



XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT-2 – Organização e Representação do Conhecimento

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

PLANNING OF AN INFORMATIONAL ENVIRONMENT AUTOMATED FOR THE EXTRACTION OF RELEVANT TERMS TO THE OVERSIGHT IN ELECTRONIC INVOICE AND ELECTRONIC CONSUMER INVOICE

Diana Maria da Camara Gorayeb. UNB.

Cláudio Duque Gottschalg. UNB.

Modalidade: Trabalho Completo

Resumo: o planejamento de um modelo de arquitetura da informação para sistemas automatizados busca identificar termos e atributos qualificadores dos produtos descritos em documentos eletrônicos de Notas Fiscais Eletrônicas e Notas Fiscais de Consumidor Eletrônicas e apoiar a gestão do fisco. Utiliza a interface da ciência da informação com a ciência da computação reunindo um modelo de arquitetura da informação e um modelo de referência de arquitetura de sistemas distribuídos *Reference Model of Open Distributed Processing* para a criação de espaços informacionais. A integração inclui a etapa de tecnologia da informação e comunicação e utiliza um modelo ontológico para descrever processos, entidades e itens de dados no contexto de normas e regras do fisco, da compreensão do usuário, da definição dos termos de interesse, obtenção dos dados em um Data Lake e aplicação de algoritmos inteligentes, com a validação e criação de um sistema de aplicação e geração de resultados para a apresentação da informação.

Palavras-Chave: Ciência da informação. Arquitetura de informação. Ontologia. Ciência da Computação. Data Lake.

Abstract: *the planning of an informational architecture model for automated systems seeks to identify terms and qualifying attributes of the products described in electronic documents of Electronic Invoices and Electronic Consumer Invoices and to support tax management. It uses the interface of information science with computer science, bringing together an information architecture model and a reference model of architecture of distributed systems Reference Model of Open Distributed Processing for the creation of informational spaces. The integration includes the information and communication technology stages and uses an ontological model to describe processes, entities and data items in the context of tax rules and regulations, user understanding, definition of terms of interest, obtaining data in a Data Lake and application of intelligent algorithms, with the validity and creation of a system of application and generation of results for the presentation of the information.*

Keywords: *Information Science. Information Architecture. Ontology. Computer Science. Data Lake.*



1 INTRODUÇÃO

Muito se discute a representação, a organização e recuperação de informação em ambientes informacionais e a necessidade de prover recursos para elaborar de atividades de gestão, melhorar a qualidade do trabalho e atender aos recursos humanos com a promoção de canais eficientes de comunicação da informação. A tecnologia é sem dúvida um recurso poderoso para atender essas expectativas pois permite a representação por meio digital de um objeto ou fato real utilizando de atributos e de suas relações para a descrição desses objetos. As representações da informação, segundo Moraes e Arcello (2000, p. 8), “são instrumentos de ordenação e hierarquização da estrutura social e identificam o grupo ou meio que as produziu e que as consome” assim, a tecnologia se utilizará da visão do meio, em determinado contexto e conteúdo para representar informações e prover conhecimento.

Com especial interesse sobre documentos eletrônicos, a área fazendária estadual utiliza a Nota Fiscal Eletrônica (NF-e) e a Nota Fiscal de Consumidor Eletrônica (NFC-e) para acompanhar movimentação de produtos e serviços, definir a alíquota do Imposto sobre a Circulação de Mercadorias e Serviços (ICMS) a ser aplicada sobre uma operação comercial, definir o cálculo do Preço Médio Ponderado ao Consumidor Final (PMPF) utilizado para o cálculo do imposto sobre mercadorias em regime de substituição tributária ou mesmo buscar os menores preços praticados no mercado para determinados produtos para balizamento de preço em licitações governamentais. Entretanto, o desafio no documento fiscal eletrônico é identificar o conteúdo nos campos não estruturados buscando na descrição o nome e atributos do produto e que por muitas vezes se confundem na marca, tipo e aspectos como tamanho, cor, textura, embalagem e volumetria, inviabilizando a rastreabilidade do produto na cadeia de venda, prejudicando o controle da arrecadação, correta tributação e ganho do tesouro e ainda superar problemas identificados como: volume e o grande número de documentos e imensa massa de dados a serem analisados; inconsistência e a falta de metadados e informações relacionadas com o documento e com o produto; e complexidade e a existência de polissemia e falta de padronização da terminologia para a categorização na descrição dos produtos (nomes e atributos) com abreviações e descrições ambíguas, o que demanda esforço para compreensão e controle da informação. Para superar esses problemas é preciso representar e organizar a informação e utilizar ferramentas para recuperá-la.



Nestes complexos ambientes, segundo Evaristo e Duque (2011), a representação da informação, por meio de um processo ontológico, oferece uma estrutura conceitual resultante de extensa análise de linguagem natural apresentando uma resposta mais eficiente às consultas de usuários pois visa, principalmente, aumentar a precisão no processo de busca em sistemas de recuperação da informação. Recuperação da informação (RI), segundo Gonzalez e Lima (2003, p. 18), trata da “indexação, busca e classificação de documentos e textos com o objetivo de satisfazer necessidades de informação de seus usuários expressas através de consultas”. Para os autores são componentes iniciais da RI: o usuário, a necessidade de informação, o conjunto de dados disponíveis, a estratégia de consulta e a base de conhecimento acumulado que servirão para a retroalimentação de novas consultas, e de fatores determinantes como: volume, qualidade e diversidade dos dados.

O objetivo deste trabalho é apresentar, dentro da área aplicada da Ciência da Informação (CI) que estuda a RI em sistemas, o planejamento de um espaço informacional automatizado para identificar, organizar e recuperar termos e atributos de produtos das notas fiscais eletrônicas relevantes à gestão do fisco estadual. Para alcançar este objetivo será necessário: prover um modelo integrado que inclui AI e arquitetura de sistemas distribuídos; descrever as atividades de cada etapa do modelo integrado; e produzir um modelo ontológico no ponto de integração entre a AI e arquitetura de sistemas distribuídos para identificar, inequivocamente, o conjunto de termos e atributos dos produtos dos documentos eletrônicos.

2 CIÊNCIA DA INFORMAÇÃO E A ORGANIZAÇÃO DA INFORMAÇÃO E DO CONHECIMENTO

Para Saracevic (1991) a CI se dedica as questões científicas voltadas para os problemas de comunicação do conhecimento e de seus registros entre os seres humanos no contexto social, institucional ou individual do uso e das necessidades de informação. Zins (2007) afirma que CI é o estudo das perspectivas mediadoras do conhecimento, incluindo aspectos e condições cognitivas, tecnológicas e sociais que facilitem a disseminação do conhecimento humano, desde o gerador até o usuário. Atualmente, a informação está associada ao volume de conteúdo disponível em aplicações tecnológicas, principalmente na *web* para Cedro e Duque (2020, p. 4) “tem-se potencializada essa necessidade de compartilhamento, interatividade e criação de informações, possibilitando o individuo ser além de receptor, o gerador de informações”.



É no contexto desta ciência que se encontra a necessidade de representação da informação e a interação com aspectos como a necessidade de organizar, armazenar e recuperação a informação, superando os problemas de imprecisão e intempetividade na comunicação.

Na Organização da Informação (OI) a análise da informação está associada ao objeto e inclui a descrição física, representação e organização do conteúdo, do suporte e da linguagem em determinado domínio do conhecimento para fins de armazenamento e recuperação da informação. Por outro lado, Organização do Conhecimento (OC) “visa à construção de modelos conceituais do mundo que se constituem em abstrações da realidade” (BRASCHER, CAFÉ, 2008, p. 6). A representação, segundo Zins (2007), descreve e reproduz a informação, que é um conjunto de símbolos que representam o conhecimento, o conhecimento por sua vez reproduz os pensamentos que o indivíduo acredita como verdadeiros e que serão documentados manual ou eletronicamente, por meio de linguagens naturais, artificiais, registros textuais, sonoros, imagens, suportes manuscritos, impressos e eletrônicos e “materializam-se por metadados relacionados ao objeto” (VICTORINO, PINHEIRO, SANTOS, 2015, p. 235).

O Processo de Organização da Informação (POI) 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 documentos e publicações que demandam linguagens sofisticadas para armazenamento, recuperação e comunicação mais precisos, rápidos e fidedignos. Para os autores Frizon e Baptista (2015, p. 166) as linguagens “servem à representação dos documentos funcionando como mediação entre a informação, em sua expressão original, e o usuário, em sua busca por dados, informação e conteúdos específicos de sua área de interesse”. Contando com suporte multidisciplinar de lógica, linguística e computação, linguagens sofisticadas resultam em um Processo de Organização do Conhecimento (POC) e, segundo HJØRLAND (2012), inclui atividades como sistemas de catalogação, análise de assunto, indexação e encontram-se em diferentes em Sistemas de Organização do Conhecimento (SOC) como tesouros e ontologias e outros sistemas de metadados.



2.1 Sistemas de Organização do Conhecimento (SOC)

Ontologias, segundo Duque (2005) são conjuntos de asserções, afirmações categóricas, que definem as relações entre conceitos e estabelecem regras lógicas de raciocínio sobre eles. Para o autor as ontologias permitem a geração de índice da coleção de documentos, aumentam a qualidade quando empregadas em um sistema de classificação, expandem as consultas dos usuários quando utilizadas na recuperação da informação e em conjunto com a IA permitem o intercâmbio dos dados gerados e editados por um modelo em outro modelo distinto. Em Mori (2009, p. 62) ontologia relaciona conceitos representados formal e consensualmente dentro de um determinado domínio, “são tecnologias importantes, quando bem empregadas, para a redução do impacto da variação terminológica”. Gonzalez (2010, p. 25) define ontologia como: “conceitualização formal de um domínio ou de uma parcela de realidade com a qual podem operar diferentes aplicações de software”. A conceitualização serve como vocabulário que permite que o recurso informacional seja utilizado e reutilizado, interoperável e aplicado por motores de busca de palavras-chaves, ou seja, com termos essenciais e que representam uma área de conhecimento para a RI.

Segundo Gruber (1995) a ontologia especifica explicitamente o conceito, são ideias básicas que serão formalizadas por meio de classes de objetos, instâncias, atributos, métodos, planos, estratégias, processos de raciocínio e relações que podem existir entre os conceitos e coisas como subclasse-de e parte-de, funções e axiomas definidos sobre estes conceitos e suas relações. Neste contexto a representação e organização da informação identificam quais elementos são essenciais, quais suas características e seus relacionamentos que representam os objetos informacionais, quais seus modificadores e conectivos, quais são os graus de similaridade entre os objetos e a expectativa de resposta que o usuário busca nas suas consultas.

2.2 Arquitetura da Informação (AI)

Arquitetura da Informação (AI) projeta ambientes informacionais compartilhados a fim de facilitar tarefas, acesso à conteúdos e a apropriação desses espaços pelos usuários (LUZ, 2020). Segundo Rosenfeld, Morville e Arango (2015), o termo foi introduzido por Saul Wurman em 1976 como uma ciência e arte de criar instruções para espaços informacionais com o propósito de visualizar a informação como sistemas de organização (forma de categorizar e hierarquizar o conteúdo utilizando ontologias para eliminar heterogeneidade e ambiguidade),



de navegação (forma de ajudar o usuário a entender onde está e para onde pode ir dentro de um ambiente de informação), de pesquisa (compostos de ferramentas de busca e algoritmos de recuperação, utilizam os metadados para apresentar conteúdos classificados) e de rótulos (títulos e rótulos para conteúdos por meio de links, termos indexados e palavras-chaves, componentes que ajudam o usuário a encontrar o caminho da informação).

É preciso entender a informação a ser tratada nesses sistemas, conhecer os objetivos do negócio, para buscar os recursos disponíveis para o projeto estrutural do ambiente de informacional. Rosenfeld, Morville e Arango (2015) apresentam uma ecologia da informação composta de uma relação entre conteúdo, usuário e contexto, elementos essenciais para a compreensão de ambientes informacionais e detalhados como: conteúdo, o tipo de documentos ou dados, a natureza e o volume de dados, a propriedade, o formato estruturado, metadados etc.; usuário, o público receptor com seu comportamento em busca da informação; e contexto, as metas do negócio, tecnologias, culturas, políticas e regras e as circunstâncias onde o objeto informacional de interesse está inserido.

Em geral, os modelos de AI apresentam etapas que organizam e descrevem a informação por meio de padrões preocupando-se com a entrega da informação em canais interativos que atendam as expectativas dos usuários utilizando-se de forte infraestrutura tecnológica, algoritmos inteligentes, integração de plataformas de informação, desenvolvimento de *big datas*, e de outras formas de organização e apresentação do conhecimento, levando em consideração o contexto e os modelos de negócios de gestão.

Oliveira (2014) apresenta um modelo de AI com seis etapas: 1 – Fenomenologia do contexto; 2 – Mapeamento e categorização das essências; 3 – Relacionamentos complexos; 4 – Delineamento da ecologia informacional complexa; 5 – Padronização; 6 – Implementação e avaliação, todas no sentido de levantamento do contexto, organização dos elementos essenciais e demonstração de seus relacionamentos, desenho da visualização da informação e padrão de metadados visando um projeto final.

Lima-Marques e Macedo (2006) apresentam um modelo de AI em três níveis: 1 – Nível de metamodelagem: para entendimento do contexto; 2 – Nível de modelagem: para identificação, captura, representação, organização e comunicação da informação; 3 – Nível de aplicação: para o uso da informação.



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

Este último modelo será o elemento de estudo para o planejamento apresentado neste trabalho, adaptado ao desenvolvimento de uma aplicação inteligente apoiada em alta tecnologia.

2.3 Contribuição da Ciência da Computação (CC)

Existe uma interface entre a CI e a Ciência da Computação (CC), a primeira trata da natureza da informação e sua comunicação para uso pelos humanos, e a segunda, ciência da computação, trata de algoritmos que transformam informações para uso em dispositivos eletrônicos. Carvalho (2013, p. 39) afirma sobre CI e CC que “ambos os objetos são interrelacionados e não competidores, mas complementares”. Para Saracevic (1991, p. 10) “a base da relação entre CI e ciência da computação reside na aplicação dos computadores e da computação na recuperação da informação”. A necessidade de processamento computacional, criação e distribuição de base de dados, faz com que a tecnologia seja uma ferramenta eficaz para o acesso e disseminação da informação, para recuperar a informação. Entretanto, os recursos utilizados em sistemas computacionais precisam ser selecionados, mantidos e evoluídos para convergirem e atenderem determinada demanda informacional, os modelos de desenvolvimento de software são instrumentos utilizados pela ciência da computação para alcançar este fim. O modelo de referência de arquitetura de sistemas distribuídos *Reference Model of Open Distributed Processing* (RM-ODP) da ISO/IEC 10746:1998, descreve as estruturas de um sistema distribuído a partir de propriedades externas e dos detalhes de componentes internos utilizando, para isso, cinco pontos de vista: organização, informação, computação, engenharia e tecnologia. O ponto de vista da



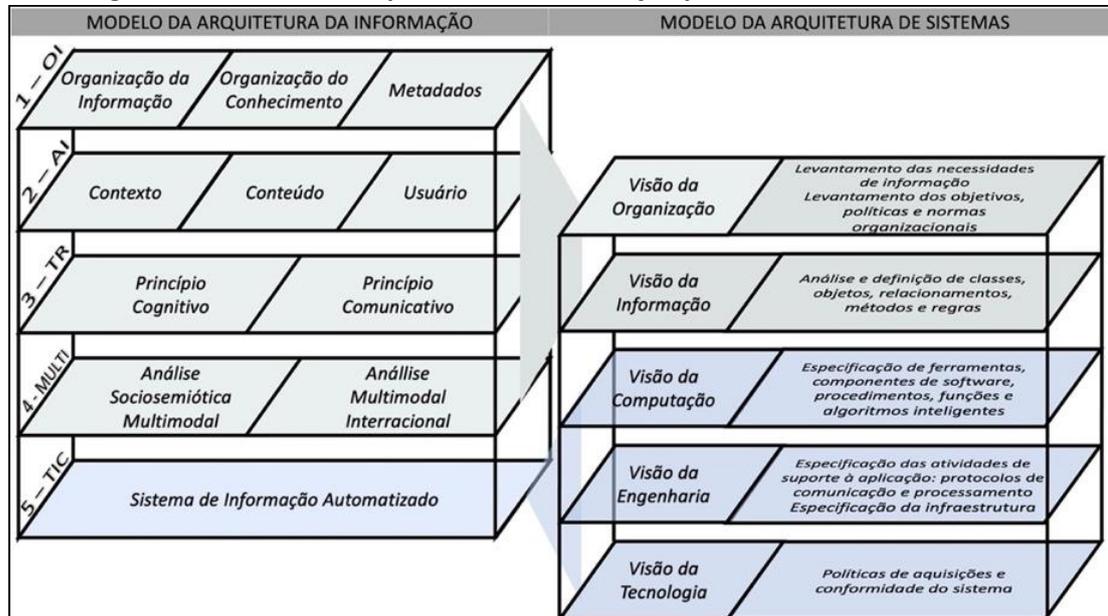
organização apresenta os requisitos do negócio e do domínio, as políticas, regras e normas nas quais as funcionalidades serão desenvolvidas, além dos usuários envolvidos, os interessados, fornecedores etc. O ponto de vista da informação auxilia a compreensão da informação e apresenta uma análise através de classes, objetos, relacionamentos e outros elementos da organização do conhecimento. Farooqui, Logripro e Meer (1995) afirmam que os demais pontos de vista: computação, engenharia e tecnologia são especificações da construção do sistema automatizado e do desenvolvimento dos códigos e algoritmos. Para o autor, a visão da computação especifica ferramentas computacionais, componentes de software, procedimentos e funções e outras estruturas de linguagem e mecanismos reaproveitáveis capazes de implementar as funcionalidades e permitir a comunicação entre os componentes de software. O ponto de vista da engenharia apresenta a arquitetura sistemas com softwares, hardwares, infraestrutura de comunicação, processamento e serviços distribuídos em rede, armazém de dados, máquinas, softwares e todas as interfaces computacionais do sistema. Por fim, o ponto de vista da tecnologia instância as especificações da arquitetura de sistemas, apresenta as tecnologias e plataformas escolhidas, políticas de aquisições, especificações de software e hardware necessárias a implantação do sistema. Neste trabalho, esses três últimos pontos de vista serão utilizados para complementar o modelo arquitetural adotado.

2.4 Metodologia e desenvolvimento

O estudo foi desenvolvido a partir de pesquisas qualitativas na área da CI no campo da RI focados em processos ontológicos e na área de CC e a integração com a AI. A partir da revisão da literatura e análise documental foi elaborado um planejamento para solucionar os problemas da pesquisa como o grande número de documentos, inconsistência e a falta de metadados e de padronização de termos para a categorização na descrição dos produtos.

A metodologia escolhida integra dois modelos arquiteturais de informação e de sistemas distribuídos. O ponto de integração exclui a 5ª. Etapa no modelo Orlandi (2019), etapa de “Gamificação”, uma sugestão de especialistas da CI e registrada no trabalho de Orlandi (2019) como Modelo para Criação de Espaços Informacionais. Neste contexto foi inserida uma nova 5ª. Etapa com a contribuição da Tecnologia da Informação e Comunicação (TIC) e com o pilar de Sistema de Informação Automatizado. Este pilar será o ponto de integração para o modelo da arquitetura de sistemas RM-ODP, conforme Figura 1:

Figura 1 – Modelo de arquitetura da informação para sistemas automatizados



Fonte: Planejamento inspirado nos modelos de arquitetura de Orlandi (2019) e de arquitetura de sistemas RM-ODP.

O Modelo para Criação de Espaços Informacionais revisado e expandido será utilizado para a criação de qualquer ambiente informacional automatizado, como por exemplo para gestão do fisco e definição de necessidades informacionais da fiscalização, para a definição de termos, atributos e relações que atendam estas necessidades e que, ao final, recupere e apresente a informação de forma condizente com a expectativa dos usuários.

Pela insuficiência de atributos no modelo RM-ODP para a AI a proposta conta com a robustez do modelo Orlandi (2019) com quatro etapas: OI, AI, TR e MULTI além da própria contribuição computacional que detalha uma arquitetura de sistemas distribuídos com a componentização, algoritmos, infraestrutura, aquisições tecnológicas etc., conforme descritos nas visões computacional, engenharia e tecnologia.

A integração descreve os objetos informacionais de acordo com a realidade na qual estão posicionados e com a representação de suas principais características, respondendo por exemplo: Quais objetos são relevantes para representar termos, seus atributos e relacionamentos para a gestão do fisco? Em que contexto eles existem? Quais metadados expressam o padrão desses objetos? Qual SOC representa os objetos e a sua realidade? Qual comportamento do usuário? Quais os interesses do usuário na organização? Como



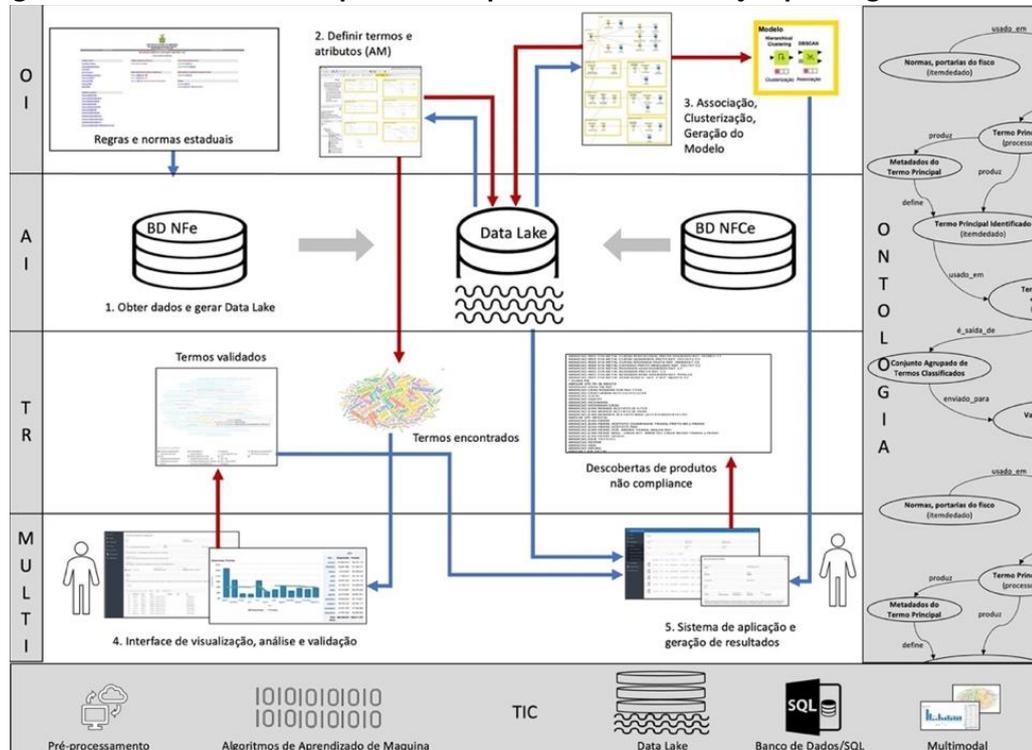
comunicar a informação ao usuário? Quais os recursos e como serão alimentadas e retroalimentadas as informações pelas consultas dos usuários?

As três últimas visões do modelo RM-ODP associadas à nova 5ª. Etapa de TIC ajudam a definir igualmente a informação utilizando a linguagem computacional como produto oriundo da tecnologia, convergindo o ambiente informacional projetado para um ambiente de software automatizado, utilizando tecnologias capazes de aprender automaticamente, processar a linguagem natural dos textos e implementar uma visão de conhecimento computacional para os dados disponíveis à gestão do fisco.

A expansão do modelo com a inclusão de novos pilares permite a organização do conhecimento sem inconsistências ou ambiguidades, com controle dos sinônimos e relações semânticas bem definidas.

Uma vez apresentado o planejamento para elaboração de espaços informacionais automatizados, o passo seguinte é descrever as atividades necessárias às etapas do modelo conforme Figura 2 abaixo:

Figura 2 – Detalhamento do plano de arquitetura da informação para a gestão do fisco.



Fonte: Elaborado pelos autores (2022).

A partir da Figura 2 foi elaborado o Quadro 1 para detalhar as atividades necessárias para alcançar o “Sistema de aplicação e geração de resultados”. Como exemplo foram



utilizados produtos de medicamentos encontrados em NFC-e:

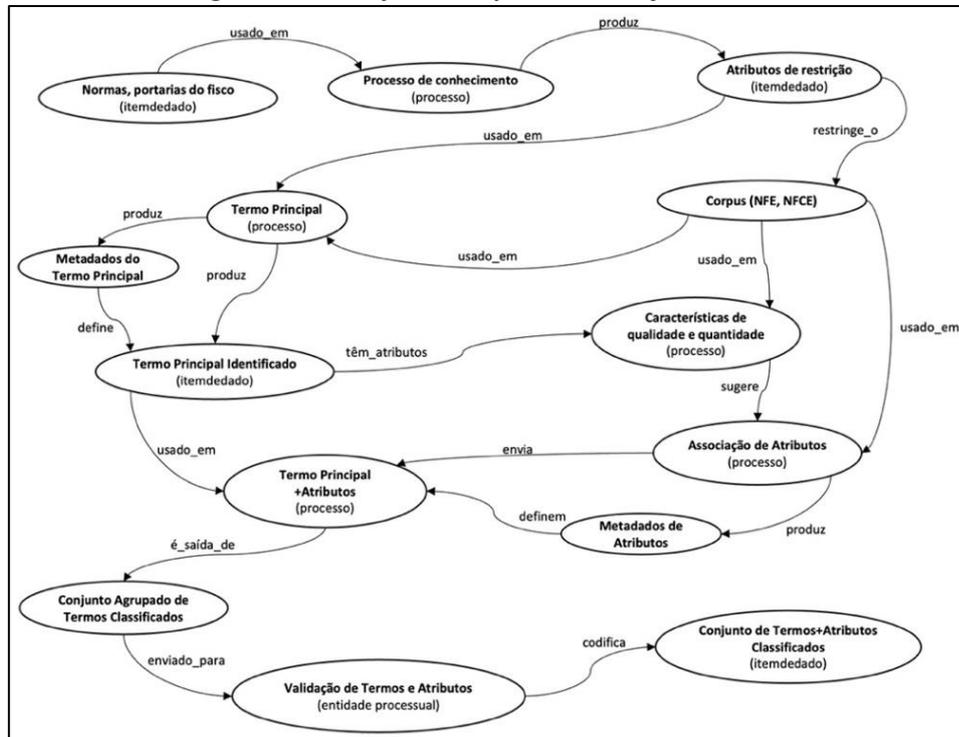
Quadro 1 – Etapas do plano de arquitetura da informação para a gestão do fisco.

Atividade	Descrição da Atividade
1. Obter dados e gerar Data Lake	A partir das bases de dados de NF-e e NFC-e gerar o <i>Data Lake</i> com as informações pertinentes ao objeto de estudo como regras e normas do fisco e aplicar filtros para produzir uma base inicial de pesquisa (conforme modelo ontológico da Figura 3).
2. Definir termos e atributos (AM)	<p>Através de algoritmos inteligentes, como algoritmos de associação de regras, será gerado um conjunto de termos e atributos usados para tokenização e pré-processamento como entrada na etapa 3 (conforme modelo ontológico da Figura 3).</p> <p>Os pré-requisitos desta etapa constituem a identificação dos atributos do objeto de estudo, por exemplo: “Medicamentos” possuem os seguintes atributos: “Nome do Medicamento ou Princípio Ativo”, “Complemento”, “Dosagem e Volumetria”, “Forma Farmacêutica”, “Marca”, “Quantidade e Embalagem”.</p> <p>Exemplo: “DIPIRONA MONOIDRATADA GTS 500 MG 20 ML”: Nome do Medicamento = “DIPIRONA”, Complemento = “MONOIDRATADA”, Dosagem e Volumetria = “500MG”, Forma Farmacêutica = “GTS”, Marca = “”, Quantidade e Embalagem = “20ML”.</p>
3. Associação, clusterização, geração do modelo	Os dados de notas fiscais serão pré-processados com base nos termos identificados na etapa anterior, a seguir clusterizados para gerar o modelo de clusterização com a lista de termos e suas aproximações (conforme modelo ontológico da Figura 3).
4. Interface de visualização, análise	O usuário, especialista do fisco estadual, validará os termos e suas aproximações associados a cada cluster gerado podendo, inclusive, categorizar vários clusters sob um único grupo, por exemplo, as “DIPIRONAS” congregam os clusters de “DIPIRONAS”, “NOVALGINAS”, “NEOSALDINAS”, “DORALGINAS”, “LISADOR” etc.
5. Sistema de aplicação e geração de resultados	Uma vez que o modelo esteja validado ele pode ser aplicado no <i>Data Lake</i> para geração de resultado e respostas a perguntas do fisco em muitas interfaces a partir do interesse do usuário ou do comportamento dos dados, dependendo da relevância do objeto informacional. Por exemplo: Qual o preço médio da dipirona 500MG comprimido? Qual a quantidade de Neosaldinas vendidas no mês passado? Que contribuinte está vendendo Dipirona? Ou qualquer outra questão a ser respondida.

Fonte: Elaborado pelos autores (2022).

A Figura 3 abaixo apresenta o modelo ontológico para as atividades de integração do modelo de arquitetura da informação para sistemas automatizados com a produção de artefatos e relacionamentos necessários à coleta das informações relevantes, expectativas da gestão para o fluxo da informação, identificação e classificação dos produtos.

Figura 3 – Planejamento para elaboração do SOC.



Fonte: Elaborado pelos autores (2022).

Inicia com o levantamento das normas e documentos da gestão do fisco, com o conhecimento da base de dados e o levantamento das necessidades e prioridades do fluxo informacional. Estes elementos, denominados <itens de dados>, serão utilizados no <processo> de conhecimento da gestão do fisco e determinarão os recortes utilizados no processo de seleção dos documentos. Quando aplicados estes recortes (filtros), será elaborado o <processo> de criação do termo principal de interesse da gestão do fisco, que uma vez executado na base de dados de documentos irá produzir a lista de termos principais. Ainda sobre a lista de termos principais serão agregados outros termos qualificadores e quantificadores extraídos também da base de dados, tratados como sinônimos e atributos, saída de um <processo de criação de termo principal + atributos + sinônimos>. A definição do termo principal, atributos e sinônimos deve ser flexível a ponto de alcançar o maior número de descrições de produtos contidas no documento eletrônico da nota fiscal, capaz de representar adequadamente o produto no contexto em que está sendo pesquisado. Nesta fase, haverá o <processo> de validação com especialistas da área do fisco estadual bem como nos documentos de normas e portarias utilizadas. A validação permitirá o projeto prosseguir pelas fases computacionais, codificando as regras de classificação de termos principais



obtidas, extraíndo e gerando uma nova fonte informacional antes inacessível à gestão tanto pela complexidade de extração, quanto pela manipulação do volume do dado e variedade de termos. Este planejamento irá apoiar as respostas aos questionamentos iniciais pois os objetos relevantes à gestão do fisco estarão relacionados aos produtos descritos nas notas fiscais eletrônicas e serão identificados por termo principal mais lista de atributos mínimos para descrever por exemplo: nome e marca e embalagem e volumetria, cor, sabor etc.

3 CONSIDERAÇÕES FINAIS

Os modelos de AI pesquisados concentram-se em fornecer meios robustos de representar, organizar e recuperar a informação, entretanto, eles não incluem a descrição das atividades tecnológicas pertinentes à CC nem o ponto onde estas atividades estarão integradas ao modelo de AI. Este trabalho revisou e expandiu o modelo de AI para Criação de Espaços Informacionais de Orlandi (2019) integrando-o ao modelo de arquitetura de sistemas distribuídos RM-ODP para criação de ambiente informacional automatizado capaz de extrair informações descritas em linguagem natural em documentos eletrônicos da gestão do fisco estadual e representá-las a partir de metadados como nomes e atributos em linguagem computacional.

O planejamento incluiu a nova etapa de TIC com o objetivo de não limitar o modelo de AI diante da interdependência entre a Ciência da Informação com a Ciência da Computação e da evolução tecnológica já incorporada aos pilares da CI e que, atualmente, não convivem mais em modelos separados. A nova etapa permite aplicação em documentos eletrônicos buscando um “Sistema de aplicação e geração de resultados” retroalimentado para o aprimoramento das consultas dos usuários.

A integração entre os modelos está representada nas atividades do modelo ontológico onde é possível acompanhar as etapas para representar, organizar e recuperar a informação.

A utilização de algoritmos de regras de associação e clusterização ajudam nas etapas iniciais de identificação dos termos relevantes bem como seus atributos principais que caracterizam esse termo de forma mais independente do especialista.

REFERÊNCIAS

BRASCHER, M; CAFÉ, L. Organização da informação ou Organização do conhecimento. **ENANCIB**, São Paulo, SP, 2008.



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.

CEDRO, L. F. dos A.; DUQUE, C. G. A disseminação da informação científica como garantia da legitimidade dos resultados de ensaios clínicos. **Revista Digital de Biblioteconomia e Ciência da Informação**. v. 18, Campinas, SP. 2020.

DUQUE, C. G. **SiRILiCO uma proposta para um sistema de Recuperação da Informação baseado em Teorias da Linguística Computacional e Ontologia**. 2005. 120 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

EVARISTO, E.; DUQUE, C. G. Recuperação da Informação em vídeo por meio de análise multimodal. In: DUQUE, C. G (org). **Ciência da Informação Estudos e Práticas**. Brasília, DF: Thesaurus Editora de Brasília Ltda, 2011. p. 237-250.

FAROOQUI, K.; LOGRIPRO, L.; MEER, J. de. The ISO Reference Model for Open Distributed Processing – an introduction. **Computer Networks and ISDN Systems**, v. 27, p. 1215-1229. July, 1995.

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

GONZALEZ, J. A. M. Palavra, Termo, Conceito: das linguagens documentárias até os vocabulários semânticos para Web. In FREITAS, L. S. de; MARCONDES, C. H.; RODRIGUES, A. C. (org). **Documento: gênese e contextos do uso**. Niterói: EdUFF, 2010. p. 11-34.

GONZALEZ, M.; LIMA, V.L.S.de. Recuperação de Informação e Processamento da Linguagem Natural. XXIII Congresso da Sociedade Brasileira de Computação, Campinas, 2003. **Anais do III Jornada de Mini-Cursos de Inteligência Artificial**, 2003.

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

HJØRLAND, B. Facet analysis: the logical approach to knowledge organization. **Information Processing and Management**, v. 49. Dinamarca: Elsevier Ltd, 2012. p.545–557.

ISO/IEC 10746:1. **Information technology — Open Distributed Processing — Reference model**: Overview. Suíça. 1998.

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.



LUZ, C. dos S. **Arquitetura da Informação: do conteúdo à experiência do usuário**. São Paulo: Feed Consultoria, 2020.

MORAES, A. F.; ARCELLO, E. N. O conhecimento e sua representação. **Informação & Sociedade**: Estudos, v. 10. n.2, 2000. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/91200>. Acesso em: 30 maio 2022.

MORI, A. **Modelagem de conhecimento baseada em ontologias aplicadas às políticas públicas de habitação**. 2009. 109 f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, DF, 2009.

OLIVEIRA, H. P. C. de. **Arquitetura da informação pervasiva: contribuições conceituais**. 2014. 202 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciência, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Marília, SP, 2014.

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.

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. **International Conference on Conceptions of Library and Information Science**: historical, empirical and theoretical perspectives, aug.26-28, 1991.

VICTORINO, M. de C; PINHEIRO, M. S; SANTOS, R. F. Organização da informação e do conhecimento em sistemas de informação transacionais para o seu reuso em sistemas de apoio a decisão. In BAPTISTA, D. M; ARAÚJO Jr., R. H. de. (org). **Organização da informação abordagens práticas**. Brasília, DF: Thesaurus, 2015. p. 219-247.

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