26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

XXVI ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO - XXVI ENANCIB

ISSN 2177-3688

GT 5 – Política e Economia da Informação

ARQUIVOLOGIA CONTEMPORÂNEA E A INDEXAÇÃO DE DOCUMENTOS DIGITAIS

CONTEMPORARY ARCHIVAL SCIENCE AND THE INDEXING OF DIGITAL DOCUMENTS

Romário Penna Forte de Amorim - Universidade Federal da Bahia (UFBA)
Isabella de Melo Moreira - Universidade Federal da Bahia (UFBA)
Renan Pereira Oliveira - Universidade Federal da Bahia (UFBA)
Bárbara Coelho Neves - Universidade Federal da Bahia (UFBA)

Modalidade: Trabalho Completo

Resumo: As tecnologias da informação e comunicação, no contexto das práticas profissionais da Ciência da Informação, revelaram novos desafios e a necessidade de uma revisão teórica e conceitual do próprio status da informação no século XXI e de sua relação com outras áreas do conhecimento. Especificamente no campo da Arquivologia, o advento do documento digital demandou o desenvolvimento de novas técnicas de indexação. A proposta deste estudo é realizar uma breve revisão da literatura sobre o tema, apresentando algumas proposições estratégicas para estabelecer uma articulação interdisciplinar entre a Arquivologia e a Ciência da Computação em estudos voltados à recuperação da informação (RI) no contexto dos documentos digitais. Abordam-se a Técnica de Índice Invertido, a Indexação Vetorial (*embeddings*) e a Indexação com Técnicas de Processamento de Linguagem Natural (PLN), além da Recuperação de Informação com Geração Aumentada (RAG) como estratégia de recuperação semântica ancorada em evidências documentais. A pesquisa utilizou a metodologia bibliográfica e concluiu que sistemas de recuperação de informação, no âmbito da Arquivologia, precisam atender às especificidades contemporâneas da área, além de dialogar com os demais sistemas arquivísticos (SIGAD e RDC-Arq). Conclui-se, ainda, que as técnicas de indexação e as arquiteturas de recuperação aumentada representam uma evolução significativa dos procedimentos de recuperação da informação, tornando-se o letramento digital uma competência essencial para os profissionais da Arquivologia.
Palavras-chave: Indexação de Documentos Digitais; Sistema de Recuperação de Informação (SRI); Recuperação de Informação com Geração Aumentada (RAG); Inteligência Artificial; Arquivologia Contemporânea.

Abstract: Information and communicatin technologies, within the context of professional practices in Information Science, have introduced new challenges and highlighted the need for a theoretical and conceptual reassessment of the very status of information in the twenty-first century, as well as its relationship with other fields of knowledge. Specifically within the

field of Archival Science, the advent of digital records has required the development of new indexing techniques. This study aims to provide a brief literature review on the subject, presenting strategic propositions to foster an interdisciplinary articulation between Archival Science and Computer Science in studies focused on information retrieval (IR) in the context of digital records. The discussion addresses the inverted index technique, vector indexing (embeddings), and indexing based on Natural Language Processing (NLP), as well as Retrieval-Augmented Generation (RAG) as a semantic retrieval strategy grounded in documentary evidence. The research adopted a bibliographic methodology and concluded that information retrieval systems, within the scope of Archival Science, must address the specific requirements of the field while also interacting with other archival systems, such as Electronic Records Management Systems and Trusted Digital Repositories for Archival Records. Furthermore, the study concludes that indexing techniques and retrieval-augmented architectures represent a significant advancement in information retrieval procedures, making digital literacy an essential competency for archival professionals.
Keywords: Indexing of Digital Documents; Information Retrieval Systems (IRS); Retrieval-Augmented Generation (RAG); Artificial Intelligence; Contemporary Archival Science.

1 INTRODUÇÃO

As práticas arquivísticas, conforme Cruz Mundet (2024), são atividades milenares, realizadas desde as antigas civilizações que desenvolveram técnicas e habilidades para registrar informações nos mais diversos tipos de suporte (tábua de argila, pergaminho, papel, etc.); com o passar do tempo houve o entendimento da necessidade de desenvolver estratégias para a preservação de tais instrumentos (documentos), com o objetivo de possibilitar o acesso às informações de modo atemporal.

Ao considerarmos que a arquivologia, contemporânea ao contexto de sociedade da informação, mantem em evidência princípios como proveniência, organicidade e integridade, legitima a demanda por estudos e discussões que ampliem a compreensão de procedimentos arquivísticos no contexto de documentos digitais. Assim, o problema de pesquisa parte do questionamento de como funciona as técnicas de indexação de documentos digitais implementadas em Sistema de Recuperação de Informação (SRI).

Tal pressuposto norteia ao seguinte entendimento: produzir, tramitar e acessar documentos digitais requer conhecimentos que ultrapassam as fronteiras da arquivologia clássica, exigindo análises meta-arquivísticas, ou seja, é preciso ainda refletir sobre a dependência de tecnologias computacionais (*hardwares* e *softwares*) e a maneira como atuam para concretizar atividades arquivísticas no contexto digital.

Doreneles, Correa e Flores (2023), ao apresentarem a proposição de que a gestão de documentos, a preservação e o acesso e transparência ativa, assim como todas as funções

arquivísticas, devem ser aplicadas ao ambiente eletrônico de produção e acumulação de documentos de instituições públicas ou privadas, corroboram com o entendimento de Bellotto (2006) e de Marques e Medeiros (2018) sobre a urgência para a formação arquivística contextualizada na interdisciplinaridade entre a arquivologia e o campo das tecnologias da informação e comunicação (TIC).

Destarte, as proposições apresentadas também visam contribuir nesta perspectiva de enlace e interdisciplinaridade "Pombiana"[1] entre a arquivologia e a ciência da computação; adotando como estratégia, estudos voltados ao campo de recuperação de informação (RI), com recorte em técnicas de indexação de documentos digitais para a implementação de recursos de inteligência artificial (IA) voltadas ao aprimoramento de sistemas de recuperação de informação (SRI).

Embasados na conjunção das proposições de Bellotto (2006), Marques e Medeiros (2018) e Olga Pombo (2005), assumimos a interdisciplinaridade entre a arquivologia e a ciência da computação em uma perspectiva de letramento arquivístico digital, que fundamentam a apresentação de estratégias para o aprimoramento de SRI por meio da adoção de técnicas de indexação que implementem tecnologias que compõem a "Biota de IA[2]".

Como procedimento metodológico, foi adotada uma pesquisa bibliográfica, com abordagem qualitativa, na qual foram consultadas e selecionadas como fontes de informação, publicações em suporte papel (livros e artigos), apresentadas em ementas das disciplinas (2024-2025) do Programa de Pós-Graduação em Ciência da Informação (PPGCI/UFBA), além de artigos disponíveis nos repositórios digitais da ScieLo, da Base de Dados em Ciência da Informação (BRAPCI), do *Web of Science* no portal de periódicos da CAPES e da Biblioteca Digital da Sociedade Brasileira de Computação (SBCOpenLib).

A estratégia de busca foi norteada por publicações que contemplassem a atividade de indexação no âmbito da RI, tanto no campo da Ciência da Informação quanto no campo da Ciência da Computação. No âmbito da CI optamos por um recorte temporal mais dilatado

[1] Olga Pombo (2005) ao apresentar o entendimento de interdisciplinaridade como um espaço intercalar, sem que haja uma interpretação equivocada de sobreposição nem fusão entre as áreas.

[2] A expressão "Biota de IA" é aplicada como metáfora para referir-se aos diversos termos e tecnologias que são empregados em sistemas computacionais que implementam a Inteligência Artificial (IA), a exemplo do aprendizado de máquina (*Machine Learning*), redes neurais artificiais (RNA), aprendizado profundo (Deep Learning), processamento em linguagem natural (PLN), *Large Language Model* (LLM), algoritmos, *big-data*, web semântica, Inteligência Artificial Generativa, entre outros.

(1970 a 2025), por ser o campo provocador do estudo e de onde partiram os pressupostos da indexação voltada a SRI em uma perspectiva arquivística. Publicações, a exemplo de Cesarino (1985), Lancaster (1993, 2004) e Fujita (2012, 2019), entre outros, constituíram o conjunto universo de referenciais teóricos.

No campo da CC, como estratégia de busca, foi considerado o requisito de hodiernidade tecnológica, ou seja, técnicas de indexação de documentos voltada a RI atravessados por recursos de IA generativa em uma perspectiva contemporânea. Essa perspectiva direcionou a um recorte temporal mais reduzido (2015 a 2026), principalmente por considerar este intervalo mais promissor para a localização de publicações científicas produzidas também a partir de perspectivas da Arquivologia.

Norteado pelo objetivo da pesquisa, de apresentar técnicas de indexação de documentos digitais voltados a recuperação de informação, a partir de uma revisão de literatura e buscando estabelecer uma articulação de interdisciplinaridade entre a arquivologia e a Ciência da Computação. Nos repositórios digitais foi aplicada uma estrutura de descritor contendo os termos: “Técnicas de indexação de documentos digitais” + “Recuperação de informação” + “Arquivologia”.

O total de publicações localizadas, nos 4 repositórios citados, acrescido dos livros disponibilizados em bibliotecas físicas da Universidade Federal da Bahia, tomando como primeira estratégia de corte o título das publicações, totalizaram 180 títulos, os quais passaram por nova etapa de mineração, principalmente, através da leitura do resumo, do sumário e da introdução, formando um subconjunto de 42 publicações que apresentavam a atividade de indexação de documentos digitais no âmbito da arquivologia mais 26 publicações da Ciência da Computação, totalizando o universo de 68 referenciais teóricos consultados, que deu origem ao subconjunto de referenciais citados no desenvolvimento da pesquisa.

2 SISTEMA DE RECUPERÇÃO DE INFORMAÇÃO (SRI)

Sistema de recuperação de informação (SRI) remete ao serviço de referência, área de estudo com amplos debates na biblioteconomia, mas também essencial aos demais ambientes informacionais a exemplo de arquivos e centros de documentação. “O objetivo principal de um sistema de RI é recuperar todos os documentos que são relevantes à

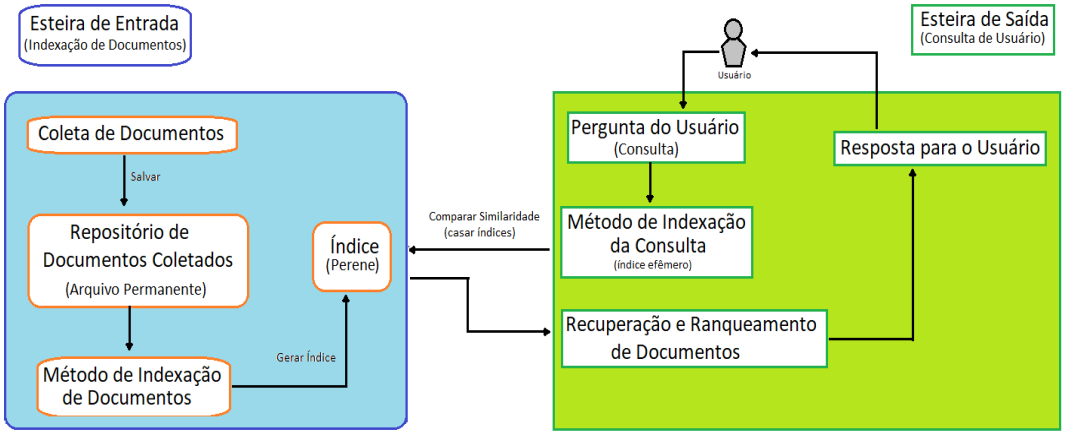
necessidade de informação do usuário e, ao mesmo tempo, recuperar o menor número possível de documentos irrelevantes" (Baeza-Yates e Ribeiro-Neto, 2013, p. 4).

Em Saracevic (1999), ao citar Mooers (1951), observa-se que o entendimento sobre uso de máquinas para realizar atividades de RI não é uma discussão recente no âmbito da CI.

> [...] a recuperação de informação abrange além dos aspectos intelectuais da descrição da informação e sua especificação para busca, também quaisquer sistemas, técnicas ou máquinas que sejam empregados para realizar a operação (Mooers, 1951 *apud* Saracevic, 1999, p. 1057) (tradução nossa).

Com o aprimoramento de tecnologias computacionais e o crescimento na produção de documentos digitais, o entendimento de RI vai incorporando novas perspectivas. Cesarino (1985) apresenta RI como "parte de um modelo de comunicação dentro de um contexto social-cultural-histórico" (Cesarino, 1985, p.159), a partir do qual SRI pode ser definido como "[...] um conjunto de operações consecutivas executadas para localizar, dentro da totalidade de informações disponíveis, aquelas realmente relevantes. Para isso, executam as funções de seleção, análise, indexação e busca das informações" (Cesarino, 1985, p.157).

Figura 1 - Sistema de Recuperação de Informação (SRI)

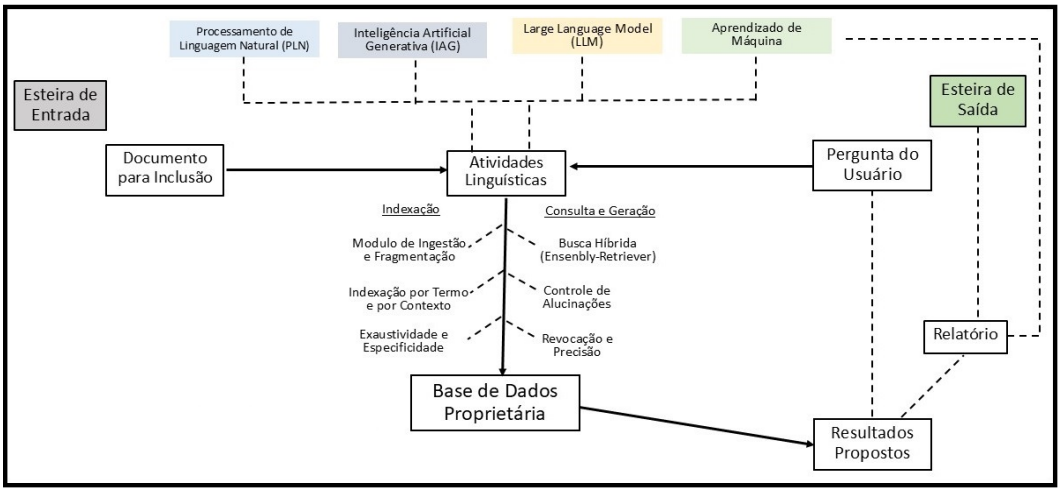


Fonte: elaborada pelos autores (2026), inspirada em Baeza-Yates e Ribeiro-Neto (2013)

No entanto, para recuperar informações disponíveis na base de dados é necessário o desenvolvimento de estratégias que estabeleçam uma análise e representação do conteúdo de cada documento e direcionem a execução de procedimentos, transformando informações em termos de busca, ou seja, a indexação das informações, que irá viabilizar a localização e recuperação da informação desejada dentro de um grande volume de texto.

A partir do fluxograma apresentado na Figura 2, observa-se a atuação das tecnologias digitais como uma espécie de espinha dorsal de atividades linguísticas que intercala os fluxos de entrada (indexação da base de dados) e de saída (consulta do usuário).

Figura 2 - IA como Espinha Dorsal em SRI



Fonte: elaborada pelos autores (2026)

> A recuperação da informação (RI) lida com a representação, o armazenamento, a organização e o acesso a itens de informação. **Devendo proporcionar ao usuário fácil acesso à informação na qual está interessado**" [...] "Usuários de sistemas modernos de RI, como usuários de máquinas de busca, têm necessidades de informação em diferentes níveis de complexidade. [...] Dada a consulta do usuário, o objetivo maior do sistema de RI é recuperar informações que sejam úteis ou relevantes para o usuário. A ênfase está na recuperação de informação, não na recuperação de dados (Baeza-Yates; Ribeiro Neto, 2013, p. 4, grifo dos autores)

A estratégia de implementação de recursos que compõem a "Biota de IA" como aprimoramento de SRI para arquivos é chancelada pelas proposições de Baeza-Yates e Ribeiro-Neto quando reforçam duas características essenciais a todo SRI no âmbito da arquivologia, a) interesse de usuários em buscar por informações, não apenas por dados e b) a facilidade de uso do sistema e acesso às informações.

3 INDEXAÇÃO NO CONTEXTO DE DOCUMENTOS DIGITAIS

Esta seção propõe-se a apresentar, de forma didática, algumas das principais estratégias de indexação voltadas a recuperação de informações em documentos digitais. Para otimizar a explanação, entendemos que o índice é construído a partir do conjunto de informações existentes nos documentos (base de dados), através da implementação de um, ou de uma combinação, de métodos de indexação que podem ser classificados como:

Tabela 1- Métodos de Indexação Digital

Indexação manual	A análise e extração de termos e conceitos é feita por humanos, usando linguagens documentárias para a elaboração do índice.
Indexação automática	Realizada por algoritmos que analisam o texto completo e extraem termos relevantes.

Indexação semiautomática	A extração é realizada de forma automática, por algoritmos, porém é o processo é supervisionado através da revisão humana.

Fonte: elaborada pelos autores (2026)

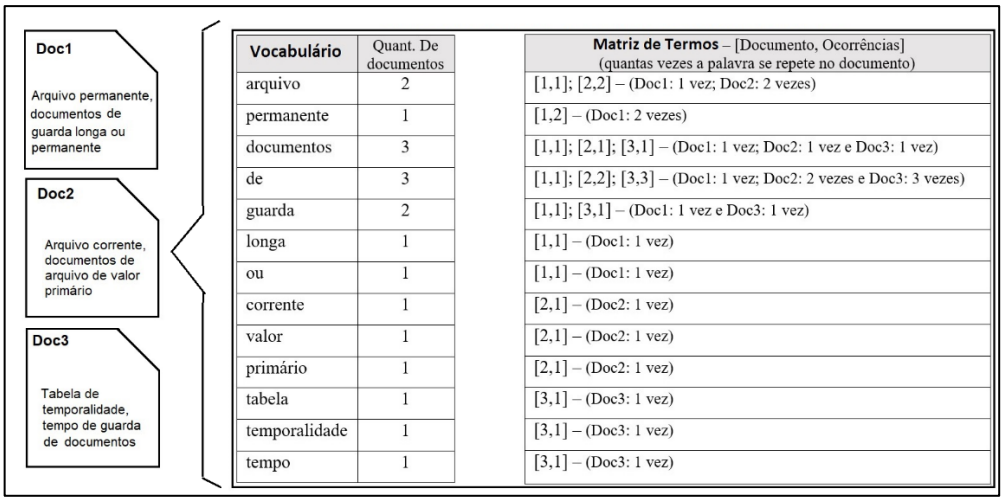
Para esta pesquisa, consideramos o processo de indexação automática a partir de recursos já implementados em plataformas de Inteligência Artificial Generativa, a exemplo do Chat GPT e do DeepSeek. Dentre as diversas técnicas de indexação para documentos digitais, foi observado, nos referenciais teóricos do campo da CC, alta recorrência das seguintes técnicas de indexação: Índice Invertido, Indexação Vetorial (*Embeddings*) e Processamento de Linguagem Natural (PLN). Esta constatação direcionou ao entendimento de considera-las as mais relevantes para serem apresentadas, observando também o potencial em contribuir no espectro de interdisciplinaridade e letramento digital no âmbito da arquivologia.

3.1 TÉCNICA DE INDEXAÇÃO: ÍNDICE INVERTIDO

A técnica denominada Índice Invertido (*Inverted Index*) "é um mecanismo orientado a palavras para a indexação de uma coleção de texto a fim de acelerar a tarefa de busca" (Baeza-Yates; Ribeiro-neto, 2013, p. 342). Do ponto de vista computacional, o índice invertido utiliza, então, estruturas de dados otimizadas que armazenam mapeamentos rápidos entre termos e documentos, permite assim que a recuperação seja em tempo logarítmico ou constante, o que o torna viável ainda que seja para bases de dados de grande escala. Trata-se de uma das técnicas de indexação mais comuns em mecanismos de busca. O índice invertido mapeia cada palavra de forma única, gerando um vocabulário que também registra em quantos documentos a palavra aparece. Paralelamente é gerada uma lista de ocorrências (matriz de termos) que indica quantas vezes a palavra ocorre em cada documento.

Tomemos como exemplo a Figura 3 (apenas para exemplificar o funcionamento do índice invertido), onde são apresentados os documentos (Doc_1, Doc_2 e Doc_3) seguido do *vocabulário* e da *matriz de termos*, gerados automaticamente para no processo de indexação.

Figura 3 - Índice Invertido

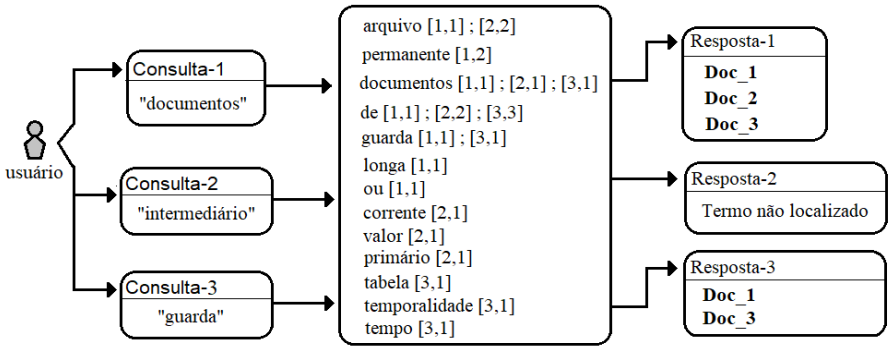


Vocabulário	Quant. De documentos	Matriz de Termos – [Documento, Ocorrências] (quantas vezes a palavra se repete no documento)
arquivo	2	[1,1]; [2,2] – (Doc1: 1 vez; Doc2: 2 vezes)
permanente	1	[1,2] – (Doc1: 2 vezes)
documentos	3	[1,1]; [2,1]; [3,1] – (Doc1: 1 vez; Doc2: 1 vez e Doc3: 1 vez)
de	3	[1,1]; [2,2]; [3,3] – (Doc1: 1 vez; Doc2: 2 vezes e Doc3: 3 vezes)
guarda	2	[1,1]; [3,1] – (Doc1: 1 vez e Doc3: 1 vez)
longa	1	[1,1] – (Doc1: 1 vez)
ou	1	[1,1] – (Doc1: 1 vez)
corrente	1	[2,1] – (Doc2: 1 vez)
valor	1	[2,1] – (Doc2: 1 vez)
primário	1	[2,1] – (Doc2: 1 vez)
tabela	1	[3,1] – (Doc3: 1 vez)
temporalidade	1	[3,1] – (Doc3: 1 vez)
tempo	1	[3,1] – (Doc3: 1 vez)

Fonte: elaborada pelos autores (2026)

O vocabulário (coluna à esquerda no índice) registra de forma única a ocorrência da palavra, além de informar em quantos documentos ela foi encontrada; já a matriz de termos (coluna à direita no índice) informa quantas vezes a palavra aparece em cada documento. A Figura_4 apresenta de forma suscinta a recuperação de informação a partir do método de índice invertido em um processo de busca que implementa o modelo Booleano[3]:

Figura 4 - Consulta por índice Invertido



Fonte: elaborada pelos autores (2026)

Consulta-1: o usuário procura pelo termo “documentos”, logo, a busca apresentará como resposta Doc_1, Doc_2 e Doc_3, pois nos 3 documentos foi identificada a ocorrência do referido termo, ou seja, o validador Booleano é igual a “1” (verdadeiro) para os 3 documentos.
Consulta-1: o usuário busca pelo termo “intermediário”; neste caso o validador Booleano será igual a 0 (zero ou falso) para Doc1, Doc2 e Doc3, indicando que o termo não ocorre em nenhum dos 3 documentos.

[3] Modelo Booleano: em recuperação de informação (RI) refere-se a um sistema clássico, baseado na teoria dos conjuntos e lógica matemática (AND, OR, NOT), que recupera documentos com base na presença (1) ou ausência (0) do termo que está sendo procurado.

Consulta-3: a consulta busca pelo termo “guarda”; neste caso o validador Booleano será “1” apenas para Doc_1 e Doc_3, que são apresentados como resposta a demanda do usuário.

Pela facilidade de implementação e baixo consumo de recursos, a técnica de índice invertido tem uma aplicação expressiva em motores de busca, especialmente na web, no entanto pode-se destacar com principais limitações: a) a não ordenação dos termos; b) a localização apenas por palavra exata, ou seja, não realiza busca por semelhança nem contexto e c) não realiza ranqueamento de resultado para listar primeiro os documentos mais prováveis de atender a demanda informacional do usuário.

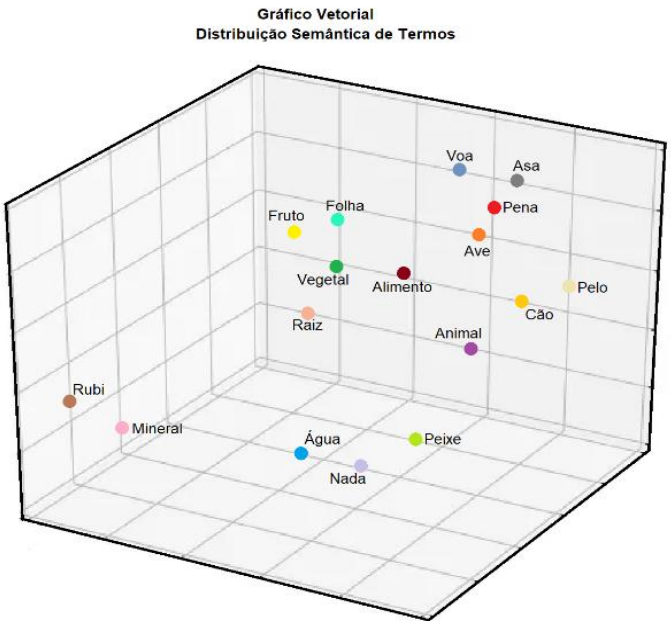
3.2 TÉCNICA DE INDEXAÇÃO VETORIAL (*EMBEDDINGS*)

A indexação vetorial (*embeddings*) é uma técnica empregada em buscas semânticas modernas que transforma texto em representações numéricas (vetores) capazes de capturar o significado contextual por similaridade entre termos. Os termos são indexados em agrupamentos (*clusters*) que os distribui por proximidade de conceito ou contexto, sendo armazenados em um banco de dados vetorial.

Do ponto de vista terminológico, a representação vetorial (embeddings) e a indexação vetorial são etapas distintas de um mesmo pipeline de recuperação de informação. Assim, o “*embedding*” corresponde à transformação do texto em vetores numéricos que capturam relações semânticas, a indexação vetorial é à construção de estruturas de dados que viabilizam a busca eficiente desses vetores em grandes coleções, como exemplo de índices aproximados como o HNSW (Hierarchical Navigable Small World). Zhou *et al.* (2023) descrevem esse processo como parte do paradigma “index-retrieve-rerank”, ou seja, a indexação de documentos pode ocorrer por meio de um índice invertido esparso (sparse inverted index) ou de um índice vetorial denso (dense vector index), evidenciando que a representação vetorial é ingrediente para a indexação, e não sinônimo dela.

A distribuição por proximidade semântica permite localizar informações por similaridade, ou seja, a localização não estará restrita apenas às palavras-chave. Para exemplificar a indexação vetorial a Figura_5 representa uma distribuição vetorial pela proximidade dos seguintes termos: asa, mineral, animal, ave, peixe, pena, rubi, voa, nada, pelo, fruto, raiz, folha, cão, vegetal, água e alimento.

Figura 5 - Gráfico Vetorial



Fonte: elaborada pelos autores (2026)

Em uma indexação vetorial (baseada em word *embeddings* ou semântica distribucional), as palavras são distribuídas em um espaço multidimensional onde a proximidade semântica é calculada através de métricas de similaridade, sendo a distância de cosseno a mais frequentemente utilizada, conforme descrito em Baeza-Yates e Ribeiro-Neto (2013). A distância de cosseno mede o ângulo entre dois vetores, permitindo identificar similaridade mesmo quando os termos não são idênticos. Dessa forma, a proximidade física entre os vetores representa a similaridade de significado ou de contexto, superando as limitações da busca por palavras-chave exatas.

A seguir observa-se que os termos são agrupados em três núcleos semânticos principais, que podem ter a seguinte leitura/interpretação:

- **Agrupamento biológico:** Aproxima os termos com a seguinte hipótese de distribuição: "Cão" e "Pelo" estarão próximos; "Ave", "Pena", "Asa" e "Voa" formarão um subgrupo, enquanto "Peixe" estará em uma área adjacente, mas separada de "Voa"; já "Animal" agirá como um hiperônimo (conceito mais amplo) próximo a todos eles.
- **Agrupamento botânico:** Aproxima outros termos, que formam um subgrupo (cluster), com a seguinte distribuição: os termos "Raiz", "Folha" e "Fruto" estão agrupados próximos a "Vegetal" enquanto o termo "Alimento" atuará como uma ponte entre o os termos “Vegetal” e o “Animal”.
- **Agrupamento natureza bruta:** Posiciona os termos “Água” e “Rubi” com proximidade a “Mineral”. Estes termos podem ser qualificados como “outliers” (pontos distantes ou muito específicos).

Em resumo, as palavras serão organizadas pela similaridade semântica: coisas que aparecem juntas no mesmo contexto (como vegetal/raiz ou asa/voa) terão vetores (*embeddings*) muito próximos.

3.3 TÉCNICA DE INDEXAÇÃO POR PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O Processamento de Linguagem Natural (PLN) é uma técnica que une a ciência da computação mais algoritmos de IA e linguística (Caseli; Nunes, 2024), que entre as diversas aplicações também pode ser implementada para indexar documentos através de um conjunto de procedimentos computacionais denominado pré-processamento.

O pré-processamento, etapa inicial do PLN, pode ser entendido como “processo de limpeza e preparação do texto para normalização. Essa etapa ajuda a reduzir os ruídos no texto e melhora o desempenho do processo de busca” (Haddi; Liu; Shi, 2013, tradução nossa). Ou seja, podemos entender o pré-processamento como uma etapa de normalização ou, processo de transformação do texto bruto (não estruturado) em um formato estruturado que se torna possível através da execução de um conjunto de procedimentos que aprimoram a estratégia de RI. São etapas que compõem o pré-processamento:

Tokenização: Segmentação do texto em unidades menores (Ramisch; Ramisc; Villavicêncio, 2024, p. 100). *Tokens* possuem granularidade fina, ou seja, um token normalmente é constituído por uma palavra ou parte de uma palavra.

Remoção de *Stopwords*: Eliminação de palavras muito comuns, a exemplo de artigos, preposições, conjunções e verbos de ligação ("o", "as", “que”, "de", “é”, etc.) que não trazem valor semântico. “o principal argumento a favor da remoção de *stop words* é a redução do tamanho do vocabulário pois elas representam cerca de 40% das ocorrências de palavras em um corpus.” (Moreira, 2023, p. 458)

***Stemming*/Lematização**: Redução de palavras ao seu radical (ex: "digitar", "digitando", “digitou” são reduzidas ao radical "digit"), permitindo a recuperação de variações do mesmo termo, o que inclui variações em gênero, número, grau.

Reconhecimento de Entidades Nomeadas (NER): Identifica e indexa automaticamente categorias de informação como pessoas, locais, organizações e datas. Conforme abordado por Caseli e Nunes (2024), esta técnica é fundamental no PLN pois estrutura informações não categorizadas, permitindo que sistemas de RI localizem documentos com maior precisão. Em

SRI para arquivos o NER possibilita, por exemplo, recuperar todos os documentos relacionados a uma pessoa específica ou a um período histórico, facilitando pesquisas estruturadas e contextualizadas.

Conforme Zhang e Xiao (2024) o reconhecimento de entidades nomeadas trata-se uma tarefa fundamental do PLN, voltada à identificação e classificação de entidades em texto não estruturado, e não como uma técnica de indexação propriamente dita. Santoso *et al.* (2021) reforçam essa diferenciação quando aplicam NER na extração de conceitos para a construção de ontologias, sendo então a etapa anterior e distinta da indexação e da recuperação.

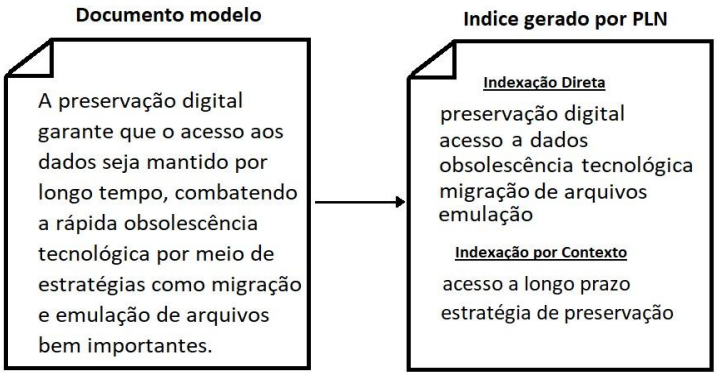
Nesse caso, o pipeline de PLN aplicado à indexação de documentos digitais pode ser compreendido em etapas sequenciais: (i) pré-processamento textual (tokenização, remoção de stopwords, stemming/lematização); (ii) representação vetorial (embeddings); (iii) extração de informação (NER, análise semântica); (iv) indexação propriamente dita (estruturação dos vetores ou termos em índices pesquisáveis); e (v) recuperação (busca e ranqueamento de resultados).

Análise Semântica/Sentimentos: Identifica o contexto do documento, ou seja, vai além da estrutura gramatical (sintaxe) e busca “compreender” o significado e a intenção no contexto do documento.

Os modelos de inteligência artificial utilizados em aplicações de PLN são previamente treinados com grandes volumes de dados textuais. Dessa forma, conforme apresentado por Caseli e Nunes (2024), esses modelos desenvolvem representações estatísticas da linguagem, permitindo identificar padrões linguísticos, relações entre termos e estruturas semânticas presentes nos documentos.

Em uma interpretação simplificada, pode-se afirmar que os modelos de PLN são capazes de reconhecer aspectos relacionados ao vocabulário e às estruturas gramaticais de um idioma, realizando análises textuais que identificam termos, classificações gramaticais (artigos, pronomes, substantivos, adjetivos, verbos etc.) e relações de contexto informacional. Considerando esse entendimento, a Figura_6 apresenta uma simulação de um possível resultado de indexação baseada em técnicas de PLN.

Figura 6 - Indexação por PLN

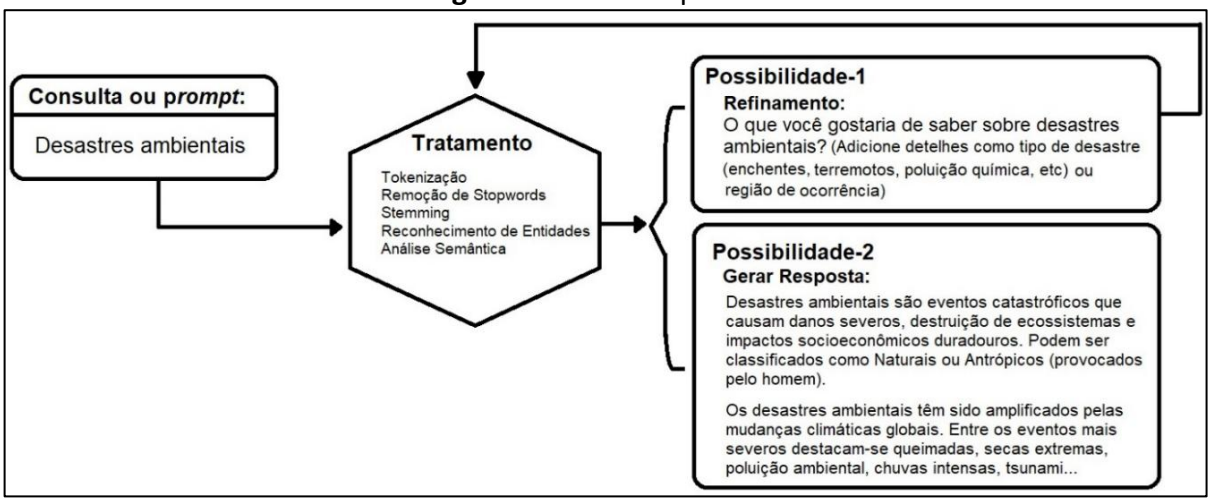


Fonte: elaborada pelos autores (2026)

No modelo apresentado, a técnica de PLN realiza um processo de indexação com critérios muito similares aos utilizados por humanos, dentre os quais é possível identificar: a) O assunto: preservação digital; b) O objetivo: acesso; c) O problema ou desafio: obsolescência tecnológica e d) A técnica ou estratégia: emulação e migração. Esta técnica de indexação facilita a recuperação de informação em repositórios digitais, permitindo que o documento seja localizado tanto por termos específicos quanto por contexto.

A Figura_7 esquematiza, de forma suscinta, um SRI que implementa técnica de indexação com o uso de PLN e IA generativa:

Figura 7 - Consulta por PLN



Fonte: elaborada pelos autores (2026)

O usuário faz uma pergunta (Consulta ou prompt); o SRI realiza o tratamento da pergunta do usuário para em seguida casá-la com os as informações (já indexadas) do banco de dados. A análise semântica permite ao desenvolvedor (programador) incluir rotinas de programação que identificam o nível de especificidade da pergunta, dessa forma o SRI poderá ser direcionado a estratégia de refinamento (Possibilidade-1) ou gerar automaticamente uma

resposta com base na pergunta do usuário (Possibilidade-2). Também é possível unir as duas estratégias, gerando uma resposta inicial e solicitando mais detalhes para o refinamento da resposta; simulando um diálogo entre usuário e SRI.

Em Possibilidade-2, observa-se a atuação da análise semântica realizada pela indexação com PLN, onde termos que não estão presente nas palavras-chave da pergunta do usuário (desastres e ambientais) compõem a resposta apresentada por possuírem relação semântica com o tema ou temática abordado (ecossistema, impactos socioeconômicos, queimadas, secas extremas, etc). “Os algoritmos de PLN podem decompor o texto em palavras-chave, frases e até mesmo ideias, tornando qualquer pesquisa mais eficiente e relevante” (Khalifa, 2025, p. 2).

A indexação baseada em PLN para recuperação de informação também supera as limitações da busca por palavras-chave exatas, pois permite que o SRI “compreenda” o significado, o contexto e até mesmo a intenção por traz dos textos, transformando dados não estruturados em conteúdos organizadas, facilitando o acesso. Por exemplo, ao buscar por “desastres ambientais”, o SRI poderá apresentar como resposta documentos que falem sobre “mudanças climáticas”, “alagamentos”, “deslizamento de terra”, etc. mesmo estes termos não estando na pergunta do usuário.

3.4 IA GENERATIVA, RECUPERAÇÃO SEMÂNTICA E RECUPERAÇÃO DE INFORMAÇÃO COM GERAÇÃO AUMENTADA (RAG)

IA Generativa refere-se a modelos capazes de produzir novos conteúdos como textos, imagens ou áudios, a partir de padrões aprendidos em grandes volumes de dados, distinguindo-se de modelos discriminativos, voltados apenas à classificação de dados já existentes (Lewis *et al.*, 2020). No contexto da indexação e recuperação de documentos digitais, a RI com Geração Aumentada (RAG) combina um mecanismo de recuperação de informação com um modelo de linguagem generativa, fazendo com que a geração de respostas esteja condicionada a documentos recuperados de um índice vetorial denso, não apenas pelo conhecimento paramétrico do modelo (Lewis *et al.*, 2020). A recuperação semântica, por sua vez, consiste na busca por similaridade de significado entre a consulta do usuário e os documentos indexados, superando a exigência de correspondência exata de palavras-chave típica do índice invertido (ver seção 3.1).

Ainda é preciso considerar que documentos arquivísticos digitais podem possuir uma estrutura com alta complexidade, a exemplo de enlaces estabelecidos por referências

cruzadas, o que demanda por SRI com abordagens mais sofisticadas. A RI com geração aumentada (RAG) representa uma evolução significativa nos procedimentos de indexação e busca, ao combinar o poder de recuperação contextual com a capacidade interpretativa de sistemas de inteligência artificial generativa. Zhang *et al. (*2025) demonstraram a eficácia de abordagens *multi-round* RAG na análise de documentos jurídicos, com estrutura complexa, apresentando dependências entre cláusulas, referências cruzadas e ambiguidades contextuais.

No âmbito das proposições apresentadas, entendemos que, no âmbito de arquivos, a arquitetura RAG pode implementar ciclos iterativos de recuperação (teste, refinamento e melhoramento em várias etapas consecutivas), seguindo o modelo proposto por Zhang *et al.* (2025) e a lógica de refinamento iterativo discutida por Lewis *et al.* (2020), na qual uma consulta ou tarefa de análise é inicialmente processada, gerando uma primeira rodada de recuperação de documentos ou metadados relevantes; em seguida, um modelo de linguagem examina o conteúdo recuperado, identificando inconsistências, lacunas de contexto ou dependências não satisfeitas; caso sejam identificadas, o sistema refina sua estratégia de busca e executa novas rodadas de recuperação até que o contexto se torne suficientemente completo para uma análise precisa.

Esse processo iterativo se diferencia fundamentalmente das abordagens tradicionais de RI por palavras-chave, pois permite que o sistema compreenda não apenas a relevância léxica, mas também a relevância semântica e contextual dos documentos recuperados; o que pode ser percebido como uma estratégia para a garantia do princípio da organicidade arquivística, já que, em uma proposição baseada no refinamento apresentado por Zhang *et al.* (2025), o SRI teria a “capacidade” de identificar inconsistência, conflitos entre dados ou mesmo contradições registradas em documentos que compõem um fundo arquivístico.

Ainda é possível estabelecer uma relação com a garantia do princípio da proveniência ao observar que a arquitetura RAG reduz o risco de alucinações (respostas geradas sem lastro documental), uma vez que a geração de respostas fica ancorada em conteúdos efetivamente recuperados na base de dados local, não apenas no conhecimento paramétrico do modelo (Lewis *et al.*, 2020). Essa ancoragem amplia tanto a transparência quanto à origem da informação recuperada, possibilitando a rastreabilidade.

4 CONSIDERAÇÕES FINAIS

O campo de RI, no contexto da arquivologia, apresenta especificidades provenientes de pressupostos da área, que refletem a complexidade de princípios e requisitos arquivísticos. SRI precisam "dialogar" com demais sistemas arquivísticos informatizados, a exemplo de SIGAD e RDCArq, para que possa apresentar, sempre que solicitado, informações referentes a cadeia de custódia, proveniência e rastreabilidade de documentos no fundo arquivístico, particularmente aqueles que compõem a proposição de solução à pergunta do usuário e apresentam interdependências (relação orgânica).

As técnicas de indexação de documentos arquivísticos digitais representam uma evolução significativa nos procedimentos de RI no âmbito de arquivos, ao combinar o poder da análise semântica com a capacidade interpretativa de sistemas de inteligência artificial generativa, possibilitando a recuperação por contexto, que busca uma compreensão mais completa entre a demanda informacional do usuário e os documentos.

No âmbito da interdisciplinaridade, ao considerar o letramento arquivístico digital como a capacidade crítica de compreender, utilizar e interagir de forma ética e eficaz com as tecnologias digitais, é necessário destacar que SRI, como sistema, é composto por um conjunto de elementos interconectados, no qual a atividade de indexação, em um processo de sinergia, trabalha conjuntamente com outros componentes para atingir o objetivo específico: recuperar de forma fácil e rápida, dentro do universo de informações existentes, apenas àquelas efetivamente relevantes para atender a demanda informacional do usuário.

Desta forma, este estudo representa uma oportunidade promissora de contribuir com debates no processo de interdicisplinaridade entre a arquivologia contemporânea à sociedade digital e as TIC, em uma perspectiva de letramento arquivístico digital que contempla, como ponto de discussão, o aprimoramento de SRI para ambientes arquivísticos digitais com o uso de tecnologias que compõem a "Biota de IA" em conformidade com princípios e requisitos da arquivísticos.

REFERÊNCIAS

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação de informação:** conceitos e tecnologias das máquinas de busca. 2. ed. Porto Alegre. Bookman. 2013.

BELLOTTO, Heloísa Liberalli. **Arquivos permanentes**: tratamento documental. 4. ed. Rio de Janeiro: Editora Fundação Getúlio Vargas, 2006.

CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (org.). **Processamento de Linguagem Natural**: Conceitos, Técnicas e Aplicações em Português. 2 ed. BPLN, 2024. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao. Acesso em: 12 dez. 2025.

CESARINO, Maria Augusta da Nóbrega. Sistemas de recuperação da informação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, 1985. Disponível em: https://periodicos.ufmg.br/
index.php/reb/article/view/36507/28553. Acesso em: 24 maio 2026.

DORENELES, Sânderson Lopes; CÔRREA, Renato Fernandes; FLORES, Daniel. Subsídios para política pública de gestão arquivística de documentos digitais: revisão sistemática de literatura. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 28, e46277, 2023. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/46277. Acesso em: 10 set. 2026.

FUJITA, Mariângela Spotti Lopes. A política de indexação para representação e recuperação da informação. *In*: GIL LEIVA, Isidoro; FUJITA, Mariângela Spotti Lopes (org.). **Política de indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012. p. 17–28. Disponível em: DOI: https://doi.org/10.36311/2012.978-85-7983-199-7.p17-28. Acesso em: 10 set. 2026.

FUJITA, Mariângela Spotti Lopes; RODRIGUEZ, Sonia Maria Troitiño. Política de indexação em arquivos de instituições de saúde. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, [*s. l.*], v. 14, n. 1, 2019. Disponível em: DOI: https://doi.org/10.22478/ufpb.1981-0695.2019v14n1.44760 Acesso em: 10 set. 2026.

HADDI, Emma; LIU, Xiaohui; SHI, Yong. The Role of Text Pre-processing in Sentiment Analysis. **Procedia Computer Science**, [*s. l.*], v. 17, p. 26-32, 2013. Disponível em: https://doi.org/10.1016/j.procs.2013.05.005. Acesso em: 03 nov. 2025.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática. Tradução de Antonio Agenor Briquet de Lemos. Brasília: Briquet de Lemos/Livros, 1993. 347p.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática. 2. ed. rev. e atual. Tradução de Antonio Agenor Briquet de Lemos. Brasília: Briquet de Lemos/Livros, 2004. 452p.

LEWIS, Patrick; PEREZ, Ethan; PIKTUS, Aleksandra; PETRONI, Fabio *et al*. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. **Advances in Neural Information Processing Systems**, [*s. l.*], v. 33, p. 9459-9474, 2020. Disponível em: https://arxiv.org/abs/2005.11401. Acesso em: 10 set. 2026.

MARQUES, Juliana Ferreira; MEDEIROS, José Washington de Morais. “Do segredo ao acesso”: entre a custódia documental e a gestão da informação à luz dos paradigmas arquivísticos. **Perspectivas em Gestão e Conhecimento**, [*s. l.*], v. 8. n. 3. p. 4–19. set. dez. 2018. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/pgc/article/view/38746/21764. Acesso em: 08 fev. 2025.

MOREIRA, Viviane P. Recuperação de informação. *In*: CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (org.). **Processamento de linguagem natural**: conceitos, técnicas e aplicações em português. 1. ed. [*S. l.*]: BPLN, 2023. cap. 16.

MUNDET, José Ramón Cruz**. Breve História de Los Documentos.** Alianza Editorial. Madri - Espanha. 2024. Disponível em https://www.alianzaeditorial.es/primer_capitulo/breve-historia-de-los-documentos.pdf. Acesso em: 12 dez. 2025.

POMBO, Olga. Interdisciplinaridade e integração dos saberes. **Liinc em Revista**, Rio de Janeiro, v. 1, n. 1, p. 3–15, mar. 2005. DOI: https://doi.org/10.18617/liinc.v1i1.186. Disponível em: https://revista.ibict.br/liinc/article/view/3082. Acesso em: 10 set. 2026.

RAMISCH, Renata; RAMISC, Carlos; VILLAVICÊNCIO, Aline**.** Expressões multipalavras: fogo de palha ou osso duro de roer? *In*: CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe. **Processamento de Linguagem Natural**: conceitos, técnicas e aplicações em português. 2. ed. São Carlos: [*S. l.*], 2024. p. 98-120.

SANTOSO, Joan; SETIAWAN, Esther Irawati; PURWANTO, Christian Nathaniel; YUNIARNO, Eko Mulyanto *et al*. Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory**. Expert Systems With Applications**, v. 176, 114856, 2021. Disponível em: https://doi.org/10.1016/j.eswa.2021.114856. Acesso em: 10 set. 2026.

SARACEVIC, Tefko. Information science. **Journal of the American Society for Information Science**, v. 50, n. 12, p. 1051-1063, 1999. Disponível em: DOI: https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1051::AID-ASI2>3.0.CO;2-Z. Acesso em: 10 set. 2026.

ZHANG, Ying; XIAO, Gang. Named Entity Recognition Datasets: A Classification Framework. **International Journal of Computational Intelligence Systems**, v. 17, n. 1, 2024. Disponível em: https://doi.org/10.1007/s44196-024-00456-1. Acesso em: 10 set. 2026.

ZHOU, Yu-Jia; YAO, Jing; DOU, Zhi-Cheng; WU, Ledell; WEN, Ji-Rong. DynamicRetriever: A Pre-trained Model-based IR System Without an Explicit Index. **Machine Intelligence Research**, v. 20, n. 2, p. 276-288, 2023. Disponível em: https://doi.org/10.1007/s11633-022-1373-9. Acesso em: 10 set. 2026.

ZHANG, Wenlong; LI, Yunshen; LOU, Jiong; ZHOU, Qiang et al. Análise abrangente de documentos legais: uma abordagem RAG multi-rodada. *In*: CONFERÊNCIA INTERNACIONAL

ACM ICMR DE PESQUISA E DESENVOLVIMENTO EM RECUPERAÇÃO DE INFORMAÇÃO, 48., 2025, Chicago, IL. **Anais** [...] New York: ACM, 2025. p. 1840-1847. Disponível em: DOI: 10.1145/3731715.3733451. Acesso em: 10 set. 2026.