos de informações entregues; possibilita o desenvolvimento de novas ontologias de produto a partir o modelo 2203NFCe; e colabora de forma inequívoca para melhoria contínua da importância da Ciência da Informação aplicada na área de auditoria.

6.2 Contribuições

Esta seção apresenta as contribuições que esta pesquisa de doutorado trouxe para a área da Ciência da Informação. Ao longo desta pesquisa, foram identificadas e desenvolvidas diversas contribuições que ampliam o entendimento sobre o impacto dos Requisitos de Arquitetura da Informação nas técnicas de Mineração de Texto e Ambientes de Aprendizado de Ontologia para construção de modelos ontológicos. A seguir, são categorizadas e descritas as contribuições desta tese:

Contribuições teóricas, apresentadas a partir de evidências experimentais sobre:

1. Dificuldades de interpretação da descrição do produto e associação com o NCM do produto descrito no cupom fiscal;

2. Dificuldade de identificar o preço unitário do produto quando a venda se refere a quantidades de pacotes do produto;

3. Dificuldades técnicas, filosóficas e terminológicas de um artefato ontológico da área da Ciência da Informação para interpretar a informação e gerar conhecimento no campo da auditoria;

4. Pouca disponibilidade de especialistas fiscais e tributários para desenvolver soluções tecnológicas para extração, organização e apresentação de conhecimento;

5. Um estudo exploratório para identificar os tipos de recursos referenciados nos temas de Arquitetura da Informação e Ontologias e as razões para o compartilhamento de tais referências;

6. Como o uso de Requisitos Arquiteturais de Dados pode dar suporte às tarefas de levantamento de fontes informacionais, organização da informação para as ferramentas de Mineração de Texto;

7. Como o uso do Ambiente de Aprendizado de Ontologias, utilizando algoritmo de aprendizado não supervisionado, permite minerar o texto a partir da frequência, pertinência e ranque dos termos que descrevem o produto;

8. Como o uso de Metadados foi eficiente para classificar o texto e determinar conceitos não explícitos da descrição do produto e que serviram para criar classes inovadoras para ontologia e constituir relações entre dados estruturados e não estruturados que envolvem a descrição e venda do produto;

9. Como a construção de um artefato ontológico permitiu a extração de informação por meio das respostas às questões de competência sobre a descrição do produto, gerando conhecimento inovador para a área de auditoria fiscal.

Contribuições práticas:

1. Desenvolvimento de uma abordagem ontológica semiautomática, baseada em técnicas da Arquitetura da Informação com Mineração de Texto e do Ambiente de Aprendizado de Ontologias com algoritmos de seleção e ranque, para apoiar a identificação de termos e relações fundamentais que representam o produto cerveja e que não estão devidamente documentadas no campo descrição do produto na NFC*e*, sem a presença do especialista.

2. Desenvolvimento, dentro da ontologia 2203NFCe, de módulo de venda do produto e módulo de descrição do produto, com capacidade de reuso para outras ontologias de domínio em contexto organizacional.

3. Geração de um Repositório com códigos em formato JSON contendo informações validadas da descrição do produto cerveja com capacidade de aplicação imediata no treinamento de sistemas automatizados transacionais para incrementar as atuais ferramentas de auditoria.

4. Publicações aceitas e publicadas:

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, 2022. **Anais**

[...]. Porto Alegre, RS, 2022. Disponível em: 2023.https://enancib.ancib.org/index.php/enancib/xxiienancib/paper/view/870. Acesso em: 13 maio 2025.

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. DOI: 10.21728/p2p.2024v11n1e-7124. Acesso em: 13 maio 2025.

GORAYEB, D. M. da C.; DUQUE, C. G. Ontologia 2203NFC-e: Sentença completa do produto cerveja no campo descrição do produto da NFC-e (Ontologia 2203NFC-e: Frase completa do produto cerveja no campo descrição do produto da NFC-e). *In:* SEMINÁRIO DE PESQUISA ONTOLÓGICA NO BRASIL (ONTOBRAS), 17.; CONSÓRCIO DE DOUTORADO E MESTRADO EM ONTOLOGIAS (WTDO), 8., 2024. **Anais** [...]. Vitória, ES, CEUR-WS, 2024. Disponível em: https://ceur-ws.org/Vol-3905/paper8.pdf. Acesso em: 13 maio 2025.

5. Publicações aceitas e ainda não publicadas:

GORAYEB, D. M. da C.; DUQUE, C. G. Modelo de Arquitetura da Informação para Ontology Learning no ambiente de Nota Fiscal de Consumidor Eletrônica (NFC-e). Submetido para revisão na Revista Digital de Biblioteconomia e Ciência da Informação (RDBCI). No prelo.

6.3 Trabalhos Futuros

Embora esta tese tenha abordado a construção do modelo de ontologia 2203NFCe para identificação da descrição do produto cerveja na NFC-*e*, utilizando técnicas e ferramentas no campo da Ciência da Informação, ela também permite diversas outras oportunidades para futuras investigações e desenvolvimentos que podem expandir e aprofundar ainda mais o entendimento do tema. Nesta seção, são apresentadas algumas sugestões para trabalhos futuros, cada um deles com potencial para apoiar as tarefas realizadas por diferentes áreas de auditoria governamental.

Novas pesquisas podem focar no aperfeiçoamento e refinamento dos módulos da ontologia 2203NFCe, especialmente o módulo de descrição de venda para outra categoria de bebida como vinho ou um segmento diferente, como por exemplo a categoria combustíveis ou produtos de limpeza que possuem ambiguidades linguísticas, palavras e descrições com múltiplos significados como sabão, sabonete, sabão líquido, detergente etc., com falta de dados rotulados e os recursos semânticos limitados. Testes em contextos diferentes irão revestir o modelo ontológico de robustez, expandindo classes, relações e regras.

Há viabilidade para pesquisa que aborde a análise de usabilidade e impacto da ontologia 2203NFCe com os usuários, desenvolvendo uma interface para automatizar o pré-processamento de uma transação de venda (ou um arquivo com várias transações de venda simultâneas), permitindo a entrada na ontologia em uma camada multimodal, para que os auditores possam, ao final, validar os dados no modelo.

Outras pesquisas podem trabalhar no aprimoramento dos componentes aqui propostos. Automatizar o processo de busca de fontes de informação com buscas inteligentes e recomendadores baseados em conceitos do mesmo domínio de interesse, a partir de diferentes repositórios de dados que facilitem a captura do conhecimento, mas também apoiem a sua organização, documentação, compartilhamento e uso, em que as informações relevantes ao desenvolvimento da ontologia possam ser armazenadas e facilmente acessadas. Uma outra abordagem no aprimoramento dos componentes propostos seria acrescentar, no Ambiente de Aprendizado da Ontologia e além do algoritmo de aprendizado não supervisionado Apriori, um método de classificação/clusterização dos elementos(instâncias) das classes descobertas. Por exemplo, 12x269ml, 269mlcx12x e cx1269 são elementos da classe Pacote, todos eles repetem o conjunto de caracteres “12x” o que determina que existe um pacote com a seguinte interpretação: 12xcerveja (12 unidades de cerveja) e que esses termos podem ser classificados como similares. Neste caso, poderia ser utilizado um algoritmo de clusterização para classificar as diversas variantes, sem necessariamente depender da semântica. Este processo acrescentaria uma tarefa a mais para cada classe descoberta, mas poderia melhorar a assertividade e ajudar no processo automático de entrada na ontologia.

Por fim, acredita-se que novas pesquisas possam explorar o enriquecimento semântico com a inclusão do módulo de gestão e controle tributário com conceitos de contexto fiscal como: destaque de alíquotas, o valor do imposto, regras de isenções fiscais etc. melhorando e permitindo a geração de novas informações que poderiam validar a escrituração do produto no conteúdo da nota fiscal, apontando as divergências em relação às classes de descrição do produto e às regras tributárias e aos cálculo de ICMS devido, entre outras pautas fiscais, como cálculo da Substituição Tributária, créditos presumidos, créditos relativos à desistência de venda, preparação e comprovação da base de cálculo de Substituição Tributária, suspensão de lançamento de imposto, lançamento de obrigação fiscais acessórias, detecção de

omissão de valores e tantas outras atividades para ações de gestão e controle do fisco estadual.

Acredita-se ainda ser possível pesquisa sobre a evolução do modelo híbrido, explorando técnicas LLM sobre uma grande quantidade de dados de texto validados pela ontologia para análise de documentos em linguagem natural, criação de chatbots para respostas especializadas como explicação de tributos, verificação de inconsistências e interpretação de notas fiscais.

Como conclusão geral, pode-se afirmar que a investigação aqui conduzida representa uma importante contribuição para as áreas que tratam da Arquitetura da Informação e Organização e Representação do Conhecimento e que, de alguma forma, auxiliam em uma maior aproximação entre a Recuperação da Informação, Engenharia Ontológica e a Ciência da Informação, cujo futuro esperado seja na direção de ontologias como artefatos ágeis e totalmente acessíveis, que entreguem como resposta produtos cada vez mais aplicáveis em organizações e permitam o uso da informação nas mais diversas áreas do conhecimento humano.

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. CRCPress, New York, 2014.

AGRAWAL, R.; SRIKANT, R. **Fast algorithms for mining association rules**. *In:* VLDB '94: PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES. Santiago: Morgan Kaufmann Publishers Inc., 1994. p. 487–499.

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. 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**, Rio de Janeiro, 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**. 2010. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista "Júlio Mesquita Filho", Marília, 2010.

AMAZONAS. SEFAZ. **Decreto Nº 20.686, de 28 de dezembro de 1999.** APROVA o Regulamento do Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação - ICMS e dá outras providências. Manaus: SEFAZ, 1999. Disponível em: https://sistemas.sefaz.am.gov.br/get/Normas.do?metodo=viewDoc&uuidDoc=cc3888c0-e1b9-4433-b513-3c0f29cc625a. Acesso em: 13 jun. 2023.

AMAZONAS. SEFAZ. **Decreto Nº 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.

AMAZONAS. SEFAZ. **Lei Complementar Nº 19, de 29 de dezembro de 1997.** INSTITUI o Código Tributário do Estado do Amazonas e dá outras providências. Manaus: SEFAZ, 1997. Disponível em: https://sistemas.sefaz.am.gov.br/get/Normas.do?metodo=viewDoc&uuidDoc=ac5497ab-4fbd-4bc8-a51b-32405f61dc30#:~:text=L%20E%20I%3A,e%20do%20C%C3%B3digo%20Tribut%C3%A1rio%20Nacional. Acesso em: 13 jun. 2023.

AMAZONAS. SEFAZ. **Lei Complementar Nº 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%AAncias. Acesso em: 28 jul. 2023.

AMAZONAS. SEFAZ. **Resolução Nº 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.

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.

BARROS, L. A. **Curso básico de terminologia**. São Paulo: Edusp, 2004.

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.

BORGELT, Christian. **Efficient implementations of Apriori and Eclat**. *In:* WORKSHOP ON FREQUENT ITEM SET MINING IMPLEMENTATIONS (FIMI). Porto, 2005. Disponível em: https://borgelt.net/papers/efficient_apriori.pdf. Acesso em: 8 maio 2025.

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.

BRASCHER, M.; CAFE, L. Organização da Informação ou Organização do Conhecimento?. *In:* ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 9., 2008. **Anais** [...]. 2008.

BRASIL. **Lei Complementar Nº 87, de 13 de setembro de 1996.** Dispõe sobre o imposto dos Estados e do Distrito Federal sobre operações relativas à circulação de mercadorias e sobre prestações de serviços de transporte interestadual e intermunicipal e de comunicação, e dá outras providências. Brasília, DF: Presidência da República, 1996. Disponível em: https://legis.senado.leg.br/norma/572842/publicacao/15713990. Acesso em: 13 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: 4 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.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. “Ontology learning from text: An overview,” Ontol. **Learn. from text Methods**, Eval. Appl., vol. 123, p. 3–12, 2005.

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 Acção 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. 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.

CHOUKRI, D. A semantic-based variables selection for ontology learning taking jaccard alignment as case. **Procedia Computer Science**, 2014, n. 37, p. 56–63. DOI: https://doi.org/10.1016/j.procs.2014.08.012. Acesso em: 20 maio 2024.

CHOUKRI D. Using Hamming Similarity to Map Ontology Learning: A New Data Mining System. *In*: PROCEEDINGS of the 2013 Research in Adaptive and Convergent Systems, New York, USA: Association for Computing Machinery, 2013. p. 82–87. DOI: 10.1145/2513228.2513232.

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.

CONSELHO NACIONAL DE POLÍTICA FAZENDÁRIA**. Convênio ICMS 142/18, de 14 de dezembro de 2018.** Macapá: CONFAZ, 2018. Disponível em: https://www.confaz.fazenda.gov.br/legislacao/convenios/2018/CV142_18. Acesso em: 13 jun. 2023.

CONSELHO NACIONAL DE POLÍTICA FAZENDÁRIA. **Convênio ICMS 143/06.** Macapá: CONFAZ, 2006. Disponível em: https://www.confaz.fazenda.gov.br/legislacao/convenios/2006/CV143_06. Acesso em: 13 jun. 2023.

CONSELHO NACIONAL DE POLÍTICA FAZENDÁRIA. **Convênio ICMS 199, de 20 dezembro de 2010.** Brasília, DF: CONFAZ, 2010. Disponível em: https://www.confaz.fazenda.gov.br/legislacao/convenios/2010/CV199_10. Acesso em: 13 jun. 2023.

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.

DU, R.; AN. H.; WANG, K.; LIU, W. A short review for ontology learning: Stride to large language models trend. **ArXiv preprint,** 2404.14991, 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-SHAPIRO, 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**, 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.

FRIZON, G. A.; BAPTISTA, D. M. Indexação e representação: uma reflexão diante das novas tipologias documentais. *In*: BAPTISTA, D. M.; ARAÚJO JUNIOR, H. (org). **Organização da informação**: abordagens e práticas. Brasília, DF: Thesaurus Editora de Brasília Ltda, 2015. p. 159-187.

FROSSARD, D. ICMS Genérico. Rio de Janeiro, Editora Ferreira, 2011.
GUARINO, N. Formal Ontology and Information Systems. *In*: PROCEEDINGS of FOIS. [*S.l.*]: IOS Press, 1998.

GARSHOL, L. M. Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all. **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.

GHOZI, M. R. N.; DJUNAIDY, A.; VINARTI, R. A.; RAKHMAWATI, N.A. Semi-Automatic Ontology Generation for Infectious Disease Domain from Text Data. *In:* INTERNATIONAL CONFERENCE ON INFORMATION & COMMUNICATION TECHNOLOGY AND SYSTEM (ICTS), 14., 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. Diposnível em: 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, 2022. **Anais** [...]. Porto Alegre, RS, 2022. Disponível em: 2023.https://enancib.ancib.org/index.php/enancib/xxiienancib/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.

GOTTSCHALG, C.D.; VILELA, P. J. Ontologias: um tipo único de sistema de organização do conhecimento. *In:* ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 2018. **Anais** [...]. Londrina, PR, 2018.

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.

GHIDALIA, S. *et al*. Combining machine learning and ontology: A systematic literature review. **ArXiv preprint,** 2401.07744, 2024.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3. ed. Amsterdam: Elsevier, 2012.

HAHSLER, M. *et al*. The arules R-package – A computational environment for mining association rules and frequent item sets. **Journal of Statistical Software**, [*S. l.],* v.

14, n. 10, 2005. Disponível em: https://www.jstatsoft.org/article/view/v014i10. Acesso em: 8 maio 2025.

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**, [*S. l.*], ano 7, v. 101, n. 2, p. 1-23, 2023. DOI 10.3390/bdcc7020101. Acesso em: 7 maio 2023.

HASSAN, B. A.; RASHID, T. A. Artificial intelligence algorithms for natural language processing and the semantic web ontology learning. **ArXiv preprint**, 2108.13772, 2021.

HAMOUDA, I.; CHOURABI, O.; BOUGHZALA, I. An ontological framework for knowledge sharing in healthcare. **Gestión Del Conocimiento**, [*S. l.*], n. 14, v. 26, p. 123–132, 2016. Disponível em: https://doi.org/10.16925/me.v14i26.1620. 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, n. 1, p. 45-58, 2009. Disponível em: https://journalofia.org/volume1/issue1/04-hinton/jofia-0101-04-hinton.pdf. Acesso em: 13 maio 2023.

HJØRLAND, B. Theories are Knowledge Organizing Systems (KOS). **Knowledge Organization**, [*S. l.*], n. 42, v. 2, p.113-128, 2015.

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

INSTITUTO RUI BARBOSA. **Normas de Auditoria Governamental (NAGS)**. Tocantins: IRB, 2011. Disponível em: https://irbcontas.org.br/biblioteca/normas-de-auditoria-governamental-nags/. Acesso em: 13 jun. 2023.

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.

ISO. **ISO/IEC 11179-3:2013. Information technology** — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes. 1 ed. Genebra: ISO, 2013.

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.

KORST, W. N. **Construction of Engineering Ontologies for Knowledge Sharing**. Twente: Center for Telematics and information Technology (CTIT), 1997.

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.

LARA, M. L. G. de. Diferenças conceituais sobre termos e definições e implicações na organização da linguagem documentária. **Ciência da informação**, Brasília, v.33, n.2, p. 91-96, 2004.

LARA, M. L. G. de. de. Linguagem documentária e terminológica. **Transinformação**, Campinas, v. 16, n. 3, p. 231-240, 2004.

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.

LISI, F. A. Building Rules on Top of Ontologies for the Semantic Web with Inductive Logic Programming. **Theory and Practice of Logic Programming**, v. 8, n. 3. 2007.

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.

MAEDCHE, A.; STAAB, S. Discovering Conceptual Relations from Text. *In*: Proceedings of the European Conference on Artificial Intelligence, 14., W. Horn, **Anais** [...]. IOS Press, 2000, p. 321–325.

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. 2000.

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 (HEMONTO). 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.** Texas, USA: IEEE Frontiers in Education Conference (FIE); College Station, 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.

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

NOVAES, D. Reflexões linguísticas para a organização hierárquica de conceitos em tesauros. *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/infa.v17i2.1056. Acesso em: 5 jul. 2023.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **Transformando nosso mundo: a Agenda 2030 para o Desenvolvimento Sustentável.** Brasília: Nações Unidas, 2015. Disponível em: https://brasil.un.org/pt-br/91863-agenda-2030-para-o-desenvolvimento-sustent%C3%A1vel. Acesso em: 26 ago. 2025.

ORLANDI, T. R. C. **Um modelo de Arquitetura da Informação, apoiado pela multimodalidade, para capacitação de profissionais de alto desempenho**. 2019. 215 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, DF, 2019.

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/biti/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.

INTERNATIONAL CONFERENCE ON COMPUTER AND COMMUNICATION SYSTEMS (ICCCS), 4., Singapura, 2019. **Anais** [...]. Singapura, 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. **Lecture Notes in Computer Science**, v. 12731. p 647–663, 2021.

SCHUNEIDER, L. F. **Mineração de Dados** - Conceitos. Porto alegre: Universidade Federal do Rio Grande do Sul, 2002.

SEFAZ/AM. Guia Prático Escrituração Fiscal Digital - EFD – Versão 2.0.4. 2011. 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.

SHEARER, C. The New Blueprint for Data Mining. **Journal of Data Warehousing**, v. 5, n. 4, p. 13-22, 2000.

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** […]. Florianopolis, 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.

VIZCAINO, A.; RUIZ, F.; PIATTINI, M.; GARCIA, F. Using REFSENO to represent knowledge in the software maintenance process. Proceedings. International Workshop on Database and Expert Systems Applications, 15. **Anais** [...]. Espanha, p. 488-493, 2004.

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 Syste ms Journal**, [*S.l.*], v. 38, n. 2-3, p. 454-470 1999.

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

ZOUAQ, A.; GASEVIC, D.; HATAL, M. Towards open ontology learning and filtering. **Information Systems**, [*S.l.*], 2011, n. 36, v. 7, p. 1064–1081.

ANEXO A – SOLICITAÇÃO PARA COLABORAÇÃO NA PESQUISA

Ao Ilustríssimo Senhor,

Secretario 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 lídimo 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 analise 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

AMAZONAS
GOVERNO DO ESTADO

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

Assinado digitalmente por: LUIZ AURELIO CARVALHO LEITE em 27/06/2022 às 11:01:57 conforme MP no- 2.200-2 de 24/08/2001. Verificador: 9A4F.7820.FDAB.8FD2
Assinado digitalmente por: ALAN CESAR MONTEIRO CORREA em 27/06/2022 às 09:14:09 conforme MP no- 2.200-2 de 24/08/2001. Verificador: 9A4F.7820.FDAB.8FD2

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

AMAZONAS
GOVERNO DO ESTADO

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

Ao GSEFAZ,

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.

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.

Gabinete do Secretário Executivo da Receita, em Manaus, 29 de agosto de 2022.

DARIO JOSÉ BRAGA PAIM
Secretário Executivo da Receita

Assinado digitalmente por: DARIO JOSE BRAGA PAIM em 01/09/2022 às 10:13:59 conforme MP no- 2.200-2 de 24/08/2001. Verificador: C59D.737D.9ACD.0EF3

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

Secretaria de Fazenda