DOI: 10.5433/1981-8920.2025v30n2p242

RECUPERANDO ESPECIALISTAS EM ENERGIAS RENOVÁVEIS POR MEIO DE TAXONOMIA FACETADA E TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL: UM EXPERIMENTO DE MINERAÇÃO DE DADOS ACADÊMICOS APLICADOS POR PESQUISADORES DAS UNIVERSIDADES ESTADUAIS DA BAHIA

RECOVERING RENEWABLE ENERGY SPECIALISTS THROUGH FACETED TAXONOMY AND NATURAL LANGUAGE PROCESSING TECHNIQUES: AN EXPERIMENT IN ACADEMIC DATA MINING APPLIED BY RESEARCHERS FROM BAHIA STATE UNIVERSITIES

Eduardo Manuel Freitas Jorge[a]
Gleidson Meireles Costa[b]
Victor Hugo Jesus Oliveira[c]
Alex Álisson Bandeira Santos[d]
Gesil Sampaio Amarante Segundo[e]

RESUMO

Objetivo: Este artigo propõe uma solução para a recuperação de informações textuais em um banco de dados acadêmico, utilizando técnicas de processamento de linguagem natural para identificar especialistas em energias renováveis. A solução emprega uma taxonomia facetada e uma plataforma de mapeamento de competências. **Metodologia**: A pesquisa segue uma abordagem experimental, estruturada nas seguintes etapas: 1) Identificação do problema e definição dos objetivos; 2) Busca e revisão sistemática de artigos sobre energias renováveis para a formação do vocabulário de controle; 3) Construção da taxonomia de energias

[a] Doutor em Difusão do Conhecimento pela Universidade Federal da Bahia (UFBA). Docente do Departamento de Ciências Exatas e da Terra da Universidade Estadual da Bahia (UNEB), Salvador, Brasil. E-mail: ejorge@uneb.br

[b] Graduando em Engenharia de Produção pela Universidade Federal do Recôncavo da Bahia (UFRB), Cruz das Almas, Brasil. E-mail: geu_costa@outlook.com

[c] Graduando em Engenharia de Sistemas pela Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brasil. E-mail: victorhugodejesusoliveira@gmail.com

[d] Doutor em Energia e Ambiente pela Universidade Federal da Bahia (UFBA). Docente da Universidade SENAI CIMATEC, Salvador, Brasil. E-mail: alexandre.ribeiro@fieb.org.br

[e] Doutor em Física pela Universidade de São Paulo (USP). Docente da Universidade Estadual de Santa Cruz (UESC). Ilhéus, Brasil. E-mail: gesil.amarante@gmail.com

renováveis usando o método 101; 4) Implementação do mecanismo de busca; 5) Análise dos dados dos pesquisadores especialistas. Os dados foram catalogados na plataforma simcc.uesc.br, incluindo informações como número de publicações, resumos do Lattes, índices de relevância e instituições dos pesquisadores. **Resultados:** O desenvolvimento de um motor de busca e de uma solução analítica permitiu correlacionar pesquisadores com a taxonomia de energias renováveis. A aplicação da taxonomia facetada como filtro resultou em 550 requisições na base de dados. **Conclusões**: A utilização da taxonomia facetada e o desenvolvimento do motor de busca proporcionaram uma recuperação de especialistas em energias renováveis, demonstrando a eficácia da abordagem proposta na combinação automática de termos para melhorar a busca e análise de informações acadêmicas.

Descritores: Buscas de informação. Processamento da linguagem natural. Mineração de dados.

1 INTRODUÇÃO

A Bahia destaca-se como polo nacional de energias renováveis, impulsionada por seus recursos naturais abundantes (Guimarães, 2023). Com 687 usinas e 23,5 GW de potência instalada, o estado lidera a geração eólica no país (Observatório da ETP, 2022). A proliferação de parques eólicos e solares consolida a Bahia como referência em energia limpa e sustentável (Sinergia Bahia, 2019).

Os investimentos em P&D na área eólica impulsionaram avanços tecnológicos e a produção em larga escala (Jannuzi, 2003). Na Bahia, o expressivo investimento em pesquisa estimula a colaboração entre academia, indústria e governo. As universidades estaduais, como a Universidade do Estado da Bahia (UNEB), a Universidade Estadual de Feira de Santana (UEFS), a Universidade Estadual de Santa Cruz (UESC) e a Universidade Estadual do Sudoeste da Bahia (UESB), desempenham um papel central nesse ecossistema de inovação.

A pesquisa em energias renováveis, concentrada em centros de pesquisa de ponta, tem mobilizado um contingente expressivo de pesquisadores universitários dedicados a estudos avançados em eficiência energética e tecnologias sustentáveis. Dada a relevância dessa área para a pesquisa científica, a identificação e caracterização desses profissionais é crucial para mapear as capacidades nacionais de pesquisa e direcionar

investimentos. A dispersão e a falta de integração dos dados acadêmicos em múltiplas bases dificultam a construção de um panorama preciso e atualizado sobre os pesquisadores do setor, comprometendo a recuperação eficiente e sistemática da informação (Jorge *et al.*, 2020).

A Recuperação de Informação (RI) oferece um conjunto de métodos e técnicas para lidar com a heterogeneidade e o volume de dados presentes em diversas bases de dados. Conforme Mooers (1950), a RI engloba os processos cognitivos envolvidos na descrição e busca de informações, independentemente dos sistemas ou ferramentas utilizados. No contexto acadêmico, ferramentas como o Somos (Somos, 2023), Stela Experta (StelaExperta, 2023) e o Sistema de Mapeamento de Competências Científicas da Bahia (SIMCC) (Santos *et al*., 2024) – desenvolvido pelos autores deste trabalho – são exemplos de sistemas de RI que auxiliam na classificação, organização e busca de informações. Essas ferramentas compartilham características comuns, como a possibilidade de realizar buscas utilizando termos-chave e operadores *booleanos* (E, OU).

O SIMCC, em particular, integra dados de diversas fontes, como a base do currículo lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a lista das revistas e a sua avaliação pelo sistema Qualis quadrienal 2017-2020 que é uma classificação feita pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) para estratificar os periódicos científicos em diferentes áreas do conhecimento e *Journal Citation Reports* (JCR) é uma base de dados da *Clarivate Analytics* que fornece métricas e indicadores para avaliar revistas científicas, principalmente na área de Ciências, proporcionando uma visão mais abrangente da produção científica.

Apesar das ferramentas supracitadas terem buscas eficientes existe uma limitação em todas elas quando é necessário realizar buscas com uma grande variação terminologia. Entende-se por grande variação terminológica quando diferentes termos ou palavras são usados para se referir ao mesmo conceito, ou quando o mesmo termo é usado para significar coisas diferentes, causando confusão e dificultando a busca de informações em um campo ou

contexto.

Este artigo propõe um motor de busca capaz de lidar com variações terminológicas, exemplificando seu uso na identificação de pesquisadores das Universidades Estaduais da Bahia na área de Energias Renováveis.

A solução utiliza uma taxonomia facetada para facilitar a recuperação controlada e organizada das informações, com foco em temas específicos. Na seção 3 do artigo, são apresentadas pesquisas correlatas e posteriormente detalha-se a metodologia baseada na Taxonomia 101 (Laubheimer, 2022) por oferecer um equilíbrio ideal entre praticidade e confiabilidade, solucionando questões de padronização ao mesmo tempo em que mantém os dados necessários para análises aprofundadas. Essa abordagem beneficia gestores e pesquisadores ao tornar a análise mais precisa e eficaz no processo de identificação de especialistas.

2 DESENVOLVIMENTO

Soluções aplicando conceitos taxonômicos são ferramentas úteis para a estruturação de informações e desempenham um papel fundamental no entendimento da organização de uma área de conhecimento, bem como na forma como ela se relaciona e interage com outras áreas (Aganette; Alvarenga; Souza, 2010). As taxonomias podem:

> [...] representar conceitos através de termos; melhorar comunicação entre especialistas e outros públicos; propor formas de controle da diversificação e oferecer um mapa do processo de conhecimento. É, portanto, um vocabulário controlado de uma determinada área do conhecimento e um instrumento que permite alocar, recuperar e comunicar informações dentro de um sistema [...] (Terra *et al.*, [20--?], p. 01).

No âmbito dos estudos da Ciência da Informação, esse tipo de estrutura tem sido considerado uma ferramenta de Recuperação da Informação (RI), com destaque também para seu papel na navegação (Ferreira, 2010). Em síntese, trata-se de um recurso valioso na organização de dados, facilitando a localização de conteúdos relevantes por meio de uma disposição lógica que orienta o usuário de forma mais eficiente.

Outra técnica aplicada a esse desavio é o uso de Processamentos de Linguagem Natural (PLN) no intuito de aprimorar a assertividade, eficiência e precisão das buscas. A técnica PLN é reconhecida como uma solução para desafios relacionados ao reconhecimento e à reprodução da linguagem humana. Evers e Finatto (2016) destacam alguns exemplos notáveis dessa abordagem, tais como a capacidade de reconhecer contexto, realizar análise sintática, semântica, léxica e morfológica, criar resumos, extrair informações, gerar traduções automáticas e interpretar significados. Além disso, a técnica de PLN pode "aprender" conceitos a partir de textos processados. Gonzalez e Lima (2003) complementam essa perspectiva, citando os níveis nos quais o PLN visa solucionar problemas computacionalmente.

Essa abordagem tem implicações significativas para a pesquisa e o desenvolvimento de sistemas inteligentes baseados em linguagem natural. Estes níveis são:

> fonético e fonológico: do relacionamento das palavras com os sons que produzem; • morfológico: da construção das palavras a partir unidades de significado primitivas e de como classificá-las em categorias morfológicas; • sintático: do relacionamento das palavras entre si, cada uma assumindo seu papel estrutural nas frases, e de como as frases podem ser partes de outras, constituindo sentenças; • semântico: do relacionamento das palavras com seus significados e de como eles são combinados para formar os significados das sentenças; • pragmático: do uso de frases e sentenças em diferentes contextos, afetando o significado (Gonzalez; Lima, 2003).

Neste contexto, buscando uma solução para a seguinte questão: "Como selecionar um conjunto de pesquisadores especialistas em Energias Renováveis?". Este artigo propõe uma solução de recuperação de informação textual, baseada em taxonomia e PLN, para realizar buscas aprofundadas na base de dados acadêmica dos pesquisadores das Universidades Estaduais da Bahia. O critério para determinar a profundidade das buscas e a eficiência é o número de combinações de termos da área pesquisada, realizadas pelo motor de busca, frente a uma busca realizada manualmente.

Na solução proposta, o processo de recuperação de informações passa a contar com uma estrutura categorial como interface de entrada, substituindo os tradicionais termos conjugados. O objetivo é facilitar a identificação de

pesquisadores, artigos e instituições acadêmicas relacionados ao tema Energias Renováveis. Diante da inexistência de uma classificação específica para esse domínio, optou-se pela construção de uma estrutura facetada, elaborada com base no método Taxonomy 101 (Laubheimer, 2022).

Para análise da efetividade da solução na seção 4.1 apresenta-se um experimento aplicado e a análise dos resultados da busca na base de dados do SIMCC.

3 APLICAÇÕES CORRELATADAS

No contexto da recuperação de dados, a taxonomia facetada tem se mostrado uma ferramenta valiosa em diversas produções. Maculan e Lima (2011), em "Taxonomia Facetada Navegacional: Agregando Valor às Informações Disponibilizadas em Bibliotecas Digitais de Teses e Dissertações", investigaram a implementação de uma taxonomia facetada navegacional (TAFNAVEGA), construída a partir da organização lógica da estrutura textual de trabalhos acadêmicos.

O mecanismo compreende um conjunto de categorias temáticas fundamentais que agrupam e interrelacionam os diversos conteúdos dos documentos. O algoritmo concebido para a extração dos conceitos que alimentam as categorias fundamentais temáticas da taxonomia facetada navegacional é empregado no processo de indexação do recurso informacional (Maculan; Lima, 2011).

Maculan e Lima (2011) utilizaram a TAFNAVEGA, desenvolvida a partir de 290 trabalhos (62 teses e 228 dissertações) defendidos entre 1998 e 2009 no Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais. Por meio de amostragem não-probabilística intencional, selecionaram um corpus de estudo composto por 41 documentos (12 teses e 29 dissertações) da linha de pesquisa “Organização e Uso da Informação”, disponíveis no banco de dados da biblioteca digital da instituição, onde cada resumo foi indexado pelo conjunto final CAFTE.

Ao final da análise, foram obtidos 168 diferentes termos de indexação,

que alimentaram as CAFTE, categorias utilizadas na etapa da implementação tecnológica.

Os resultados indicam que o mecanismo da Taxonomia Facetada Navegacional pode facilitar a exploração, busca e recuperação dos conteúdos dos documentos, proporcionando acesso a dados como teorias, métodos e instrumentos de coleta de dados. Este mecanismo auxiliou pesquisadores em suas atividades profissionais, permitindo maior visibilidade ao conteúdo disponível na biblioteca digital de teses e dissertações, sem sobrecarregar o usuário com informações. Isso demonstra o potencial das taxonomias para a recuperação de informação, validando sua utilidade para a disseminação de conhecimento. A presença de metadados, que o SIMCC possui, ligados às produções acadêmicas, proporciona ainda mais oportunidades para que interfaces como essa possam viabilizar a criação de novas conexões entre profissionais, empresas e instituições.

O estudo “Utilização de Taxonomias Facetadas para Aprimoramento da Recuperação de Informação em Acervos Culturais: Avaliação das Práticas do Instituto Brasileiro de Museus” (Coelho Júnior; Lemos, 2023) oferece uma análise sobre a implementação de taxonomias facetadas em coleções museológicas gerenciadas pelo Instituto Brasileiro de Museus. O foco da pesquisa reside na avaliação da conformidade do uso de vocabulários controlados nos metadados dessas coleções, com ênfase na harmonização com as diretrizes de catalogação. Para isso, é conduzido um processo de alinhamento dos metadados utilizados pelo IBRAM com as diretrizes do referido guia, seguido pelo desenvolvimento de um script em Python para o processamento dos dados das coleções.

Em seu estudo Coelho Júnior e Lemos (2023) destacaram a importância crítica do uso adequado de taxonomias e vocabulários controlados para garantir a qualidade e eficiência na recuperação de informações em acervos culturais. No entanto, eles também observaram que há margem para melhorias na aplicação padronizada dessas ferramentas entre os museus geridos pelas instituições relevantes. Isso sublinha a necessidade de aperfeiçoamento nesta área para facilitar uma melhor organização e acessibilidade das coleções

museológicas.

Ao comparar com o uso de taxonomias propostas no SIMCC, conclui-se que a aplicação da metodologia Taxonomia 101 oferece um equilíbrio ideal entre praticidade e confiabilidade, resolvendo assim o problema de padronização enquanto mantém dentro da aplicação os dados necessários para análises aprofundadas tanto para os gestores, quanto para pesquisadores que estão desenvolvendo suas pesquisas.

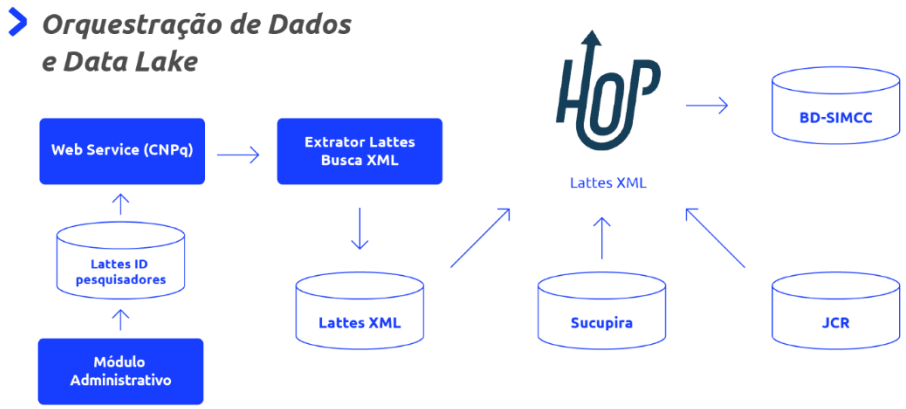
4 O SIMCC – SISTEMA DE MAPEAMENTO DE COMPETÊNCIAS CIENTÍFICAS DA BAHIA

É uma plataforma de mapeamento de competências em rede para apoiar os gestores e a comunidade acadêmica na análise de informações da produção dos pesquisadores da Bahia e foi concebido como uma solução de RI desenvolvido para facilitar a busca por termos de pesquisa dos pesquisadores.

O SIMCC envolve a área da pesquisa em Ciência de Dados, com a especificação de uma arquitetura e o desenvolvimento de processos e soluções para: a) Integração e orquestração de dados volumosos em estrutura semiestruturada; b) Buscas usando linguagem de processamento natural textual; c) Soluções analíticas para apoiar os gestores da pós-graduação e comunidade acadêmica na tomada de decisão.

A arquitetura do SIMCC (Figura 1) foi projetada para automatizar a ingestão incremental de currículos Lattes no formato XML, disponibilizados pelo CNPq. A figura 1 representa um esquema com o *Data orchestration* que é a automação e coordenação do fluxo de dados entre sistemas para garantir que estejam disponíveis e preparados para análise. Na figura 1 também está representado o *Data lake* que é um repositório que armazena grandes volumes de dados brutos, estruturados ou não, para exploração futura.

Figura 1 – Arquitetura Geral SIMCC



Fonte: Criada pelos autores (2023)

O processo consiste no *Extract, Transform, Load*, implementado com Apache Hop, que extrai dados de uma máquina autorizada Web Service (CNpq) e realiza a transformação desses dados e os carrega em uma base relacional integrada. As bases utilizadas incluem informações do Qualis que está na base Sucupira da Capes (contendo todas os periódico avaliados para a quadrienal 2017-2020), JCR e outros agregadores.

Para a indexação e busca, o sistema utiliza o PostgreSQL, com recursos de PNL e Python (NLTK). A interface web, desenvolvida em React JS, oferece uma barra de pesquisa que sugere termos com base em um dicionário construído a partir das produções acadêmicas indexadas.

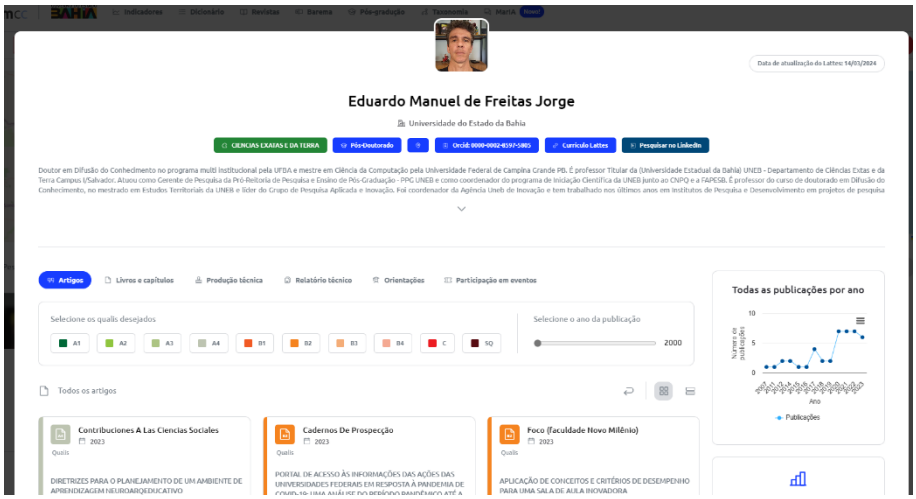
Figura 2 – Busca do SIMCC

Fonte: Criada pelos autores (2023)

A pesquisa permite a combinação de múltiplos termos, incluindo palavras-chave, resumos e áreas de atuação dos pesquisadores (Figura 2). Ao selecionar um pesquisador, é possível visualizar detalhadamente suas publicações e métricas, bem como um resumo de sua produção científica (Figura 3).

Além dessa visão individualizada, a plataforma oferece ferramentas para análises mais amplas. Painéis interativos em Power BI permitem explorar dados estatísticos sobre instituições e programas de pós-graduação, alimentados por dados exportados do SIMCC. Esses painéis, juntamente com os perfis individuais, proporcionam uma visão abrangente da produção científica.

Figura 3 – Perfil detalhado do pesquisador



Fonte: Criada pelos autores (2023)

Embora a plataforma ofereça uma busca para termos com uma ou duas palavras, a identificação de pesquisadores em áreas com terminologias complexas que demandam um vocabulário extenso e com conexões recorrentes ainda representa um desafio. As próximas seções apresentam uma proposta para superar essa limitação.

5 PROCEDIMENTOS METODOLÓGICOS

Para a realização do projeto de identificar um conjunto de pesquisadores especialistas em “Energias Renováveis” utilizou-se a metodologia baseada na pesquisa experimental de Gil (2002). De forma iterativa e incremental desenhou-se o percurso metodológico apresentado na figura 4 que será explicado na sequência dessa seção.

A presente pesquisa iniciou com a identificação de uma lacuna terminológica na área de Energias Renováveis, não foi encontrada uma taxonomia pré-existente para essa área. Para supri-la, realizamos uma revisão sistemática da literatura, buscando artigos científicos publicados nos últimos 10 anos em bases de dados renomadas como IEEE Xplore, ScienceDirect, Scopus, *Google Scholar* e *Web of Science*.

Figura 4 – Etapas da pesquisa

Fonte: Criada pelos autores (2023)

A seleção dos artigos priorizou publicações em revistas com Qualis superior a A e fator de impacto superior a 2.0, O protocolo de seleção dos artigos seguiu a metodologia PRISMA (*Preferred ReportingItems for Systematic Reviews and Meta-Analyses*), que assegura transparência e reprodutibilidade na revisão sistemática. Essa abordagem permitiu construir uma taxonomia precisa e atualizada para a área, fundamentada em pesquisas de alto impacto.

Ressalta-se que, no momento da elaboração deste artigo, o sistema Qualis está passando por uma reformulação significativa, com a adoção de novos critérios de avaliação centrados na qualidade individual dos artigos científicos. Apesar dessas mudanças, a plataforma SIMCC não sofre grandes perdas, pois sua estrutura se fundamenta em bases consolidadas como o JCR, que continua sendo um dos principais referenciais de qualidade na avaliação da produção científica. Ademais, o JCR permanece como um insumo relevante na própria reformulação do Qualis, o que garante a continuidade metodológica da plataforma.

Com um primeiro vocabulário de controle definido, a etapa 3 empregou a metodologia Taxonomia 101 para construir uma taxonomia específica para o domínio das Energias Renováveis. A partir dos artigos selecionados na revisão sistemática, foram extraídos os termos-chave, conforme demonstrado na

coluna "termos" da Tabela 1. A fim de garantir a abrangência das buscas, os termos foram traduzidos para o inglês. Uma equipe realizou uma análise crítica dos termos, excluindo aqueles excessivamente específicos ou alheios ao escopo da pesquisa, resultando em uma taxonomia refinada e adequada para as buscas em português e inglês.

Tabela 1 – Amostra dos artigos selecionados

Nº	*Revista / Titulo*	*Termos (PT-BR)*	*Ano*	*Qualis*	*JCR*
1	Cadernos de Saúde Pública Chemical Characteristics of Pollutants from Biomass Burning and Combustion of Fossil	Poluentes; Matéria Particulada; Doenças do Trato Respiratório	2023	A1	2.8
2	International Journal of Hydrogen Energy Energy Systems with Hydrogen Storage: Sizing, Optimization, and Energy	Sistema de Energia Renovável Híbrido; Otimização; Dimensionamento;	2022	A1	7.2
3	Global Food Security The Battle for Biomass: A Systematic Review of Food-Feed-Fuel Competition	Bioenergia; Cascata de Biomassa; Uso Sustentável de Recursos; Uso Circular de Biomassa.	2022	A2	8.9
4	Ecological Informatics Energy Consumption and Environmental Degradation Nexus: A Systematic Review and Meta-Analysis of Fossil Fuel and Renewable	Energia de Combustíveis Fósseis; Energia Renovável; Consumo de Energia; Degradação Ambiental;	2022	A1	5.1
5	Renewable and Sustainable Energy Reviews Operational Planning of Renewable and Sustainable Energy Systems Considering	Energia Renovável; Quantificação de Incerteza; Demanda	2021	A1	15.9
6	Energy Policy The Application of Renewable Energy to Social Housing: A Systematic Review	Poluentes; Matéria Particulada; Doenças do Trato Respiratório	2018	A1	9.0

Fonte: Criada pelos autores (2023)

Além disso, foram coletadas manualmente palavras-chave em sites especializados. Portanto de maneira sistemática, os termos muito genéricos foram removidos do corpo da taxonomia, sugestões de palavras chaves identificadas pelo corpo de especialistas também foram acrescentadas. Exemplo de termos genéricos que foram retirados: “tecnologia da informação”, “doenças do trato respiratório”, “degradação ambiental”, etc. Por fim, foi feita uma última limpeza na taxonomia deixando somente os termos que se enquadram em n-gramas menores ou iguais a bigramas, devido às limitações do mecanismo de busca utilizado, o qual será detalhado na próxima seção.

Na quarta etapa, adaptou-se o motor de busca do SIMCC para permitir o uso dos termos combinados da taxonomia facetada, elaborada na etapa

anterior, visto que o motor atual tinha como entrada termos e não uma taxonomia. Outro novo recurso implementado foi a geração de um cubo multidimensional de dados (que será detalhado na próxima seção), correlacionando os termos combinados com as produções cientificas dos pesquisadores que contém no seu título os termos pesquisados.

A última interação, “etapa 5”, de análise das produções identificadas após a busca utilizando a taxonomia, objetivou verificar o nível de assertividade das produções através da ferramenta Power BI que consumiu o cubo multidimensional gerado e efetuar adequação dos termos da taxonomia e formas de possíveis combinações. Observa-se que esse processo de evolução da taxonomia é contínuo e como apresentado na figura 4, a cada ciclo do processo, a taxonomia pode evoluir gerando novas versões. Para este projeto, após três ciclos de análise dos resultados das buscas, findou-se a elaboração taxonomia de “Energias Renováveis”. Na próxima seção detalham-se os resultados obtidos nas etapas 4 e 5.

6 ANÁLISE DE RESULTADOS

Essa seção está estruturada, primeiramente, na apresentação do funcionamento do motor de busca do SIMCC, aprimorado para o uso da taxonomia facetada elaborada no projeto, posteriormente apresenta-se a última versão da taxonomia de “Energias Renováveis” que foi utilizada no experimento aplicado para identificar pesquisadores das universidades do Estado da Bahia com competências neste domínio.

O motor de busca do SIMCC, desenvolvido na linguagem Python, como detalhando na seção 2.2, realizava um conjunto de requisições ao SGBD PostgreSQL, que orginalmente permitia que o usuário na barra de busca fornecesse um ou mais termos combinados com o operador “and”. Os termos eram usados em um processo aplicado nos títulos dos artigos, resumo do lattes, títulos de livros e artigos, área de atuação e patente, seguindo os seguintes passos: a) Padronização dos termos para palavravas em minúsculas; b) Retirada de caracteres especiais “- \. ' “; c) Retirada de stop words; d)

Transformação do texto em vetor de palavras;

Caso o termo de busca seja um bigrama, por exemplo “Energia Solar”, fazer uma busca utilizando um parâmetro de ranqueamento que garante a proximidade dos termos nos títulos ou resumos. Para a montagem das consultas, utilizou-se a seguinte expressão de filtro na condição “WHERE” do PostgreSQL:

Figura 5 – Parâmetro de ranqueamento simples

Fonte: Criada pelos autores (2023)

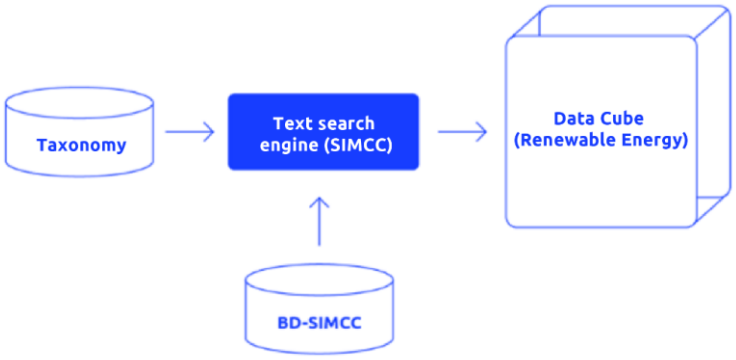
Nesta expressão, o primeiro parâmetro representa o título ou resumo, enquanto o segundo parâmetro representa o termo de interesse, como “Biomassa”, por exemplo. Para o caso de bigramas, a seguinte expressão é utilizada:

Figura 6 – Parâmetro de ranqueamento expandido

Fonte: Criada pelos autores (2023)

Nesta expressão, o primeiro parâmetro representa novamente o título ou resumo. O segundo e o terceiro parâmetros representam os termos de interesse, como “Energia” e “Solar”, respectivamente. O último parâmetro representa o valor do rank.

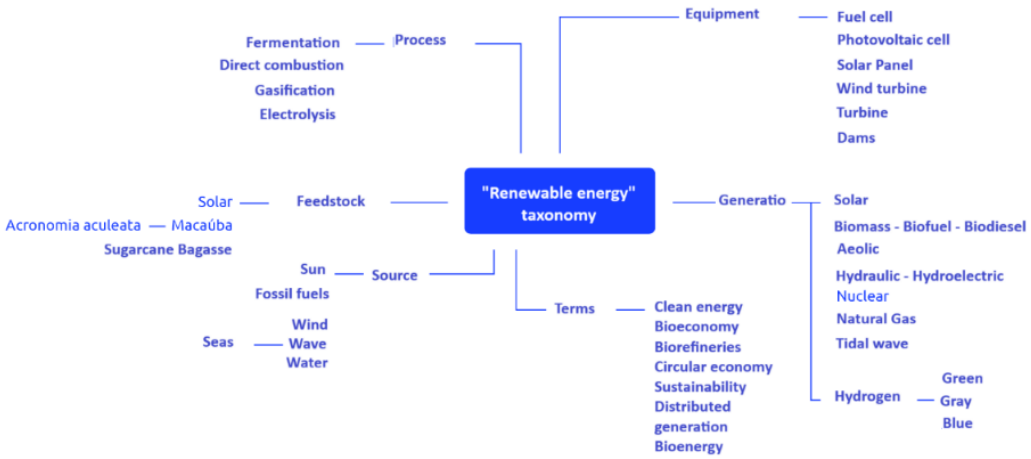
Figura 7 – Arquitetura do motor de busca



Fonte: Criada pelos autores (2023)

Como mencionado, se a busca tiver a conjugação de vários termos a expressão “and” é concatenada aos filtros montados para cada termo. No novo motor de busca, representado na figura 7, a taxonomia facetada é dada como entrada pelo usuário que não precisa definir os termos da busca. A consulta é processada através da combinação dos termos generalistas da taxonomia de “Energias Renováveis” representados pela hierarquia “termos” com as outras hierarquias. São exemplos de termos generalistas neste domínio: “Energia Limpa”, “Energia”, “Geração Distribuída”, etc. Além da hierarquia termos, na taxonomia existem os seguintes conjuntos para o processo de combinação: “processos”, “matérias-primas”, “equipamentos”, “geração” e “fontes” (figura 8)

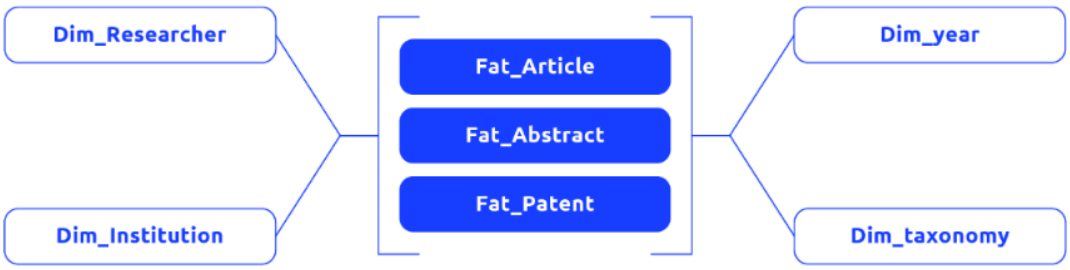
Figura 8 – Hierarquia da Taxonomia



Fonte: Criada pelos autores (2023)

Esse processo de combinação gerou 550 requisições na base de dados do SIMCC realizadas de forma automática. Ao final do processo um cubo multidimensional, ver figura 9, é gerado como resultado deste processamento. Para permitir a análise dos dados da pertinência dos termos o cubo de dados possui o seguinte recorte de dados de negócio: (i) dimensões pesquisador, instituição, ano e taxonomia; (ii) fatos artigo, resumo do pesquisador e patente.

Figura 9 – Cubo multidimensional de dados



Fonte: Criada pelos autores (2023)

7 EXPERIMENTO – PESQUISADORES DAS UNIVERSIDADES ESTADUAIS DA BAHIA COM COMPETÊNCIAS EM ENERGIAS RENOVÁVEIS

Para análise da efetividade da solução proposta foi realizado um experimento na base de dados do SIMCC. Para compreensão do experimento detalha-se as características da base de dados do SIMCC. Para os testes da

solução, a carga das informações foi realizada no mês de setembro de 2023 com o Lattes dos pesquisadores das universidades estaduais da Bahia, somando-se cerca de 4810 docentes distribuídos entre a UNEB, UESB, UEFS e UESC. Além dos currículos lattes, os artigos carregados dos pesquisadores foram categorizados através das bases Qualis (Capes, 2019) e o JCR, identificando o grau de impacto das produções. Esses 4810 pesquisadores possuem 42.196 artigos e 255 patentes com ano maior ou igual que 2013.

Após a execução do motor de busca, utilizando a taxonomia facetada em português e inglês foi gerado o cubo multidimensional de dados com os seguintes resultados discutidos na sequência. Dentre os termos com maior ocorrência nos títulos dos artigos, temos: “bagaço da cana”, “solar”, “biomassa”, “biodiesel” e “eólica”. Assim, destaca-se que esses termos não foram localizados de forma isolada, mais sim conectados com os termos adjacentes da taxonomia como explicado anteriormente.

Tabela 2 – Pesquisador (P) x Termo

Pesquisador Nº	*Biodiesel*	*Biomassa*	*Eólica*	*Solar*	*Total*
1				3	3
2	3				3
3	3				3
4			1	1	2
5			2		2
6		2			2

Fonte: Criada pelos autores (2023)

Outra visão analítica dos dados possível é fazer um cruzamento dos pesquisadores com os termos da taxonomia como pode ser observado na tabela 2. Por exemplo, o pesquisador 1 possui 3 (três) artigos conectados com o termo “Solar”.

Além da visão, em relação ao pesquisador, é possível ampliar o nível da consulta e obter a mesma relação só que por instituição de ensino. Na tabela 3, a UESC é a universidade com maior número de ocorrências de artigos em relação ao termo “Solar” com 8 (oito) artigos, já a UNEB teve o maior resultado em relação ao termo “Eólica” com 4 (quatro) artigos.

Tabela 3 – Instituição x Termo

Instituição	*Biodiesel*	*Biomassa*	*Eólica*	*Hidrogênio*	*Solar*	*Total*
UESC	5	3	2	1	8	19
UNEB	1	2	4	1		8
UEFS		2			1	3
UESB	1				1	2
Total	6	7	6	1	10	30

Fonte: Criada pelos autores (2023)

Uma análise mais detalhada também é possível listando os artigos por um termo. Na tabela 5 são listados os artigos que possuem o termo “Eólica” detalhando o ano, título, nome da revista, fator de impacto, Qualis, pesquisador e instituição. Na segunda linha desta tabela, o artigo de título “Aspectos Positivos e Negativos da Produção de Energia Eólica no Distrito do Tamboril, Morro do Chapéu Bahia” é um artigo do ano de 2020 da universidade UNEB com um estrato Qualis A2.

Nesse experimento o motor de busca utilizou a taxonomia facetada também no texto do resumo do currículo lattes dos pesquisadores. Nesta busca os cinco termos mais localizados mudaram, sendo a o termo “Nuclear”, “Biomassa”, “Eólica”, “Solar” e “Fermentação”. Por exemplo, se a busca é por pesquisadores que têm no seu resumo o termo “Hidrogênio” pode se observar que existem pesquisadores que atendem a esse critério.

Tabela 4 – Artigos que contém o termo vento

Instituição	*Revista / Titulo*	*Ano*	*JCR*	*Qualis*
UNEB	Revista Geotemas Aspectos Positivos E Negativos Da Produção De Energia Eólica No Distrito Do Tamboril, Morro Do Chapéu? Bahia	2020		A2
UESC	Journal Of Control, Automation And Electrical Systems Methodology For Assessing The Risk Of Unintentional Islanding Of Distributed Wind Generators Using Passive Schemes	2020	1.5	A4
UNEB	Revista Eletrônica De Estratégia & Negócios Análise Estratégica Do Setor De Energia Eólica No Brasil	2019		B1
UNEB	Nexos Econômicos Aspectos Econômicos E Jurídicos Que Cercam A Relação De Camponeses Com Empresas Exploradoras De Energia Eólica No Município De Brotas De Macaúbas? Bahia	2019		B2
UNEB	Rgsa (anpad) Projetos De Mdl De Energia Eólica No Nordeste Do Brasil: Perfil E Cobenefícios Declarados	2018		A3
UESC	Revista Científica Semana Acadêmica Estudo Comparativo Do Aproveitamento De Energia Eólica Na Região Nordeste Do Brasil X Alemanha	2018		B3

Fonte: Criada pelos autores (2023)

Em continuidade, a tabela 6 e 7 listas a relação de instituições versus as

ocorrências de termos nos resumos do currículo lattes dos pesquisadores. Assim, fazendo-se uma nova análise para o termo “Hidrogênio”, observa-se que só existe relação com a instituição UNEB, ao contrário do termo “Nuclear” que existe relação com as quatro instituições do estado.

Tabela 5 – Instituições x Ocorrência de termos em resumos

Instituição	*Bagaço de cana*	*Biodiesel*	*Biomassa*	*Eólica*	*Fermentação*	*Gás Natural*
UEFS	1			1	1	
UESB			2	1	1	1
UESC		1				
UNEB			1	1		
Total	1	1	3	3	2	1
Instituição	**Hidrogênio**	**Nuclear**	**Ondas; Marés**	**Solar**	**Vento**	**Total**
UEFS		2		2		5
UESB		3		1	1	8
UESC		8				9
UNEB	2	1	1		1	6
Total	2	14	1	3	2	28

Fonte: Criada pelos autores (2023)

De forma sintética o experimento identificou 42 docentes o que representa 0,93% em relação ao total de pesquisadores e 68 artigos o que representa 0,16% em relação ao total de artigos. Em relação as patentes, a busca não localizaram documentos no processo. Estes percentuais demonstram que apesar dos percentuais serem baixo em relação a base completa quantitativamente são valores importantes para conectar pesquisadores e ações com empresas e órgãos governamentais de fomento à pesquisa. Foi observado que determinados termos apresentaram uma frequência considerável, indicando uma consistência temática. A esfera da energia eólica e solar tem experimentado uma ampla disseminação no cenário científico, conforme delineado na introdução deste artigo, refletindo-se de maneira expressiva na formulação da taxonomia.

8 CONSIDERAÇÕES FINAIS

Esta pesquisa apresentou uma solução para RI no contexto das

"Energias Renováveis", temática relevante no contexto da Bahia devido as condições naturais e aos investimentos recentes nesta área. A busca por pesquisadores com competências nesse domínio é uma demanda necessária, visando conectar empresas, governo e universidades, e assim estimular e promover a tríplice hélice da inovação. A identificação de pesquisadores neste artigo aplicou as técnicas de PLN e de taxonomia facetada para buscas nas fontes de dados acadêmicas do SIMCC que já tinha resolvido o problema de carga e integração dessas bases, mas não possuía um motor de busca que usasse como filtro profundo uma taxonomia de pesquisa multifacetada.

O primeiro desafio foi a elaboração da taxonomia de "Energias Renováveis", pela não identificação da mesma. A estratégia de elaboração foi seguir o método da Taxonomia 101 que preconizava a busca por um vocabulário de controle que foi parcialmente resolvido, utilizando termos encontrados em artigos que tratavam de Revisões Sistemáticas. Para o refinamento dos termos ainda foi necessário a interação de especialistas para retirada de termos genéricos, inclusão de termos de sites especializados e do portal SIMCC. Por fim, foi necessário a retirada de termos superiores a bigrama devido a limitação atual do motor de busca.

Os resultados obtidos ao entorno da pesquisa foram o percurso metodológico proposto e a taxonomia construída. Já o principal resultado alcançado, foi o desenvolvimento de um novo motor de busca do SIMCC, apoiado no SGBD PostgreSQL devido aos recursos de PLN textual, que automatizou o processo de combinação de termos.

Esta automatização, se feita manualmente, o usuário do SIMCC teria que realizar 550 requisições na base de dados. Desta forma, o desenvolvimento desse novo motor de busca não apenas simplifica e acelera o processo, mas também otimiza significativamente a eficiência do sistema. Somado a isso, o resultado da busca gerou um cubo multidimensional de dados que possibilitou os gestores a realizarem uma série de análises através de uma ferramenta OLAP, no caso o Power BI. Na seção de detalhamento do experimento aplicado, demonstrou-se as possibilidades de análise através de tabelas e gráficos apresentados no recorte de dados pelas dimensões ano,

instituição de ensino, pesquisadores, termos sobre os fatos de artigos, resumos do lattes e patentes. No contexto geral, a presente pesquisa identificou 42 docentes e 68 artigos no experimento realizado dentro da tangência de “Energia Renováveis”.

Prosseguindo, realizou-se uma análise das instituições, exemplificando com a UNEB, que se destaca na produção de pesquisa sobre "Hidrogênio", e a UESC, que focaliza em "Nuclear", com base na frequência de ocorrência desses termos nos resumos dos docentes. Dessa forma, este estudo alcançou seus objetivos ao identificar termos relevantes e criar uma taxonomia específica para "Energias Renováveis", ao mesmo tempo em que aplicou um filtro multifacetado para apresentar uma análise gráfica abrangente sobre pesquisadores, artigos, resumos e instituições.

9 PERSPECTIVA FUTURA

Em primeiro plano, almeja-se a expansão da base de dados do SIMCC para incluir informações das demais universidades públicas do estado da Bahia, como a Universidade Federal do Sul da Bahia (UFSB), Universidade Federal do Oeste da Bahia (UFOB), Instituto Federal da Bahia (IFBA) e outras instituições relevantes. Essa iniciativa visa potencializar a plataforma, não se limitando apenas ao âmbito das universidades estaduais, mas abrangendo todo o estado. Essa ampliação contribuirá para uma visão mais abrangente e integrada do cenário acadêmico baiano, fortalecendo a utilidade e abrangência do SIMCC como uma ferramenta valiosa para a comunidade acadêmica como um todo.

Além disso, elaborar uma taxonomia representa um desafio significativo. No entanto, aprimorar uma taxonomia, especialmente no âmbito da literatura, apresenta-se como um desafio ainda mais complexo, demandando a criação de novas categorias por meio de um processo manual. Desta forma, a ideia de automatização da concepção de uma nova taxonomia alinhada ao grupo de palavras que circundam um tema como “Energias renováveis” e os termos encontrados nos títulos de artigos e resumos dos pesquisadores se torna uma

etapa futura do projeto.

Diante da crescente demanda por inteligências artificiais e o avanço do aprendizado de máquina, o projeto tem a perspectiva de ser ampliado com a introdução da estrutura de Mapeamento de Pesquisadores utilizando Inteligência Artificial (MarIA). A implementação irá utilizar modelos como o Vicuna-13B (LMSYS, 2023), um Large Language Model (LLM) de código aberto, treinado através do ajuste fino do LLaMA (META AI, 2023), um modelo de linguagem grande fundamental de última geração projetado para ajudar os pesquisadores a avançar em seu trabalho neste subcampo da IA, como detalhado em (META AI, 2023). O LLaMA, por sua vez, foi treinado em um vasto conjunto de dados e demonstra capacidades avançadas de geração de texto e resolução de problemas.

A utilização do Vicuna-13B, ajustado a partir do LLaMA, no MarIA nos permite explorar as potencialidades desse modelo, ao mesmo tempo em que nos confrontamos com os desafios de mitigar os riscos associados a essas tecnologias. Através do MarIA, buscamos desenvolver estratégias para minimizar esses problemas e maximizar os benefícios do uso de grandes modelos de linguagem em pesquisas científicas.

Assim, a implementação do MarIA visa promover uma interação maior entre os usuários e a máquina, estreitando laços e simplificando a comunicação. Além disso, essa abordagem aprimorada resultará em análises mais refinadas e em resultados mais precisos, contribuindo para o avanço da pesquisa.

AGRADECIMENTOS

Agradecemos pelo apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) que financiou a pesquisa realizada por meio de duas bolsas de Iniciação Científica e uma Bolsa de Produtividade Desen. Tec. e Extensão Inovadora do CNPq – Nível 2

REFERÊNCIAS

AGANETTE, Elisangela; ALVARENGA, Lídia; SOUZA, Renato Rocha. Elementos constitutivos do conceito de taxonomia. **Informação & Sociedade: Est.,** João Pessoa, v. 20, n. 3, p. 77-93, set./dez. 2010. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/3994. Acesso em: : 23 ago. 2024.

CAPES. **Metodologia do Qualis – Referência Quadrienal 2017-2020**. 2019. Disponível em: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/avaliacao-quadrienal/avaliacao-quadrienal-2017-2020/metodologia-do-qualis-referencia-quadrienio-2017-2020. Acesso em: 23 ago. 2024.

COELHO JÚNIOR, Abeil; LEMOS, Daniela Lucas da Silva. Tratamento da informação em acervos culturais: avaliação do uso de vocabulários controlados em coleções museológicas sob gestão do Instituto Brasileiro de Museus. **RICI: Revista Ibero-Americana de Ciência da Informação,** Brasília, v. 16, n. 1, p. 131-145, 2023. Disponível em: https://brapci.inf.br/v/219726. Acesso em: 23 ago. 2024.

EVERS, Aline; FINATTO, Maria José Bocorny. Linguística de *corpus*, léxico-estatística textual e processamento de linguagem natural: perspectiva para estudos de vocabulário em produções textuais. **Revista GTLex**, Uberlândia, v. 1, n. 2, p. 271-295, jan./jun. 2016. Disponível em: https://seer.ufu.br/index.php/GTLex/article/view/34711. Acesso em: 23 ago. 2024.

FERREIRA, Hildenise. A Taxonomia Enquanto Estrutura Classificatória: Uma Aplicação em Domínio de Conhecimento Interdisciplinar. **Ponto de Acesso,** Salvador, v. 4, n. 2, p. 131-156, set. 2010. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/4103. Acesso em: 19 de fevereiro 2024.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002.

GONZALES, Marco; LIMA, Vera L. S. Recuperação de informação e processamento da linguagem natural. *In:* CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. 23, 2003, Campinas. **Anais** [...]. Porto Alegre: SBC, 2003, p. 347-395. Disponível em: https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/mri-06---gonzales-e-lima-2003.pdf. Acesso em: 26 ago. 2024.

GUIMARÃES, Nathália Ramos. Bahia é o estado que mais produziu energia eólica no primeiro trimestre de 2023. **Brasil 61,** 06 jun. 2023. Disponível em: https://brasil61.com/n/bahia-e-o-estado-que-mais-produziu-energia-eolica-no-primeiro-trimestre-de-2023-pind234048. Acesso em: 19 de fevereiro 2024.

JANNUZZI, Gilberto de Martino. Uma avaliação das atividades recentes de P&D em energia renovável no Brasil e reflexões para o futuro. Campinas, SP:

Energy Discussion Paper nº 2.64-01/03, 2003. (Energy Discussion Paper). Disponível em: https://www.iei-brasil.org/pdf/reliei-2640103.pdf. Acesso em: 23 fev. 2024.

JORGE, Eduardo Manuel de Freitas; SANTOS, Franciele Portugal dos; CARNEIRO, Breno Pádua Brandão; MACHADO, Fernanda Almeida. Arquitetura da informação analítica para integração de dados da pesquisa e pós-graduação: um estudo de caso da Universidade do Estado da Bahia. **Informação & Informação,** Londrina, v. 25, n. 1, p. 115-140, jan./mar. 2020. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/36009. Acesso em: 23 ago. 2024.

LAUBHEIMER, Page. **Taxonomy 101:** definition, best practices, and how it complements other IA work. *In:* NNGroup, 03 jul. 2022. Disponível em: https://www.nngroup.com/articles/taxonomy-101/. Acesso em: 23 ago. 2024.

LMSYS. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. [S.l.] **LMSYS**, 2023**.** Disponível em: http://www.lmsys.com/vicuna13b. Acesso em: 23 de fevereiro 2024.

MACULAN, enildes Coura Moreira dos Santos; LIMA, Gercina Angela Borém de Oliveira. Taxonomia facetada navegacional: agregando valor às informações disponibilizadas em bibliotecas digitais de teses e dissertações. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília. **Anais** [...]. Brasília: Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2011. Disponível em: https://cip.brapci.inf.br/download/174991. Acesso em: 23 ago. 2024.

META AI. Introducing LLaMA: a foundational, 65-billion-parameter large language model. **Meta,** 24 fev. 2023. Disponível em: https://ai.meta.com/blog/large-language-model-llama-meta-ai/. Acesso em: 23 fev. 2024.

MOOERS, Calvin S. Editor's corner: "coding, information retrieval, and the rapid selector". **American Documentation,** v. 1, n. 4, p. 225-229, oct. 1950. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.5090010409. Acesso em: 25 mar. 2024.

OBSERVATÓRIO DA EPT. **Dashboard da distribuição da produção de energias renováveis no Brasil**. 25 ago. 2022. Disponível em: https://observatorioept.org.br/ept-em-numeros/painel-de-energias-renovaveis. Acesso em: 19 de fevereiro 2024.

SANTOS, M. S. dos; OLIVEIRA, V. H. de J.; JORGE, E. M. de F.; COSTA, G. de M. Solução para Mapeamento e Consulta das Competências dos Pesquisadores: uma arquitetura para extração, integração e consultas de informações acadêmicas. **Cadernos de Prospecção**. Salvador, v. 17, n. 2, p. 671–688, 2024. DOI: 10.9771/cp.v17i2.56670. Disponível em:

https://periodicos.ufba.br/index.php/nit/article/view/56670. Acesso em: 25 jul. 2025.

SINERGIA BAHIA. Bahia é o líder em energia eólica no país. **Sinergia-Ba,** 28 jul. 2019. Disponível em: https://sinergiabahia.com.br/bahia-e-o-primeiro-em-energia-eolica-no-pais/. Acesso em: 19 fev. 2024.

STELAEXPERTA. Plataforma StelaExperta. [Plataforma online]. São Paulo: **StelaTek**, 2023. Disponível em: http://www.stelaexperta.com.br/. Acesso em: 19 fev. 2024.

TERRA, José Cláudio Cyrineu; SCHOUERI, Ricardo; VOGEL, Michely Jabala M.; FRANCO, Carlos. Taxonomia: elemento fundamental para a Gestão do Conhecimento. *[S. l.]*: **Biblioteca TerraForum Consultores,** p. 1-8, [20--?]. Disponível em: http://paginapessoal.utfpr.edu.br/mansano/arquivos/taxonomia.pdf. Acesso em: 19 fev. 2024.

UNIVERSIDADE FEDERAL DE MINAS GERAIS. **Somos UFMG**. [Plataforma online]. Belo Horizonte: UFMG, [s.d.]. Disponível em: https://somos.ufmg.br/ Acesso em: 19 de fevereiro 2024.

RECOVERING RENEWABLE ENERGY EXPERTS THROUGH FACETED TAXONOMY AND NATURAL LANGUAGE PROCESSING TECHNIQUES: AN ACADEMIC DATA MINING EXPERIMENT APPLIED BY RESEARCHERS FROM BAHIA STATE UNIVERSITIES

ABSTRACT

Objective: This article proposes a solution for retrieving textual information from an academic database, using natural language processing techniques to identify renewable energy experts. The solution employs a faceted taxonomy and a competency mapping platform. **Methodology**: The research follows an experimental approach, structured in the following steps: 1) Problem identification and objective definition; 2) Systematic search and review of articles on renewable energy to form the control vocabulary; 3) Construction of the renewable energy taxonomy using the 101 method; 4) Implementation of the search engine; 5) Analysis of the expert researchers' data. The data were cataloged on the simcc.uesc.br platform, including information such as number of publications, Lattes abstracts, relevance indices, and researchers' institutions. **Result:** The development of a search engine and an analytical solution allowed correlating researchers with the renewable energy taxonomy. Applying the faceted taxonomy as a filter resulted in 550 database requests. **Conclusions**: The use of faceted taxonomy and the development of the search engine provided a recovery of experts in renewable energy, demonstrating the effectiveness of the proposed approach in the automatic combination of terms to improve the search and analysis of academic information.

Descriptors: Information searches. Natural language processing. Data mining.

RECUPERACIÓN DE ESPECIALISTAS EN ENERGÍAS RENOVABLES A TRAVÉS DE TAXONOMÍA FACETADA Y TÉCNICAS DE PROCESAMIENTO DEL LENGUAJE NATURAL: UN EXPERIMENTO DE MINERÍA DE DATOS ACADÉMICOS APLICADO POR INVESTIGADORES DE LAS UNIVERSIDADES DEL ESTADO DE BAHÍA

RESUMEN

Objetivo: Este artículo propone una solución para recuperar información textual de una base de datos académica, utilizando técnicas de procesamiento del lenguaje natural para identificar expertos en energías renovables. La solución emplea una taxonomía facetada y una plataforma de mapeo de competencias. **Metodología**: La investigación sigue un enfoque experimental, estructurado en los siguientes pasos: 1) Identificación del problema y definición del objetivo; 2) Búsqueda y revisión sistemática de artículos sobre energías renovables para formar el vocabulario de control; 3) Construcción de la taxonomía de energías renovables utilizando el método 101; 4) Implementación del motor de búsqueda; 5) Análisis de los datos de los investigadores expertos. Los datos se catalogaron en la plataforma simcc.uesc.br, incluyendo información como número de publicaciones, resúmenes de Lattes, índices de relevancia e instituciones de los investigadores. **Resultado**: El desarrollo de un motor de búsqueda y una solución analítica permitió correlacionar a los investigadores con la taxonomía de energías renovables. La aplicación de la taxonomía facetada como filtro resultó en 550 solicitudes a la base de datos. **Conclusiones**: El uso de la taxonomía facetada y el desarrollo del motor de búsqueda proporcionaron una recuperación de expertos en energías renovables, demostrando la efectividad del enfoque propuesto en la combinación automática de términos para mejorar la búsqueda y análisis de información académica.

Descriptores: Buscas de información. Procesamiento del lenguaje natural. Minería de datos.

Recebido em: 27.08.24
Aceito em: 26.06.25