

Organização e Representação do Domínio da Sustentabilidade: Um Arranjo Metodológico

Maria Carolina Coradini¹, Milena Ambrosio Telles², Eder Cavalcanti Coimbra³, Isaac Souza de Miranda Jr⁴, Gabriela Wick-Pedro⁵

Resumo. *Nos últimos anos, tem-se observado um crescente debate e uma proliferação de informações sobre sustentabilidade em todo o mundo. No Brasil, o discurso sobre sustentabilidade abrange múltiplos domínios — incluindo os setores jornalístico, corporativo e científico. Consequentemente, há uma necessidade premente de sistematizar o conhecimento acumulado sobre esse tema. Como iniciativa do Laboratório de Informação para Sustentabilidade (LIS) do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), este estudo empregou Linguística de Corpus e Processamento de Linguagem Natural (PLN) para mapear e organizar o conhecimento linguístico sobre sustentabilidade no Brasil. Utilizando um corpus de aproximadamente 50 milhões de palavras, abrangendo cinco gêneros textuais, candidatos a termo foram extraídos e posteriormente validados por especialistas do IBICT. Esses termos foram organizados em conjuntos temáticos e classificados semanticamente com apoio do Agrovoc. O mapa conceitual resultante foi transformado em uma estrutura dinâmica, publicada no portal do LIS. Este artigo apresenta o desenvolvimento desse produto terminológico por meio de uma proposta metodológica interdisciplinar, detalhando os resultados e os desafios enfrentados.*

¹Linguista, bolsista da Embrapa Meio Ambiente, atua em Linguística de Corpus e Representação do Conhecimento. Colaboradora da Comissão Permanente de Vocabulários Controlados, Agroterminologias e Agrossemântica (GTermos). E-mail: m.linacoradini@gmail.com - ORCID: <https://orcid.org/0000-0002-1691-6879>

²Analista da Embrapa, Assessoria de Serviços Estratégicos (AEST), coordenadora da Comissão Permanente de Vocabulários Controlados, Agroterminologias e Agrossemântica (GTermos) e editora do português brasileiro no Agrovoc. E-mail: milena.telles@embrapa.br - ORCID: <https://orcid.org/0000-0001-9523-9724>

³Linguista na Cognizant, colaborador da Comissão Permanente de Vocabulários Controlados, Agroterminologias e Agrossemântica (GTermos) da Embrapa. E-mail: edercocoimbra@gmail.com - ORCID: <https://orcid.org/0000-0002-7138-4967>

⁴Doutorando no Departamento de Letras da Universidade Federal de São Carlos. Pesquisador em Linguística Computacional e membro do NILC (Núcleo Interinstitucional de Linguística Computacional) e do grupo de pesquisa LeGOS (Léxico, Gramática, Opinião e Sentimento). E-mail: isc_jr@live.com - ORCID: <https://orcid.org/0000-0002-4004-3182>

⁵Linguista na Atento, atua em Análise e Processamento de Linguagem Natural e em produtos terminológicos baseados em Linguística de Corpus. E-mail: gabiwick@gmail.com. - ORCID: <https://orcid.org/0000-0002-7332-4482>

Palavras-chave: mapa conceitual; sustentabilidade; processamento de linguagem natural; linguística de corpus

1. Introdução

Nas últimas décadas, a produção de informação se intensificou, impulsionada pela facilidade de produção e pela velocidade do compartilhamento digital. Encontrar informações sobre qualquer tema hoje é simples: há inúmeros portais na *web* que oferecem acesso gratuito. Os discursos emergem de múltiplas perspectivas, como as corporativa, política e acadêmica. Além disso, a ponte entre a comunidade científica e o público em geral se fortalece à medida que a divulgação científica evolui, garantindo que descobertas científicas e técnicas cheguem à sociedade de forma rápida e acessível pelos meios digitais.

Se, por um lado, o conhecimento especializado é agora produzido e disseminado em uma escala sem precedentes, por outro, sua sistematização, organização e representação estruturada permanecem desafios centrais para a Ciência da Informação (CI). Historicamente, a área tem se dedicado ao estudo de princípios, métodos e instrumentos voltados ao tratamento e à organização do conhecimento, reconhecendo que a mera disponibilidade de informação não garante sua inteligibilidade, interoperabilidade ou recuperação eficiente (Hjørland, 2008; Dahlberg, 1978). Nesse sentido, a representação do conhecimento constitui um passo fundamental para transformar fluxos informacionais dispersos em estruturas operáveis cognitivamente e computacionalmente (Kobashi, 2008).

A Linguística, particularmente suas vertentes computacional e de *corpus*, oferece contribuições complementares e fundamentais para enfrentar esses desafios, ao fornecer métodos empíricos e ferramentas analíticas para extrair e modelar o conhecimento linguístico-conceitual contido em grandes volumes de textos especializados. Essa disciplina enfatiza o estudo sistemático de padrões linguísticos em contextos reais de uso, por meio da utilização de *corpora* representativos e de técnicas de Processamento de Linguagem Natural (PLN), permitindo a identificação de termos específicos de domínio, relações semânticas e estruturas discursivas de maneira escalável e baseada em evidências.

O domínio da sustentabilidade insere-se de forma emblemática nesse cenário. Há algum tempo, a sustentabilidade tornou-se eixo estruturante de agendas políticas globais, estratégias corporativas, programas científicos e movimentos sociais. No Brasil, o discurso sobre sustentabilidade abrange diversas esferas — jornalística, corporativa, técnica e científica - resultando em um campo discursivo heterogêneo e em constante

evolução. Essa heterogeneidade evidencia a necessidade de ferramentas capazes de mapear, modelar e representar sistematicamente o conhecimento produzido sobre o tema.

Sob a perspectiva da Organização do Conhecimento, domínios complexos e transversais, como o da sustentabilidade, exigem abordagens que integrem análise conceitual, modelagem semântica e representação formal (Smiraglia, 2014; Hjørland, 2017). O desenvolvimento de mapas conceituais, vocabulários controlados e ontologias é uma estratégia bem estabelecida para definir explicitamente as relações entre conceitos, apoiar sistemas de recuperação da informação e subsidiar arquiteturas informacionais digitais. Esses instrumentos são particularmente relevantes em contextos de Ciência, Tecnologia e Inovação (CT&I), onde a interoperabilidade semântica e a mediação qualificada da informação são essenciais para a circulação eficaz do conhecimento.

Do ponto de vista terminológico, a identificação e a organização dos termos que estruturam um domínio são uma etapa essencial para sua modelagem conceitual. A Terminologia contemporânea, especialmente a partir das contribuições da Teoria Comunicativa da Terminologia (Cabré, 1999) e das abordagens sociocognitivas (Temmerman, 2000), reconhece que os termos não são meros rótulos estáticos, mas unidades linguístico-conceituais situadas em práticas discursivas específicas. Assim, a extração e a validação de termos a partir de *corpora* especializados permitem capturar empiricamente a configuração dinâmica de um campo do saber.

O arranjo metodológico apresentado neste trabalho insere-se nesse contexto. Como parte de uma iniciativa do Laboratório de Informação para Sustentabilidade (LIS) do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), este estudo fundamenta-se em CI, Linguística de Corpus e PLN para mapear e organizar o conhecimento linguístico sobre sustentabilidade no Brasil. O estudo integra o projeto ITBr, desenvolvido pelo Grupo de Pesquisa ESFERA, que visa promover um ecossistema de informação em ciência, tecnologia, inovação e sustentabilidade, por meio de ações e produtos voltados à infraestrutura informacional desses domínios.

Neste trabalho, o mapa conceitual destaca-se como uma ferramenta para modelagem e visualização do conhecimento, organizado com base na terminologia especializada extraída de um *corpus* de quase 50 milhões de palavras, composto por textos de diversos gêneros e fontes. Entendido como um sistema estruturado de conceitos semanticamente inter-relacionados, representados por termos, atua, assim, como um instrumento de organização do conhecimento, capaz de apoiar sistemas de recuperação da informação e arquiteturas digitais.

Nesse sentido, este artigo apresenta o desenvolvimento de um mapa conceitual para a sustentabilidade no Brasil, detalhando as etapas metodológicas, os resultados e os desafios enfrentados. O trabalho busca contribuir para o debate sobre princípios, metodologias e instrumentos para o processamento, a organização e a representação do

conhecimento em domínios interdisciplinares complexos, oferecendo um modelo replicável para iniciativas semelhantes em contextos institucionais.

2. Metodologia: uma proposta interdisciplinar

O mapeamento conceitual do domínio foi desenvolvido a partir da integração da proposta metodológica de Di Felippo e Almeida (2010) — uma instanciação do modelo genérico de pesquisa em PLN proposto por Dias-da-Silva (2006) — com ferramentas da CI para análise bibliométrica e mineração de textos. Essa abordagem fundamenta o estudo terminológico em um *corpus* linguístico, utilizando Linguística de Corpus para o desenvolvimento de produtos orientados por dados, como glossários e redes de palavras (*wordnets*). Ao integrar extração automática de candidatos a termos, validação por especialistas e classificação semântica apoiada pelo Agrovoc (vocabulário controlado da Organização das Nações Unidas para a Alimentação e a Agricultura - FAO), a metodologia produz uma representação gráfica e dinâmica do domínio. Essa síntese evidencia a convergência entre práticas terminológicas, modelagem conceitual e representação do conhecimento, alinhando-se às tendências contemporâneas que integram fundamentos epistemológicos a aplicações práticas.

Ao colocar em foco o domínio especializado através de um *corpus* de referência (um conjunto pré-estruturado de textos), a metodologia inicia-se com a definição das estratégias de coleta e das expressões de busca necessárias para explorar as fontes que subsidiarão a extração do conhecimento léxico-conceitual.

Para estruturar essas expressões de busca, foi conduzido um estudo inicial utilizando o VOSViewer (CWTS, 2023) sobre a literatura global recuperada da Web of Science⁶, com base em dois termos centrais: “sustentabilidade” e “informação para sustentabilidade”. Esses testes preliminares revelaram os termos mais frequentes no domínio e seus padrões organizacionais. A rede bibliométrica apresentada na Figura 1 ilustra um dos resultados dessa análise.

⁶ <https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/>

Esses resultados proporcionaram percepções iniciais para a equipe do LIS, que atuou como especialista de domínio, e serviram como parâmetro para determinar os termos-chave utilizados na recuperação de textos especializados. Para garantir abrangência e relevância durante a compilação do *corpus*, os especialistas optaram por uma estratégia de busca ampla, combinando um conjunto limitado de termos e suas variações. Os termos “sustentabilidade”, “sustentável” e “sustentáveis” serviram como expressões de busca para os quatro tipos de fontes: notícias, acadêmicas, técnicas e páginas da *web*. Para páginas da *web*, as consultas também incluíram “desenvolvimento sustentável”, “estratégia sustentável” e “estratégias sustentáveis”. As bases de dados foram selecionadas em colaboração com os especialistas, e a estratégia de busca foi refinada para cada repositório, considerando as características específicas de cada ambiente.

Além da representatividade e da relevância, o balanceamento é outro critério crucial para a construção de um *corpus* (Sinclair, 2004). Para garantir que um *corpus* seja minimamente representativo e evite vieses, ele deve incluir uma variedade de gêneros e registros textuais. Por isso, o *corpus* da sustentabilidade foi construído com textos de três categorias discursivas: acadêmica, não-acadêmica e técnica.

Os textos acadêmicos foram recuperados por meio do Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto (Oasisbr, 2026)⁷. A porção não acadêmica foi coletada de páginas da *web* e veículos de notícias. Os textos técnicos, por sua vez, foram coletados de bases de dados institucionais brasileiras, contendo artigos e relatórios técnicos.

A coleta automatizada dos textos acadêmicos foi realizada utilizando as bibliotecas Selenium (Goldberg et al., 2025) e BeautifulSoup (Richardson, 2025), que foram empregadas para navegação, recuperação e extração estruturada dos dados disponíveis nas páginas da *web*. No entanto, como os textos estavam hospedados em arquivos PDF acessíveis por *links* externos — em vez de estarem incorporados ao conteúdo HTML — a biblioteca PyPDF2 (Fenniak; Thomas, 2025) foi necessária para extrair e processar o conteúdo textual.

Para coletar conteúdo de *sites* padrão, como blogues e portais institucionais, utilizou-se a ferramenta BootCat. Essa ferramenta extrai automaticamente os 100 textos mais relevantes para cada expressão de busca aplicada ao Google.

Para a porção jornalística do *corpus*, quatro portais de notícias *online* foram selecionados: Folha, G1, IstoÉ e UOL. Esses textos também foram extraídos automaticamente usando Selenium e BeautifulSoup e processados por meio de *scripts* de Python. Diferentemente do *corpus* acadêmico, não houve necessidade de uma biblioteca específica para PDF, pois o conteúdo jornalístico estava totalmente incorporado ao HTML das páginas. As expressões de busca pré-definidas foram aplicadas por meio do motor de busca interno de cada *site* para recuperar os artigos. A extração dos textos foi, então, realizada em cada página individualmente, utilizando expressões regulares (regex) adaptadas para identificar as *tags* HTML que continham o corpo do artigo.

Os textos técnicos foram coletados individualmente ou automaticamente, quando possível, a partir de 15 bases de dados brasileiras, associadas às seguintes instituições: Conselho Empresarial Brasileiro para o Desenvolvimento Sustentável (CEBDS)⁸, Comissão Econômica para a América Latina e o Caribe (CEPAL)⁹, Empresa Brasileira de Pesquisa Agropecuária (Embrapa)¹⁰, Instituto Brasileiro do Meio Ambiente e dos Recursos Renováveis (Ibama)¹¹, Instituto Brasileiro de Geografia e Estatística (IBGE)¹², Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio)¹³, Instituto Akatu¹⁴,

⁷ <https://oasisbr.ibict.br/>

⁸ <https://cebds.org/>

⁹ <https://www.cepal.org/>

¹⁰ www.bdpa.cnptia.embrapa.br

¹¹ <https://www.ibama.gov.br/sophia/>

¹² <https://www.ibge.gov.br/>

¹³ <https://biblioteca.sophia.com.br/terminal/9836>

¹⁴ <https://akatu.org.br/>

Instituto Ethos¹⁵, Instituto de Pesquisa Econômica Aplicada (IPEA)¹⁶, Observatório 2030¹⁷, Pacto Global¹⁸, Portal ODS Brasil¹⁹ e Serviço Brasileiro de Respostas Técnicas (SBRT)²⁰.

A coleta automatizada nas bases de dados de textos técnicos seguiu procedimento semelhante ao dos textos acadêmicos, utilizando as mesmas ferramentas de extração e processamento. Especificamente, o processo envolveu a identificação e a coleta de *links*, o *download* de arquivos e o subsequente processamento dos PDFs, utilizando as bibliotecas Selenium (Goldberg et al., 2025), BeautifulSoup (Richardson, 2025) e PyPDF2 (Fenniak; Thomas, 2025).

Após a compilação do *corpus* de referência, as etapas seguintes incluíram: conversão dos textos de vários formatos para .txt; limpeza semiautomática para remover *links*, figuras, tabelas e elementos corrompidos, bem como cabeçalhos e outros elementos pré e pós-textuais que pudessem distorcer a contagem de palavras; e, finalmente, a extração de listas de lexias simples e compostas (n-gramas), a partir das quais os termos centrais do domínio foram posteriormente identificados.

O ponto de corte para a extração das listas de palavras foi determinado pela normalização da frequência para 100.000 ocorrências, calculada como $\frac{\text{cardinalidade do corpus}}{100.000} + 1$. Com base nesse limiar, qualquer palavra com frequência igual ou superior ao valor resultante foi identificada como um candidato a termo em potencial. Dado o total de 49.864.770 *tokens*, candidatos a termo sob a forma de lexias simples (unigramas) e compostas, com frequência mínima de 500 ocorrências, foram extraídos do *corpus* utilizando o AntConc (Anthony, 2020), uma ferramenta de análise textual. Os n-gramas variaram de dois a oito elementos (2-gramas a 8-gramas).

Das listas de palavras geradas pela ferramenta foram filtrados termos da língua geral e expressões sem significado especializado no domínio. Após esse refinamento, as listas foram exportadas para planilhas para validação dos especialistas. Os resultados foram então consolidados em um único conjunto de dados para a etapa subsequente.

Termos validados por pelo menos 80% da equipe de especialistas — valor aqui tratado como indicativo de consenso — foram selecionados para classificação semântica e mapeamento conceitual. As categorias amplas do Agrovoc²¹ serviram como ponto de

¹⁵ <https://www.ethos.org.br/>

¹⁶ <https://www.ipea.gov.br/>

¹⁷ <https://observatorio2030.com.br/>

¹⁸ <https://www.pactoglobal.org.br/>

¹⁹ <https://odsbrasil.gov.br/>

²⁰ <https://www.respostatecnica.org.br/>

²¹ Disponível em <https://www.fao.org/agrovoc/top-concepts-agrovoc>.

partida. Esse vocabulário foi escolhido por sua ampla cobertura de tópicos agrícolas e alimentares, pela correspondência translinguística robusta dos mais variados termos e sua reputação consolidada como ferramenta de organização do conhecimento.

Após a validação das classificações iniciais pelos especialistas, os termos foram categorizados nas três dimensões da sustentabilidade: econômica, ambiental e social (Veiga, 2010). Termos inter-relacionados foram então mapeados para garantir conectividade lógica dentro do modelo gráfico final.

Com a validação dessas duas propostas de classificação, os dados foram organizados em planilhas de duplas (pares de termos em colunas adjacentes) para gerar representações gráficas em *softwares* especializados. Essas planilhas foram inicialmente processadas utilizando yEd e XMind, e cada resultado foi revisado e ajustado manualmente para maior legibilidade. Finalmente, os dados foram processados por meio do Vis.js, uma biblioteca JavaScript para visualização dinâmica de dados, permitindo a integração de um mapa conceitual interativo ao portal do LIS.

A metodologia descrita até este ponto está sintetizada no fluxograma apresentado na Figura 2.

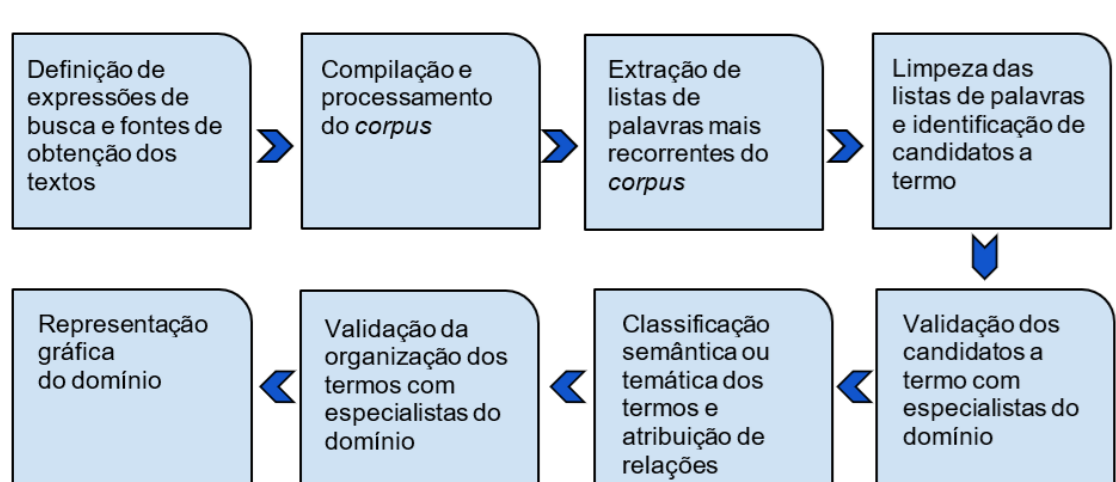


Figura 2. Fluxograma da metodologia utilizada para desenvolver o mapa conceitual.

Fonte: Os autores.

Os resultados de cada uma dessas etapas são descritos a seguir.

3. Resultados

O desenvolvimento de um mapa conceitual, seguindo a metodologia descrita anteriormente, gera produtos específicos em cada etapa. O mais significativo deles é o

próprio *corpus* de referência²² que, dada sua complexidade e importância, serve como a base primordial de todo o processo. Isso ocorre porque o *corpus* fornece todos os dados essenciais para o mapa, incluindo os termos mais frequentes do domínio, sua organização temática e suas relações semânticas.

Não surpreende, portanto, que a construção do *corpus* seja uma das etapas que envolve o conjunto mais extenso de subprocessos. Ela começa com a definição das expressões de busca, projetadas para produzir um volume ideal de material que seja relevante. Essas expressões passam, por sua vez, por testes, para verificar sua qualidade e abrangência, permitindo que sejam refinadas iterativamente. Só então os textos são recuperados, para serem subsequentemente filtrados, organizados, limpos e processados para a extração léxico-conceitual.

O *corpus* resultante compreende aproximadamente 50 milhões de palavras, distribuídas em 14.128 documentos em língua portuguesa. Esses documentos de diversos gêneros foram obtidos por meio de plataformas de acesso *online* aberto e gratuito. A Tabela 1 lista os tipos de texto e as fontes consultadas, enquanto a Tabela 2 detalha as estratégias de busca e seus respectivos resultados.

Tabela 1. Informações sobre o *corpus* e distribuição dos textos

Classe	Tipo	Textos	Palavras	% Relativa
Acadêmica	Artigos	3.677	27.720.771	55,6
Não-acadêmica	Conteúdo da <i>web</i>	558	704.347	1,4
	Notícias	9.149	7.575.558	15,2
Técnica	Relatórios e outros documentos técnicos	744	13.846.094	27,8
Total	-	14.128	49.846.770	100

Fonte: Os autores

²² Disponível em: https://github.com/MapeamentoC/Informacao_Para_Sustentabilidade.

Tabela 2. Expressões de busca e resultados para recuperação de documentos.

Classe	Tipo	Fonte	Filtros de busca	Expressões de busca
Acadêmica (27.7 mi)	Artigos	Oasis.br	No título	“sustentável”, “sustentáveis”, ou “sustentabilidade”
Não-acadêmica (8.3 mi)	Conteúdo da <i>web</i> (0. 7 mi)	Bootcat/Google	Textos não-acadêmicos (verificação humana)	“sustentável”, “sustentáveis”, “sustentabilidade”, “desenvolvimento sustentável”, “estratégias de sustentabilidade”, “estratégias sustentáveis”, ou “sustentabilidade no Brasil”
	Notícias (7.6 mi)	Folha	No título	“sustentável,” “sustentáveis”, ou “sustentabilidade”
		G1		
		IstoÉ		
		UOL		
Técnica (13.8 mi)	Relatórios e outros documentos técnicos	Instituto Akatu	No título	“sustentável,” “sustentáveis”, ou “sustentabilidade”
		CEBDS		
		Instituto Ethos		
		ICMBIO		
		IPEA		
		SBRT	N/A	
		CEPAL	No título ou resumo	
		EMBRAPA	No título Tipo: relatório; técnico	
		IBAMA		
		IBGE	No título Tipo: livro	
		Observatório 2030	Todos os textos Tipo: entrevistas; estudos; outros	N/A
		Pacto Global	Todos os textos em português	N/A
		Portal ODS		

Fonte: Os autores

O *corpus* foi balanceado entre as categorias que o constituem, de modo que os textos acadêmicos representam pouco mais da metade do volume total (56%), enquanto os textos não acadêmicos, somados aos técnicos, compreendem 44% da coleção.

As listas de palavras contendo os candidatos a termo foram extraídas dessa grande coleção de textos sobre sustentabilidade. Conforme descrito na Seção 2, essas listas foram geradas utilizando um ponto de corte baseado no tamanho do *corpus*. Em razão de sua escala massiva, as listas geradas continham um elevado número de n-gramas — 29.761 no total. Estes variavam desde lexias simples (1-gramas, como 'sustentabilidade') até compostas, de dois a oito elementos (por exemplo, 'recursos hídricos'). Por essa razão, as listas brutas passaram por rodadas de limpeza automática e humana para refinar os dados antes da validação. Os resultados parciais e finais dessa etapa são apresentados na Tabela 3.

Tabela 3. Resultados da extração e do refinamento das listas de palavras

n-gramas	Número de n-gramas extraídos do corpus	Número de n-gramas depois da primeira rodada de limpeza (automática)	Número de n-gramas depois da segunda rodada de limpeza (humana)	Número de n-gramas depois da terceira rodada de limpeza, realizada por especialistas de domínio (humana)
1-gramas	9.231	2.990	1.168	172
2-gramas	14.218	766	505	387
3-gramas	4.285	509	203	150
4-gramas	930	258	36	31
5-gramas	429	165	8	6
6-gramas	275	109	1	1
7-gramas	213	98	1	1
8-gramas	180	74	0	0
TOTAL	29.761	4.969	1.922	748

Fonte: Os autores

Os 748 candidatos a termo mantidos após a terceira rodada de limpeza foram enviados ao grupo de especialistas para validação. Destes, 746 foram validados como termos da área por ao menos um especialista. Para a primeira versão do mapa conceitual,

apenas termos com pelo menos 80% de consenso entre a equipe foram incluídos, totalizando 417 termos.

O mapa conceitual foi inicialmente estruturado a partir da classificação dos termos de acordo com as categorias semânticas amplas do Agrovoc, a saber: atividades, características, disciplinas, entidades, estados, estágios, estratégias, eventos, fatores, fenômenos, grupos, localidades, medidas, métodos, objetos, organismos, períodos, processos, produtos, propriedades, recursos, sistemas, sítios, substâncias, tecnologias. Os termos foram agrupados nessas categorias e enviados para validação, resultando no seguinte panorama:

Tabela 4. Número de termos por categoria semântica ampla do Agrovoc

atividades	105	fenômenos	50	organismos	2	sistemas	24
disciplinas	3	grupos	8	processos	6	sítios	5
entidades	73	localidades	12	produtos	5	substâncias	12
estados	13	medidas	12	propriedades	30	tecnologias	2
estratégias	23	métodos	2	recursos	30		

Fonte: Os autores

A segunda etapa de classificação envolveu a associação dos 417 termos aos três pilares da sustentabilidade: econômico, ambiental e social. Verificou-se que alguns termos eram intrinsecamente transversais, situando-se nas intersecções entre dois ou todos os três pilares. Termos relacionados primariamente a uma dimensão, mas secundariamente a outra, foram alocados na primeira; essa escolha não tem a intenção de desconsiderar a natureza interconectada dessas dimensões ou dos próprios termos.

Antes de criar a estrutura conceitual dinâmica apresentada no portal do LIS, essas classificações foram organizadas em planilhas compatíveis com ferramentas de modelagem gráfica, como yEd e XMind, para facilitar a visualização do domínio mapeado. Testes foram conduzidos para identificar a maneira mais eficaz de representar esses termos. No entanto, devido à dupla classificação de alguns itens, o yEd não foi adequado para produzir uma visualização clara. Em contraste, o XMind, apesar de suas opções limitadas de layout ajustável e dinâmico, facilitou a organização do mapa.

Em relação à visualização das imagens, a versão principal do Mapa Conceitual da Sustentabilidade está disponível em alta resolução no seguinte endereço: http://lis.ibict.br/wp-content/uploads/2024/09/MC-Sustentabilidade-XMind-Abr_24.pdf

Os autores optaram por apresentar no texto as figuras das três diferentes representações do mapa da sustentabilidade, a fim de fornecer um panorama inicial da estrutura de cada versão; no entanto, todas elas podem ser visualizadas em maior resolução ou em sua versão dinâmica por meio dos *links* fornecidos ao longo do artigo.

Além disso, a lista completa dos termos incluídos no mapa conceitual está disponível no Apêndice I.

Na estrutura conceitual apresentada nas Figuras 3 e 4, os termos estão dispostos verticalmente devido às limitações de espaço e agrupados pelas dimensões da sustentabilidade. Os rótulos coloridos indicam a categoria semântica do Agrovoc para cada termo. A legenda das cores está localizada à esquerda, em ordem alfabética. A cor branca corresponde aos termos “dimensão social” e “dimensão ambiental”. Estes foram incluídos durante o processo de mapeamento para manter a consistência com o termo validado “dimensão econômica”.

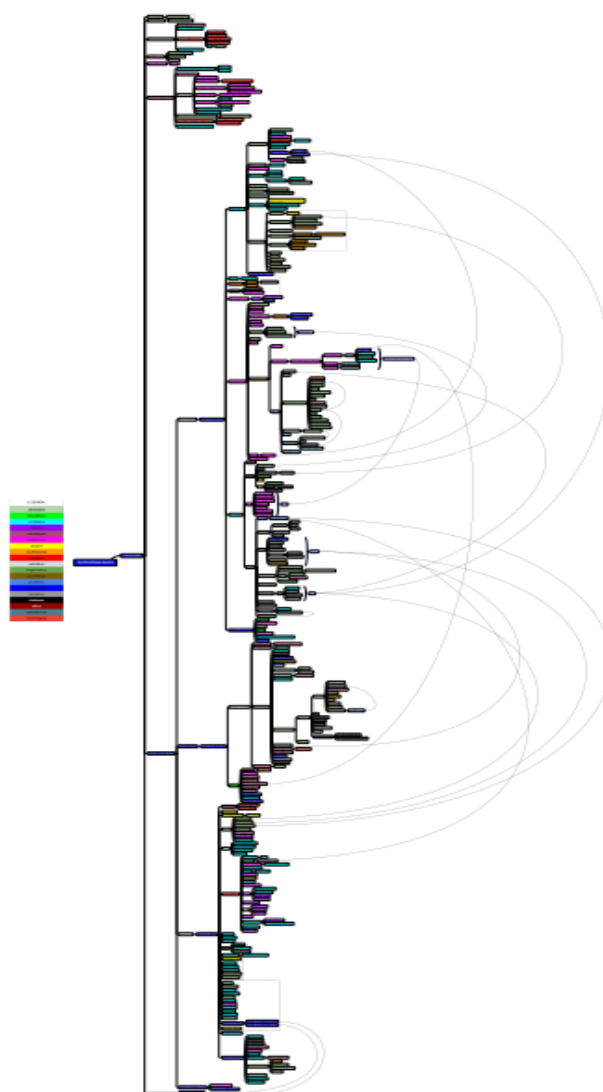


Figura 3. Representação gráfica dos termos classificados segundo as categorias semânticas amplas do Agrovoc e entre as dimensões da sustentabilidade no XMind.

Fonte: Os autores

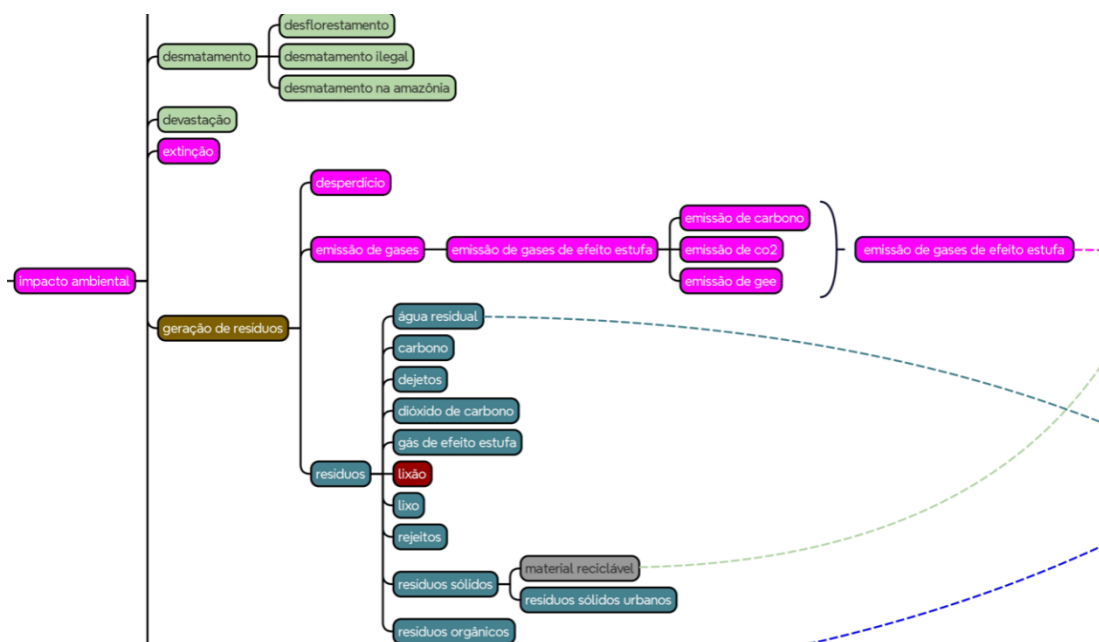


Figura 4. Visualização aproximada da representação gráfica dos termos classificados segundo as categorias semânticas amplas do Agrovoc e entre as dimensões da sustentabilidade no XMind. Fonte: Os autores

A estrutura apresenta vários níveis hierárquicos, embora as relações semânticas entre os termos permaneçam não especificadas nesta fase (a maioria se relaciona por hponímia, antonímia ou meronímia, ou como termos preferenciais e não preferenciais). O primeiro nível representa o domínio geral, indicado pelo rótulo “SUSTENTABILIDADE” em letras maiúsculas. O nível subsequente ramifica-se em “sustentabilidade”, para agrupar termos transversais, e “dimensões da sustentabilidade”. Esta última subdivide-se nas três dimensões, dispostas de cima para baixo: “dimensão ambiental”, “dimensão econômica” e “dimensão social”. Certos conjuntos de termos são contornados por linhas tracejadas para indicar agrupamentos. Da mesma forma, linhas tracejadas conectando termos distantes indicam uma relação semântica sem definir explicitamente seu tipo (por exemplo, entre “reciclagem” e “coleta seletiva”).

Representações adicionais foram elaboradas utilizando o yEd, com foco em um único critério de classificação por vez. Isso é ilustrado na Figura 5, na qual os termos estão organizados exclusivamente de acordo com as categorias semânticas do Agrovoc (também disponível como material suplementar no GitHub: https://github.com/MapeamentoC/Informacao_Para_Sustentabilidade).

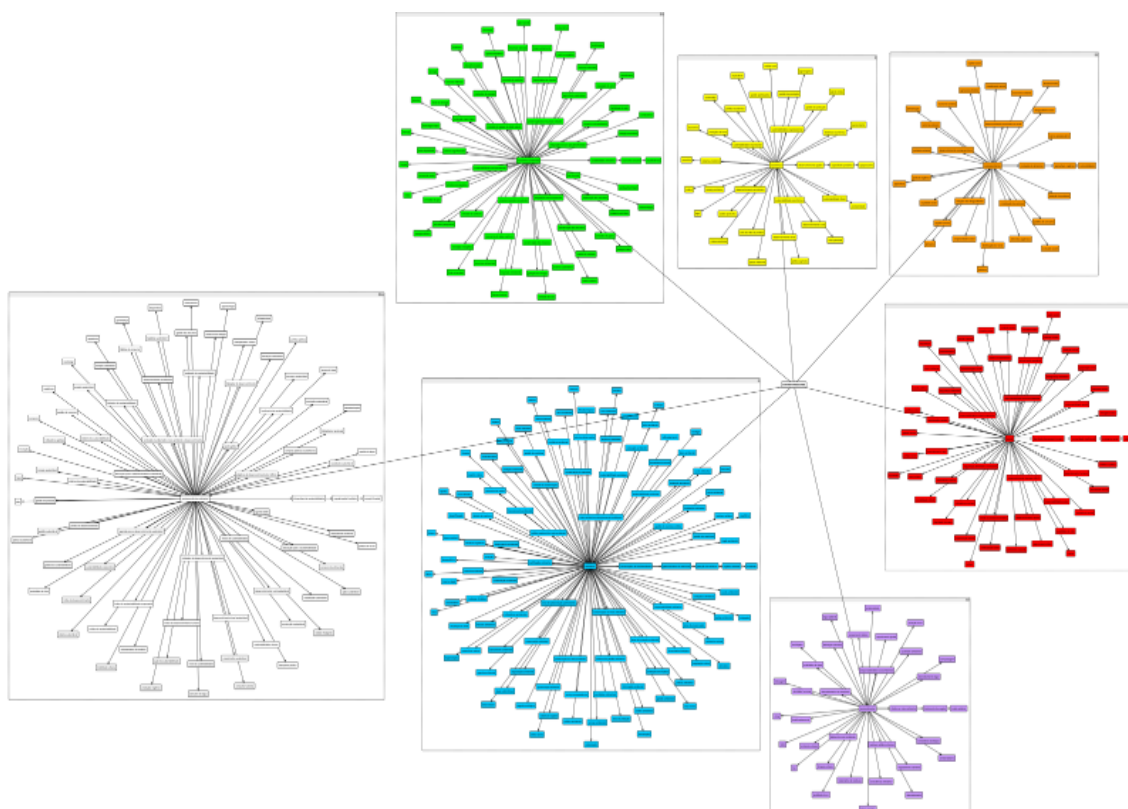


Figura 5. Representação gráfica dos termos classificados segundo as categorias semânticas amplas do Agrovoc no yEd. Fonte: Os autores

A estrutura dinâmica navegável, publicada no site do LIS, foi a versão final desenvolvida. Construída usando a biblioteca JavaScript Vis.js, ela permite que os usuários manipulem a estrutura conceitual ou grafo diretamente no navegador. Embora essa versão esteja disponível no portal (<https://lis.ibict.br/produtos/mapa-conceitual/>), ela requer desenvolvimento adicional, pois as relações semânticas entre os termos ainda não foram estabelecidas. Ainda que essa representação não inclua as classificações baseadas nas categorias do Agrovoc, ela fornece uma visualização orgânica de como os termos se integram entre as diferentes dimensões do domínio, conforme ilustrado na Figura 6.

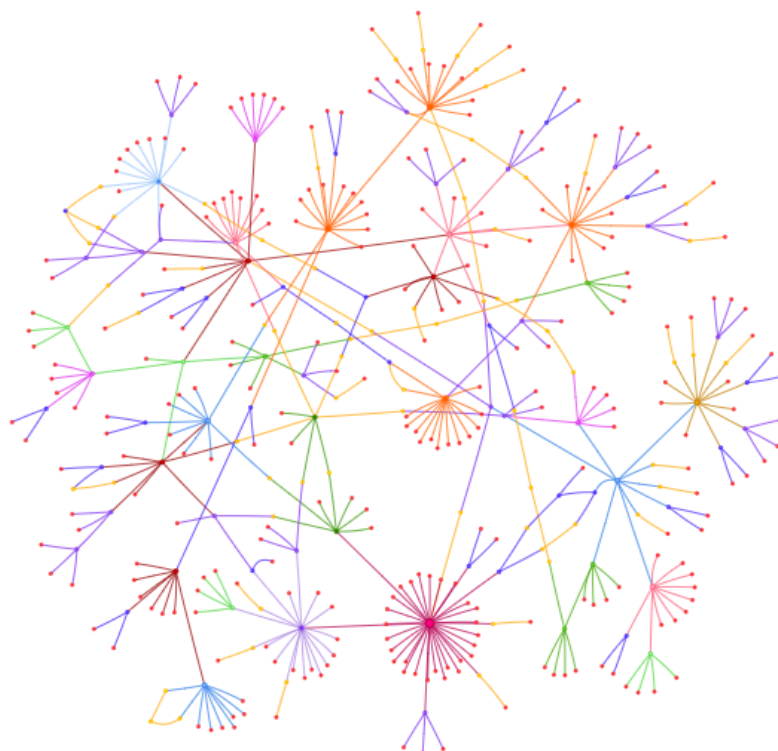


Figura 6. Mapa conceitual dinâmico publicado no Portal do LIS²³. Fonte: Os autores

4. Discussão

A construção do mapa conceitual para o domínio da sustentabilidade em português brasileiro permitiu não apenas sistematizar um conjunto significativo de termos extraídos de um *corpus*, como também observar a complexidade estrutural do domínio e as tensões inerentes à sua constituição discursiva. Sob a perspectiva da Análise de Domínio (Hjørland, 2017), os domínios do conhecimento não são segmentos neutros da realidade, mas sim configurações epistemológicas e discursivas historicamente situadas. Ao contrário de campos consolidados com fronteiras epistemológicas estáveis, a sustentabilidade é um domínio transversal, fortemente influenciado por agendas internacionais, políticas públicas nacionais, discursos corporativos e produção científica especializada. Nesse sentido, o mapa conceitual desenvolvido neste estudo não pretende representar a sustentabilidade como uma entidade ontológica estável, mas modelar a configuração terminológica que emerge do discurso brasileiro sobre sustentabilidade dentro de um escopo temporal, institucional e documental específico.

²³Disponível em: <https://lis.ibict.br/produtos/mapa-conceitual/>.

Ao adotar o *corpus* como evidência empírica da configuração discursiva do domínio, o estudo alinha-se a uma abordagem sociocognitiva da terminologia (Temmerman, 2000), na qual os termos são compreendidos como unidades linguístico-conceituais situadas em práticas sociais específicas. A heterogeneidade do *corpus*, integrando textos acadêmicos, jornalísticos e técnicos, garante que o mapa resultante reflita o discurso acadêmico consolidado e os discursos institucionais e midiáticos que moldam a construção social da área. Assim, a modelagem conceitual aqui apresentada é necessariamente situada e plural, evidenciando a sustentabilidade como um campo do saber policontextual e transversal.

4.1. Expressividade categorial e configuração do domínio

A classificação inicial dos 417 termos validados revelou uma predominância das categorias relacionadas a atividades, entidades e fenômenos. Estas apresentam 105 (25,2%), 73 (17,5%) e 50 (12%) termos, respectivamente; coletivamente, somam 228 termos (54,7%), ou seja, pouco mais da metade do total. Esse resultado sugere que o discurso da sustentabilidade, no *corpus* analisado, enfatiza práticas, mecanismos de implementação, gestão e transformação, ao mesmo tempo em que abrange um número significativo de entidades estáticas.

A predominância de termos associados a atividades, entidades e fenômenos indica que a sustentabilidade, no contexto brasileiro, é discursivamente construída com base em seus elementos tangíveis e nos atores envolvidos. Essa tendência reflete um *corpus* orientado para a descrição de agentes, eventos e realidade física, retratando o campo mais como um mapeamento de fatos e organizações do que como um conjunto estruturado de processos e estratégias. Em contraste, as categorias menos prevalentes revelam lacunas relativas ou menor consolidação terminológica em certos subcampos, como grupos e produtos; estas, em especial, podem sinalizar oportunidades para maior desenvolvimento conceitual ou expansão futura do *corpus*.

4.2. Interseccionalidade das dimensões da sustentabilidade

A segunda classificação baseada nas dimensões ambiental, social e econômica identificou um conjunto de termos categorizados como supradimensionais. Em vez de atribuir termos a intersecções entre duas dimensões, a versão final do mapa incorporou uma única categoria supradimensional. Essa categoria compreende termos cujo escopo conceitual abrange todos os três componentes do tripé da sustentabilidade. Nessa configuração, esses termos foram agrupados sob o próprio termo “sustentabilidade”, diferentemente daqueles atribuídos especificamente às dimensões ambiental, econômica ou social.

Esse resultado reforça o caráter sistêmico da sustentabilidade, um ponto frequentemente enfatizado em abordagens teóricas e refletido aqui na presença de termos

cujo escopo semântico não se restringe a uma única dimensão. Conceitos como “desenvolvimento sustentável”, “governança sustentável” e “políticas públicas”, por exemplo, operam como elementos estruturadores do domínio, articulando múltiplas esferas simultaneamente. Embora a divisão tripartite permaneça analiticamente útil, a existência de termos supradimensionais indica que certos conceitos centrais transcendem as fronteiras dimensionais e funcionam como nós integradores dentro do domínio.

Sob a perspectiva da Organização do Conhecimento, essa constatação evidencia a necessidade de estruturas de modelagem capazes de acomodar conceitos transversais sem fragmentá-los em intersecções artificiais. A decisão de adotar uma categoria supradimensional, em vez de um sistema de múltiplas classificações interseccionais, visou preservar a coerência conceitual enquanto se mantinha a clareza estrutural nas ferramentas de visualização empregadas. À luz da Terminologia Sociocognitiva (Temmerman, 2000), tais termos podem ser compreendidos como nós conceituais dinâmicos situados em práticas discursivas, exigindo estratégias de mapeamento flexíveis. A indicação de termos relacionados e agrupamentos, denotada por linhas tracejadas na visualização do XMind, visa complementar a organização hierárquica, sinalizando proximidade semântica e relações associativas dentro do domínio.

4.3. Contribuições para a Ciência da Informação e para a Terminologia

No âmbito da CI, o mapa conceitual resultante serve como um instrumento de organização do conhecimento fundamentado em evidências empíricas extraídas de um *corpus*. Ao contrário de vocabulários controlados construídos apenas por consenso de especialistas, a abordagem adotada integra a extração automática de candidatos a termo, a validação por especialistas, a classificação semântica apoiada por um padrão global (Agrovoc) e a representação gráfica. Essa estrutura sintetiza práticas da Terminologia, da Linguística de Corpus, do PLN e da Organização do Conhecimento, oferecendo um modelo metodológico replicável em outros domínios estratégicos.

Além disso, o mapa conceitual pode apoiar a arquitetura da informação em portais institucionais, fornecer base para o alinhamento com ontologias especializadas, melhorar a recuperação da informação e atuar como ferramenta mediadora entre especialistas e públicos leigos. Assim, o produto final transcende um papel meramente descritivo para se tornar parte integrante do ecossistema informacional de CT&I, fortalecendo a infraestrutura semântica do domínio.

4.4. Desafios metodológicos

Vários desafios emergiram ao longo do processo. Em primeiro lugar, a heterogeneidade do *corpus*, que, embora necessária para representar um discurso socialmente amplo, resultou em variação terminológica significativa, polissemia e diferentes níveis de

especialização, exigindo múltiplas rodadas de refinamento e validação.

Em segundo lugar, o fato do critério de frequência mostrar-se eficaz para a delimitação inicial do domínio, mas não suficiente por si só, uma vez que termos conceitualmente centrais podem ocorrer com menor frequência. A validação por especialistas foi, portanto, uma etapa decisiva para garantir a consistência semântica.

Finalmente, a representação gráfica revelou limitações técnicas nos *softwares* quanto à dupla classificação. A necessidade de ajustes manuais e da subsequente vinculação de termos aparentemente distantes indica que as ferramentas convencionais de mapeamento conceitual ainda impõem restrições quando aplicadas a domínios complexos e multidimensionais.

5. Conclusão

A construção do mapa conceitual da sustentabilidade confirma a eficácia de uma abordagem orientada por dados para a organização do conhecimento em domínios transversais e complexos. Ao expandir os princípios da Terminologia Comunicativa (Cabré, 1999) e da Análise de Domínio (Hjørland, 2017) por meio de um *corpus* de 50 milhões de palavras, este trabalho fornece uma alternativa empírica robusta aos métodos tradicionais baseados apenas no consenso de especialistas.

Os resultados confirmam que a sustentabilidade, no contexto brasileiro, é um campo primordialmente orientado para a ação e a governança, onde a interdependência dimensional exige modelos de representação flexíveis e não lineares. Em vez de um produto estático, o mapa e o *corpus* resultantes constituem uma infraestrutura semântica essencial para o ecossistema de CT&I, capaz de mediar a comunicação entre a produção científica e a formulação de políticas públicas.

Apesar de terem sido limitados ao escopo documental específico analisado, os fundamentos metodológicos estabelecidos aqui são replicáveis e escaláveis. Pesquisas futuras podem concentrar-se no refinamento das relações semânticas, na conversão deste mapa em uma ontologia formal ou na exploração de variações diacrônicas no discurso sobre sustentabilidade. Em última análise, este estudo reafirma o papel estratégico da CI na organização e representação da crescente complexidade do panorama informacional contemporâneo.

Uso de IA generativa

Os autores declaram o uso das ferramentas de inteligência artificial generativa ChatGPT – OpenAI e Gemini – Google, para suporte à tradução do português para o inglês. Todo o conteúdo foi revisado criticamente pelos autores, que assumem total responsabilidade pela versão final do texto.

Referências

- Anthony, L. (2020). *AntConc* (Versão 3.5.9) [Software de Computador]. Waseda University. <https://www.laurenceanthony.net/software>
- Cabré, M. T. (1999). *Terminology: Theory, methods and applications*. John Benjamins Publishing Company. <https://doi.org/10.1075/tlrp.1>
- CWTS. (2023). *VOSviewer* (Versão 1.6.20) [Software de Computador]. Leiden University. <http://www.vosviewer.com/>
- Dahlberg, I. (1978). A referent-oriented, analytical concept theory for INTERCONCEPT. *International Classification*, 5(3), 142–151. <https://doi.org/10.5771/0943-7444-1978-3-142>
- Di Felippo, A., & Almeida, G. M. de B. (2010). Uma metodologia para o desenvolvimento de Wordnets terminológicas em português do Brasil. *TradTerm*, 16, 365–395. <https://doi.org/10.11606/issn.2317-9511.tradterm.2010.46325>
- Dias-da-Silva, B. C. (2006). O estudo Lingüístico-Computacional da Linguagem. *Letras De Hoje*, 41(2), 103–138. <https://revistaseletronicas.pucrs.br/fale/article/view/597>
- Fenniak, M., & Thomas, M. (2025). *PyPDF2* (Versão 3.0.1) [Software de Computador]. <https://pypi.org/project/PyPDF2/>
- Goldberg, C., Burns, D., Diemol, Pujagani, & Titusfortner. (2025). *Selenium: Python language bindings for Selenium WebDriver* (Versão 4.11.0) [Software de Computador]. <https://pypi.org/project/selenium>
- Hjørland, B. (2008). What is knowledge organization (KO)? *Knowledge Organization*, 35(2–3), 86–101. <https://doi.org/10.5771/0943-7444-2008-2-3-86>
- Hjørland, B. (2017). Domain analysis. *Knowledge Organization*, 44(6), 436–464. <https://doi.org/10.5771/0943-7444-2017-6-436>
- Instituto Brasileiro de Informação em Ciência e Tecnologia. (2026). *Oasisbr – Sobre*. <https://oasisbr.ibict.br/vufind/about/home>
- Kobashi, N. Y. (2008). Fundamentos semânticos e pragmáticos da construção de instrumentos de representação de informação. *DataGramaZero*, 8(6). http://www.dgz.org.br/dez07/Art_01.htm
- Organização das Nações Unidas para a Alimentação e a Agricultura (FAO). (s.d.). Tesouro AGROVOC. Padrões para Gestão da Informação Agrícola (AIMS). <https://agrovoc.fao.org/browse/agrovoc/>
- Richardson, L. (2025). *Beautiful Soup 4* [Software de Computador]. <https://pypi.org/project/beautifulsoup4/>

-
- Sinclair, J. (2004). Corpus and text — Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (capítulo 1). Oxbow Books. <https://users.ox.ac.uk/~martinw/dlc/chapter1.htm>
- Smiraglia, R. P. (2014). *The elements of knowledge organization*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-09357-4>
- Temmerman, R. (2000). *Towards new ways of terminology description: The sociocognitive approach*. John Benjamins Publishing Company. <https://doi.org/10.7202/002265ar>
- Veiga, J. E. da. (2010). *Sustentabilidade: A legitimação de um novo valor*. Senac São Paulo.

APÊNDICE I

Lista dos 417 termos validados que integram o mapa conceitual de Informação para Sustentabilidade.

ação de sustentabilidade	cobertura vegetal	ecodesign	índice de sustentabilidade	produção orgânica
ação social	código florestal	ecoeficiência	índice de sustentabilidade empresarial	produção sustentável
ação sustentável	coleta seletiva	ecologia	industrialização	produtividade
acesso à água	combustível	economia	informação ambiental	produto orgânico
acesso à justiça	combustível fóssil	economia circular	infraestrutura	produto sustentável
agenda 2030	comissão mundial sobre meio ambiente e desenvolvimento	economia criativa	inovação	programa social
agenda ambiental	commodity	economia de baixo carbono	inovação social	proteção ambiental
agenda para o desenvolvimento sustentável	compostagem	economia solidária	inovação sustentável	proteção social
agricultura	compra sustentável	economia verde	insegurança alimentar	protocolo de quioto
agricultura familiar	compras públicas sustentáveis	ecossistema	integração lavoura pecuária	qualidade ambiental
agricultura orgânica	comunidade tradicional	ecossistema natural	integração lavoura pecuária floresta	qualidade da água
agricultura sustentável	condições ambientais	ecoturismo	interesse social	qualidade de vida
agro	condições climáticas	educação ambiental	investimento social	qualidade do ar
agroecologia	consciência ambiental	educação para a sustentabilidade	justiça	reaproveitamento
agroecossistema	consciência ecológica	educação para o desenvolvimento sustentável	justiça ambiental	reciclagem
agroindústria	conservação ambiental	efeito estufa	justiça social	recurso
agronegócio	conservação da biodiversidade	eficiência econômica	legislação ambiental	recursos ambientais
agropecuária	conservação da natureza	eficiência energética	licenciamento ambiental	recursos hídricos
agrotóxico	conservação do meio ambiente	emissão de carbono	limpeza urbana	recursos minerais
água	conservação dos recursos	emissão de co2	lixão	recursos naturais
água potável	construção sustentável	emissão de gases	lixo	recursos naturais renováveis
água residual	consumo	emissão de gases de efeito estufa	logística reversa	recursos pesqueiros
água subterrânea	consumo de água	emissão de gee	logística sustentável	recursos renováveis
alimentação	consumo de alimentos	empreendedor social	manejo	redução da pobreza

alimentação saudável	consumo de energia	energia	manejo florestal	redução das desigualdades
alimento	consumo humano	energia elétrica	manejo sustentável	reflorestamento
alimentos orgânicos	consumo sustentável	energia limpa	material reciclável	regeneração
alteração climática	contabilidade ambiental	energia renovável	matriz energética	regularização fundiária
ambientalismo	controle ambiental	equidade	meio ambiente	rejeitos
ambientalista	controle social	equidade social	meio ambiente ecologicamente equilibrado	relatório de sustentabilidade
ambiente	crédito de carbono	equilíbrio	mercado de carbono	reserva legal
ambiente natural	crédito rural	equilíbrio ambiental	mitigação	resíduos
ambiente urbano	crescimento sustentável	erradicação da pobreza	mobilidade sustentável	resíduos orgânicos
aquecimento global	crime ambiental	esgotamento sanitário	mobilidade urbana	resíduos sólidos
aquicultura	crise ambiental	espécie nativa	mortalidade infantil	resíduos sólidos urbanos
arborização	crise climática	exclusão social	movimento social	resiliência
área ambiental	crise econômica	exploração dos recursos	mudança climática	responsabilidade ambiental
área de preservação	critério ambiental	extensão rural	mudança do clima	responsabilidade social
área de preservação permanente	critérios de sustentabilidade	extinção	nível de sustentabilidade	responsabilidade social corporativa
área de proteção	cultivo	extrativismo	objetivos de desenvolvimento do milênio	responsabilidade social empresarial
área de proteção ambiental	cultura	extrema pobreza	objetivos de desenvolvimento sustentável	responsabilidade socioambiental
área degradada	cultura local	fauna	ocupação do solo	reúso
área natural	custo ambiental	fertilidade do solo	ordenamento do território	reutilização
área plantada	dano ambiental	flora	ordenamento territorial	risco ambiental
área protegida	defesa do meio ambiente	floresta	organização social	rotulagem
área verde	degradação ambiental	floresta tropical	órgão ambiental	saneamento ambiental
arranjo produtivo	dejetos	fonte de energia	pacto global	saneamento básico
assentamento	descarbonização	fonte renovável	padrão de consumo	saúde ambiental
assentamento humano	descarte	gás de efeito estufa	padrão de produção	saúde coletiva
assentamento rural	desempenho ambiental	gás natural	participação popular	saúde pública
assistência social	desenvolvimento agrário	geração de energia	participação social	segurança
aterro sanitário	desenvolvimento econômico	geração de resíduos	patrimônio cultural	segurança alimentar

atividade humana	desenvolvimento econômico e social	gerenciamento de resíduos	patrimônio natural	segurança alimentar e nutricional
ativo ambiental	desenvolvimento humano	gestão ambiental	pecuária	segurança jurídica
atmosfera	desenvolvimento local	gestão da água	pegada ecológica	selo
avaliação ambiental	desenvolvimento local sustentável	gestão da produção	pesca artesanal	serviço ambiental
avaliação da sustentabilidade	desenvolvimento rural	gestão da sustentabilidade	planejamento urbano	serviço ecossistêmico
avaliação de impacto	desenvolvimento rural sustentável	gestão de pessoas	plantio direto	sistema agroflorestal
baixo carbono	desenvolvimento social	gestão de recursos	pobreza	sistema de gestão ambiental
barômetro da sustentabilidade	desenvolvimento socioeconômico	gestão de resíduos	pobreza extrema	sistema produtivo
bem público	desenvolvimento sustentável	gestão de resíduos sólidos	política ambiental	sociedade de risco
biocombustível	desequilíbrio	gestão de riscos	política de desenvolvimento	sociedade sustentável
biodiesel	desertificação	gestão dos recursos	política econômica	sustentabilidade
biodiversidade	desflorestamento	gestão dos resíduos	política nacional do meio ambiente	sustentabilidade ambiental
bioeconomia	desigualdade social	gestão participativa	política pública	sustentabilidade corporativa
biogás	desmatamento	gestão social	política social	sustentabilidade ecológica
bioma	desmatamento ilegal	gestão sustentável	poluição	sustentabilidade econômica
biomassa	desmatamento na amazônia	governança	poluição do ar	sustentabilidade empresarial
biosfera	desperdício	governança ambiental	população rural	sustentabilidade fiscal
biotecnologia	devastação	hábitos de consumo	povo indígena	sustentabilidade organizacional
cadeia produtiva	dignidade da pessoa humana	impacto ambiental	prática agrícola	sustentabilidade social
capacidade de suporte	dignidade humana	impacto social	prática ambiental	sustentabilidade socioambiental
capacidade produtiva	dimensão econômica	impacto socioambiental	prática de sustentabilidade	sustentabilidade urbana
capital humano	dimensões da sustentabilidade	inclusão social	prática sustentável	tecnologia limpa
capital natural	dióxido de carbono	indicador de desenvolvimento	preservação ambiental	transformação social
capital social	direito ambiental	indicador de desenvolvimento sustentável	preservação do meio ambiente	transição energética
carbono	direito ao desenvolvimento	indicador de sustentabilidade	preservação dos recursos	transporte público



censo agropecuário	direito ao meio ambiente	indicadores ambientais	preservação dos recursos naturais	tratamento de esgoto
certificação ambiental	direitos fundamentais	indicadores globais	problema ambiental	tratamento de resíduos
ciclo de vida do produto	direitos sociais	indicadores nacionais	produção de alimentos	turismo sustentável
cidade inteligente	distribuição de renda	indicadores sociais	produção de bens	unidade de conservação
cidade sustentável	diversidade biológica	índice de desenvolvimento	produção de energia	uso do solo
clima	diversidade cultural	índice de desenvolvimento humano	produção mais limpa	vegetação nativa
				vulnerabilidade