

XX ENANCIB

21 a 25 Outubro/2019 – Florianópolis

A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-2 – Organização e Representação do Conhecimento

AMAZON OU AMAZÔNIA: OS PRINCÍPIOS FAIR APLICADOS AO USO DE TERMOS E CONCEITOS NA BASE SCOPUS

AMAZON OR AMAZÔNIA: FAIR PRINCIPLES APPLIED ON USING TERMS AND CONCEPTS IN THE BASE SCOPUS

Rosali Fernandez de Souza - Instituto Brasileiro de Informação em Ciência e Tecnologia
Marcos Gonçalves Ramos - Instituto Brasileiro de Informação em Ciência e Tecnologia

Modalidade: Trabalho Completo

Resumo: O presente trabalho relata uma pesquisa empírica que buscou analisar possíveis desvios semânticos, ambiguidades e falsos positivos na recuperação de documentos na base de dados de Scopus através de expressões de busca dos termos *Amazon* e *Amazônia*, no período de 2008 a 2018. A partir da seleção de amostra composta por 40 artigos mais citados de ambos os termos, realizou-se uma análise comparativa entre as palavras-chaves do autor versus as palavras-chaves indicadas pela base de dados bem como foram avaliadas as áreas de assunto recuperadas pelos termos escolhidos. Além disso, utilizou-se o software Iramuteq para análise e visualização de dados. O estudo pretende contribuir para uma melhor compreensão da tríade: conceitos, palavras-chaves e termos no que tange a questão da recuperação de informação em base de dados e verificar a aplicação dos princípios encontrável, acessível, interoperável e reutilizável, na sigla em inglês FAIR, na análise de domínio de conhecimento como parte do universo do Sistema de Organização do Conhecimento. Ademais, verificou-se a difícil reciprocidade entre as palavras-chaves dos autores e palavras-chaves da base de dados que talvez possa implicar em algum desvio semântico e ambiguidades na recuperação de documentos.

Palavras-chave: FAIR; Semântica; Organização do Conhecimento; Recuperação de Informação.

Abstract: The present paper reports an empirical research that sought to analyze possible semantic deviations, ambiguities and false positives in the retrieval of documents in the Scopus database through search expressions of the terms *Amazon* and *Amazon*, from 2008 to 2018. From the sample selection composed of 40 most cited articles of both terms, a comparative analysis was performed between the author's keywords versus the keywords indicated by the database and the subject areas retrieved by the chosen terms were evaluated. In addition, Iramuteq software was used for data analysis and visualization. The study aims to contribute to a better understanding of the triad: concepts, keywords and terms regarding database information retrieval and to verify the application of the findable, accessible, interoperable and reusable principles, in the knowledge domain analysis as part of the universe of the Knowledge Organization System. In addition, there was a difficult reciprocity between authors' keywords and database keywords that may imply some semantic deviation and ambiguities in document retrieval.

Keywords: FAIR, Semantic; Knowledge Organization; Information Retrieval.

1 INTRODUÇÃO

Em 21 de maio de 2019, o ICANN - *Internet Corporation for Assigned Names and Numbers*, que administra o sistema de endereços da internet, determinou que a empresa de tecnologia *Amazon* de Jeff Bezos tem o direito de utilizar o domínio na internet “amazon” com suas variações, apesar dos protestos de países sul-americanos que abrigam a floresta amazônica, entre eles o Brasil.

Segundo a Revista Exame¹, o Ministério Brasileiro das Relações Exteriores declarou que “devido a sua inseparável relação semântica com a selva amazônica, esta área não deveria ser de nenhuma maneira o monopólio de uma empresa”, pois a decisão deixaria de considerar a questão da “necessidade de defender o patrimônio natural, cultural e simbólico dos países e povos da região amazônica.

Sem negar a importância da questão jurídica e das relações internacionais, a pesquisa empírica aqui desenvolvida baseou-se na análise de domínio com o objetivo geral de demonstrar se a disputa dos nomes *Amazon* e *Amazônia* teria alguma influência sobre os resultados das expressões de busca na recuperação de documentos em grandes bases de dados.

Na perspectiva do estudo, entendemos que o campo da Organização e Representação do Conhecimento ao lado da Epistemologia apontam para a análise de questões importantes diante do desafio da virtualização do conhecimento armazenado e distribuído por empresas proprietárias de gigantescas bases de dados cujo acesso à informação é dado no contexto de sistemas interligados em rede com múltiplos formatos e com múltiplas etiquetas. (DODEBEI; GUIMARÃES, 2013).

O presente estudo se insere na linha de reflexão em Organização do Conhecimento armazenado e disponível em bases de dados ao lado das questões sobre classificação, sociedade e cultura apontadas por Gutiérrez, (2011):

A rede digital, juntamente com as operações de classificação incessante e rotineiramente promovidas pela cultura ocidental, atua sobre um espaço aberto em que outras civilizações e culturas; também importantes produtores de conhecimento e memórias se tornam mais vulneráveis. A “Digitabilidade” já impõe uma certa ordem lógica no mundo, porque como com qualquer outra tecnologia, é principalmente uma (ordem) “tecnológica”, um instrumento com um alcance simbólico que transfere

¹ REVISTA EXAME. Amazon ganha direito de usar nome da empresa como domínio. Disponível em: <https://exame.abril.com.br/tecnologia/amazon-ganha-direito-de-usar-nome-da-empresa-como-dominio/>. Acesso em: 21 jul. 2019.

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

imperceptivelmente os códigos da cultura que o projetou. (GUTIEEREZ, 2011, p. 2).

O movimento *e-Science* se expressa pelo o uso intensivo de dados por cientistas, técnicos e profissionais liberais em geral e mais conhecido como *Data-Intensive Science* ou *i-Science* referindo se a um conjunto de práticas objetivadas em recursos lógicos formais, instrumentos de registro de imagem e a transferência simultânea de dados em diversos formatos. Os objetivos da *i-Science* visam facilitar e integrar o acesso de usuários a sistemas de maior complexidade de organização, armazenamento e recuperação da informação agora no ecossistema digital da era moderna. (ABBOTT, 2009).

No caso da organização do conhecimento na área das Ciências Ambientais, no caso da oceanografia dos estudos florestais são especialmente importantes os satélites para a captação de imagens e os sensores sonoros, como aponta Monteiro (2015):

Os instrumentos de captação em vários pontos no planeta ou em sua órbita representam a chamada Engenharia de Imagem ou “*imagery*”, cujo impacto na produção da literatura científica é ainda muito pouco estudado em Ciência da Informação. (MONTEIRO, 2015, p.589-91.)

Armazenamento, representação e organização simétricas dos dados na nuvem do bigdata não se reduzem na esfera dos SOC's apenas à discussão de problemas lógico-matemático e padronização de códigos de acesso, mas implicam também em profundas mudanças para a metodologia das Ciências em termos da classificações de novas áreas de conhecimento (SOUZA, 2004) e de novos comportamentos dos atores envolvidos nos processos de comunicação e de produção acadêmica, sobretudo voltados para o tema da Amazônia e das Ciências Ambientais de modo amplo.

O objetivo específico do estudo foi de demonstrar a possibilidade metodológica de reutilização de dados de pesquisa em forma de resumos, palavras-chaves e termos retirados de artigos científicos indexados na Scopus, utilizando como referencial teórico e prático os princípios FAIR (na sigla original em inglês) a saber: *Findable*: para ser encontrável qualquer objeto de dados deve ser exclusivamente e persistentemente identificável; *Accessible*: os dados devem acessíveis por máquinas ou humanos mediante autorização adequada e/ou através de um protocolo bem definido; *Interoperable*: os objetos de dados podem ser interoperáveis quando os (meta) dados são formatos de dados acionáveis por máquina ou humanos que (meta) utilizam vocabulários compartilhados e/ou ontologias; e *Reuseable*: os objetos de dados são reutilizáveis somente em conformidade com os princípios anteriores. (WILKINSON *et al.*, 2016).

A comunicação relata, portanto, um estudo experimental com utilização do método qualitativo de análise de domínio aliado à análise quantitativa da Cientometria, com a utilização do programa informático de análise textual Iramuteq, gratuito e de código aberto.

2 ABORDAGEM CONCEITUAL

Barité (2001) conceitua a Organização do Conhecimento como disciplina que trata do desenvolvimento de técnicas para construção, gestão e uso, avaliação de classificações científicas, taxonomias, nomenclatura e linguagens documentais. Trata-se da visão integrada do conhecimento, na qual se associam classificações filosóficas ou científicas do saber com classificações destinadas à organização de documentos em bibliotecas, arquivos e outras unidades de informação. Em síntese, o objeto de estudo da Organização do Conhecimento é o conhecimento socializado. (BARITÉ, 2001).

Bräscher (2014) destaca que os estudos sobre Sistemas de Organização do Conhecimento (SOC) têm se voltado cada vez mais para a importância das estruturas linguísticas no que se refere a análise dos problemas de conceituação e atribuição de significado, pois servem como fonte do “mapeamento sintagmático de um dado domínio.” (BRASHER, 2014, p. 178).

Hjørland (2007) argumenta que para a definição e delimitação dos Sistemas de Organização do Conhecimento (SOC) é fundamental observar os princípios e métodos da área de semântica, pois os SOC são sistemas de relações semânticas entre conceitos, palavra e termos que se inter-relacionam de modo dinâmico e orgânico (HJØRLAND, 2007).

Neste sentido, o autor questiona a proposta realizada por Hodge (2000) sobre a abrangência dos SOC não restrita apenas às listas de relacionamento como tesouros, taxonomias, redes semânticas e ontologias, mas também encobrem a área de Bibliometria. Os mapas de citação devem ser considerados como um SOC, pois refletem comportamentos de citações e distâncias semânticas entre os autores (HODGE, 2000).

Hjørland (2007) crítica a ausência de estudos sobre semântica na área de Ciência da Informação que tradicionalmente se ocupa de questões de terminologia entre palavras e conceitos para normalização de palavras-chaves para fins unicamente de indexação de termos (HJØRLAND, 2007).

Sowa (1984) observa que conceitos e percepções são abstrações que selecionam

importantes aspectos para um determinado propósito e ignoram detalhes ou complexidades que são somente importantes para outras finalidades.

Uma mesma entidade pode ser caracterizada em diferentes modos dependendo da área e do tipo do sistema de organização de conhecimento que desejamos construir. A arbitrariedade da seleção de grupos e relacionamentos de conceitos não podem ser, entretanto, como algo totalmente aleatório. (SOWA, 1984, p. 344, tradução nossa).

Zeng (2008) destaca que os serviços de informação atuais devem observar três níveis de operação: conceitual, lógico e físico como previstos nos SOC's. Os problemas de sinonímia, hierarquia e classificação de termos e conceitos nos sistemas convencionais de informação são praticamente os mesmos como, por exemplo, os problemas discutidos nas revisões sobre o papel dos motores de busca na recuperação de dados nas redes, como também a dificuldade para organização de index de assuntos de uma publicação seriada em meio impresso ou digital (ZENG, 2008).

A indexação extrai ou atribui ao texto um conjunto de termos ou frases que funcionam como palavras-chave. Os termos ou frases atribuídas são chamados de termos de índice de linguagem controlada quando vêm de um vocabulário fixo. Os termos do índice podem ser usados como pontos de acesso ou identificadores do texto na coleção de documentos. Este formato de representação de texto é utilizado em sistemas de recuperação de informação ou filtragem (MOENS, 2002, p. 24).

Para Lync (2007), após a transição do modelo impresso para o digital, a comunicação científica no contexto atual da *e-Science* encontra-se no estágio de interconexão de dados e informação, e que já superaram muito dos processos previstos de indexação mesmo no universo dos repositórios e bibliotecas digitais. As Bases de dados e outros materiais de informação estão interligados em uma variedade de modos de comunicação e transferência de informação muito mais complexos:

O modelo de artigo científico agora autorizado a migrar para o formato digital é uma composição convergente com estrutura de todos os recursos disponíveis no ambiente digital, isto é, progressivamente o artigo em meio digital abandona o modelo do periódico eletrônico que até então se restringia a disponibilização de artigos impressos digitalizados. (LYNCH, 2007, p. 9, tradução nossa).

Bräscher (2014) se aproxima das definições de Dahlberg (1993) quando afirma que o objeto principal dos estudos da organização de conhecimento são os conceitos e não as expressões usadas para se referirem a eles. Assim, os termos, os códigos, as fórmulas ou

outros símbolos são considerados como etiquetas, isto é, elementos que respondem a uma lógica relacional no SOC ao passo que os conceitos representam relações semânticas mais amplas. (BRASHER, 2014).

Para fins de concisão e clareza nos utilizamos da definição de conceito como qualquer "unidade de pensamento" definida para a qual nos referimos nos formatos dos objetos digitais nas redes. Quando usamos o termo (meta) dados aqui, pretendemos indicar que o princípio é verdadeiro para metadados no sistema, bem como para o real, mas o princípio em questão pode ser implementado de forma independente para cada um deles (THE ROYAL SOCIETY, 2012, p.12).

3 METODOLOGIA

A pesquisa utilizou os dois métodos qualitativo, análise de domínio, e quantitativo, Cientometria, para demonstrar possíveis desvios semânticos na recuperação de informação.

Macias Chapula (1998) defini Cientometria como:

[...] um segmento da sociologia da ciência, sendo aplicada no desenvolvimento de políticas científicas. Envolve estudos quantitativos das atividades científicas, incluindo a publicação e, portanto, sobrepondo-se à Bibliometria. (MACIAS CHAPULA, 1998, p. 134).

A análise de domínio de conhecimento é um método usado para demarcar comunidades discursivas ou epistêmicas que desenvolvem diferentes gêneros de documentos, ou melhor, diferentes composições de linguagens documentárias com diversos pontos potenciais de acesso para a recuperação da informação (HJØRLAND, 2002).

As comunidades são normalmente influenciadas por várias normas epistemológicas ou tendências, mas também influenciam a construção social dos sistemas simbólicos não restritos ao domínio da comunicação científica e das humanidades, e que também interagem com a mídia e a comunicação social voltada para a esfera pública. (HJØRLAND, 2002, p. 446-7).

O propósito da pesquisa empírica foi identificar, analisar e discutir se a disputa dos nomes *Amazon* e *Amazônia* teria alguma influência sobre os resultados dos termos de expressões de busca na recuperação de documentos na base Scopus, no período de 2008 a 2018, sendo a coleta de dados realizada no dia 19 de abril de 2019 cujos resultados totais de documentos recuperados constam na tabela 1 abaixo:

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

Tabela 1: Resultado da expressão de busca (2008-2018).

Termo	Expressão	Resultado
<i>Amazon</i>	<i>TITLE-ABS-KEY (Amazon) AND PUBYEAR > 2007 AND PUBYEAR < 2019</i>	24.265
Amazônia	<i>TITLE-ABS-KEY (Amazônia) AND PUBYEAR > 2007 AND PUBYEAR < 2019</i>	7.805

Fonte: Elaborado pelos autores.

A escolha da base Scopus pertencente à Elsevier foi devido ao seu reconhecimento na comunidade científica e pelo seu padrão de qualidade e curadoria de dados, como também pela sua cobertura contendo 22.800 títulos de publicações alimentadas por mais de 5.000 editores internacionais.

O trabalho foi realizado em etapas: a primeira etapa, caracteriza-se pela comparação da quantidade de documentos recuperados pelos termos *Amazon* e Amazônia por área de assunto, conforme descritores da base Scopus.

Na segunda etapa, realizou-se a análise dos resultados das palavras-chaves enquanto metadados na recuperação dos documentos relacionados às expressões de busca de *Amazon* e Amazônia.

Convém lembrar que o objetivo geral do estudo não é em relação à padronização de termos e palavras-chaves enquanto metadados utilizados na base. Segundo JEFFERYA (2014) metadado é um dado utilizado para algum propósito dentro de um sistema e relacionado a uma camada de representação com outros termos. A padronização do metadado pode levar a construção de uma arquitetura eficiente de informação. (JEFFERYA, 2014).

Na terceira etapa, com a finalidade de analisar e ilustrar a complexidade das relações da tríade conceitos, palavras-chaves e termos selecionou-se uma amostra composta por 20 artigos mais citados para cada de um dos termos, no total de 40 artigos de acordo com os critérios de ordenação e recuperação da base Scopus.

Inicialmente, realizou-se uma análise quantitativa comparando a presença das palavras-chaves dos autores versus do índice da base de dados. Em seguida, utilizou-se as análises do programa Iramuteq (*Interface de R pour analyses Multidimensionnelles de Textes et de Questionnaires*), de análise textual, gratuito e de código aberto.

O programa Iramuteq foi criado por Pierre Ratinaud, no Laboratório de Estudos e Pesquisa Aplicada em Ciências Sociais em Toulouse, unindo as linguagens de programação estatística R e Python, no qual foram incorporadas as análises lexicais do programa Alceste. O programa possibilita diferentes análises textuais como, por exemplo, análises lexicais clássicas

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

que identificam a quantidade de palavras, a frequência média e o número de palavras com frequência igual a um.

Cabe ressaltar que o programa escolhido tem sido utilizado na Academia Brasileira desde 2013, a partir da iniciativa de grupos de pesquisa como da UNESP que o adaptou para língua portuguesa. O Iramuteq não é um método de análise de dados, e sim uma ferramenta de processamento de grandes volumes de dados de preferência de fontes textuais, sendo necessária a interpretação do pesquisador (CAMARGO; JUSTO, 2013).

Na quarta etapa, realizou-se a análise qualitativa dos resultados tendo como base a análise de domínio que possibilitou contextualizar a aplicação do princípio de reutilização de dados.

4 RESULTADO E ANÁLISE DE DADOS

O resultado da primeira etapa demonstrou que os termos recuperam diferentes volumes de documentos em cada área de assunto, conforme Tabela 2.

Tabela 2: Comparativo de resultado por área de assunto

Áreas de Assunto	<i>Amazon</i>	Amazônia
Agricultura e Ciências Biológicas	7.613	4.097
Artes e Humanidades	1.121	530
Bioquímica, Genética e Biologia Molecular	1.777	520
Negócios, Gestão e Contabilidade	938	72
Engenheiro químico	158	23
Química	398	104
Ciência da Computação	5.759	76
Ciências da decisão	590	45
Odontologia	28	1
Ciências da Terra e Planetárias	2.826	1.483
Economia, Econometria e Finanças	519	102
Energia	458	112
Engenharia	2.064	114
Ciência Ambiental	3.757	2.033
Profissões de Saúde	67	1
Imunologia e Microbiologia	1.074	245
Ciência de materiais	301	40
Matemática	1.246	32
Medicina	2.681	574
Multidisciplinar	393	150
Neurociência	166	24

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

Enfermagem	112	14
Farmacologia, Toxicologia e Farmacêutica	399	57
Física e Astronomia	485	109
Psicologia	514	21
Ciências Sociais	3.018	1.271
Indefinido	1	1
Veterinária	568	163
Total	39.031	12.014

Fonte: Elaborado pelos autores.

Diferenças foram percebidas entre o resultado total de documentos por área de assunto (tabela 2), e o resultado total da recuperação das expressões de busca dos termos *Amazon* e *Amazônia* (tabela 1).

Para o termo *Amazon*, as áreas de assunto que tiveram destaque foram: Agricultura e Ciências Biológicas em primeiro lugar com 7.613 documentos recuperados correspondendo a 20% do total de 39.031; a Ciência da Computação em segundo lugar com 5.759 ou 15% do total; Ciências Ambientais em terceiro com 3.757 ou 10% do total; e Ciências Sociais em quarto lugar com 3.018 ou 8 % do total.

As áreas de assunto mais representativas para o termo *Amazônia* foram as seguintes: Agricultura e Ciências Biológicas em primeiro lugar com 4.097 documentos recuperados correspondendo a 34% do total de 12.014; Ciências Ambientais em segundo com 2.033 ou 17% do total; Ciências da Terra e do Planeta em terceiro com 1.483 ou 12% do total; e Ciências Sociais em quarto com 1.271 ou 11% do total.

Notou-se que as principais áreas de assunto apareceram na recuperação de ambos os termos de busca *Amazon* e *Amazônia* tais sejam: Agricultura e Ciências Biológicas em primeiro lugar; Ciências Sociais em quarto lugar; e Ciências Ambientais em posições diferentes.

Convém mencionar que a área de Ciência da Computação aparece em ambos os resultados de busca, porém para *Amazon* representou 15% (5.759) total de documentos recuperados e já em *Amazônia* foi apenas de 0,63% (76) do total.

A diferença de percentual de recuperação de documentos aparentemente indica a possibilidade de ambiguidade no processo de recuperação de documentos, pois o universo semântico que envolve a empresa de comercio eletrônico internacional *Amazon.com, Inc.* parece se estender ao universo semântico do termo *Amazônia*. Neste sentido, o conceito de reutilização dos dados talvez possa estar comprometido devido a indeterminação dos usos de conceitos, descritores ou termos usados alternadamente na base de dados Scopus.

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

Cabe ressaltar que na o resultado mostrou que os artigos podem fazer parte de mais de uma área de assunto, o que dificulta uma comparação precisa, no que se refere ao grau de vinculação entre a palavra-chave e a área de assunto.

Nessa perspectiva, pode-se inferir que Amazônia é objeto de estudo de diversas disciplinas e projetos. Entretanto, vale questionar se os resultados indicam a pluralidade das áreas de assunto ou a fragmentação do conceito Amazônia e *Amazon* em miríade de termos que podem implicar diferentes resultados de busca, por conta de questões de significação que precisam ser melhor estudadas.

O resultado da segunda etapa demonstrou que o total da busca do termo *Amazon* foi de 85.772 palavras-chaves ou metadados e o total do termo Amazônia foi 34.442 metadados, na tabela 5, no apêndice. O que significa que o resultado do termo Amazônia representa aproximadamente 40% do total de *Amazon*.

As tabelas 3 e 4 apresentam as dez palavras-chaves mais frequentes nos resultados das expressões de busca.

Tabela 3: Frequência de Keywords do Resultado de *Amazon*.

Nº	Termo	Frequência
1	Brazil	4.751
2	Article	3.520
3	Human	2.463
4	<i>Amazon</i>	2.385
5	<i>Amazonia</i>	2.322
6	Humans	2.048
7	Animals	2.010
8	Female	1.945
9	Male	1.900
10	Animal	1.652

Fonte: Elaborado pelos autores.

Tabela 4: Frequência de Keywords do Resultado de Amazônia.

Nº	Termo	Frequência
1	<i>Amazonia</i>	4.046
2	Brazil	2.071
3	Article	889
4	<i>Amazon</i>	720
5	South America	658
6	Animals	609
7	Deforestation	591
8	Tropical Forest	549
9	Rainforest	541
10	Animal	522

Fonte: Elaborado pelos autores.

Verificou-se nos resultados de ambas as expressões de busca que houve a repetição de diversas palavras-chaves tais como: *Brazil*, *Article*, *Amazon*, *Amazonia*, *Animals* e *Animal*. Na tabela 3, o resultado de *Amazon* mostrou palavras utilizadas no singular e no plural como, por exemplo, *Humans* e *Human*, bem como *Animals* e *Animal*, que poderiam ser normalizadas em apenas um termo.

Convém mencionar que a falta de padronização pode gerar problemas de sinonímia, redundância e ambiguidade que, consequentemente, proporcionam falsos positivos em termos de estatísticas de frequência de palavras e de recuperação de documentos. Tal fato confirma a discrepância dos resultados totais de documentos nas expressões de busca dos termos *Amazon* e *Amazônia* em relação ao somatório dos documentos por área de assunto, presentes nas tabelas 1 e 2.

O exemplo mais evidente foi o uso do termo *Article* como palavra-chave em ambos os resultados das expressões de busca. Compete mencionar que *Article* pode recuperar documentos sem relevância, pois refere-se ao principal tipo de documento armazenado na base de dados em sua totalidade.

Na terceira etapa, o resultado do termo *Amazônia* nos 20 artigos analisados mostrou que 8 dispunham de palavras-chaves indicadas pelo autor (*author keywords*); 19 tinham palavras-chaves da base de dados (*index keywords*); 1 artigo apresentou palavras-chaves do autor, mas não constou palavras-chaves indicadas pela base de dados; e 7 tinham palavras-chaves de ambos.

Observou-se que nos resumos dos 20 artigos selecionados, o termo *Amazônia* apareceu em 16 artigos; *Amazon* em 4 artigos; e em apenas 1 artigo ocorreram ambos os termos. Nas palavras-chaves dos autores de 8 artigos, o termo *Amazônia* não foi identificado, porém o termo *Amazon* apareceu em 1 artigo.

Nas palavras-chaves da base de dados de 19 artigos, o termo *Amazônia* apareceu em 15 artigos e não apareceu em 4, porém o termo *Amazon* apareceu juntamente com o termo *Amazônia* em 2 artigos.

O resultado da análise comparativa das palavras-chaves de autores (*author keywords*) versus as palavras-chaves da base de dados (*index keywords*) mostrou que não existe correspondência entre a quantidade e a forma de representação dos conceitos originais

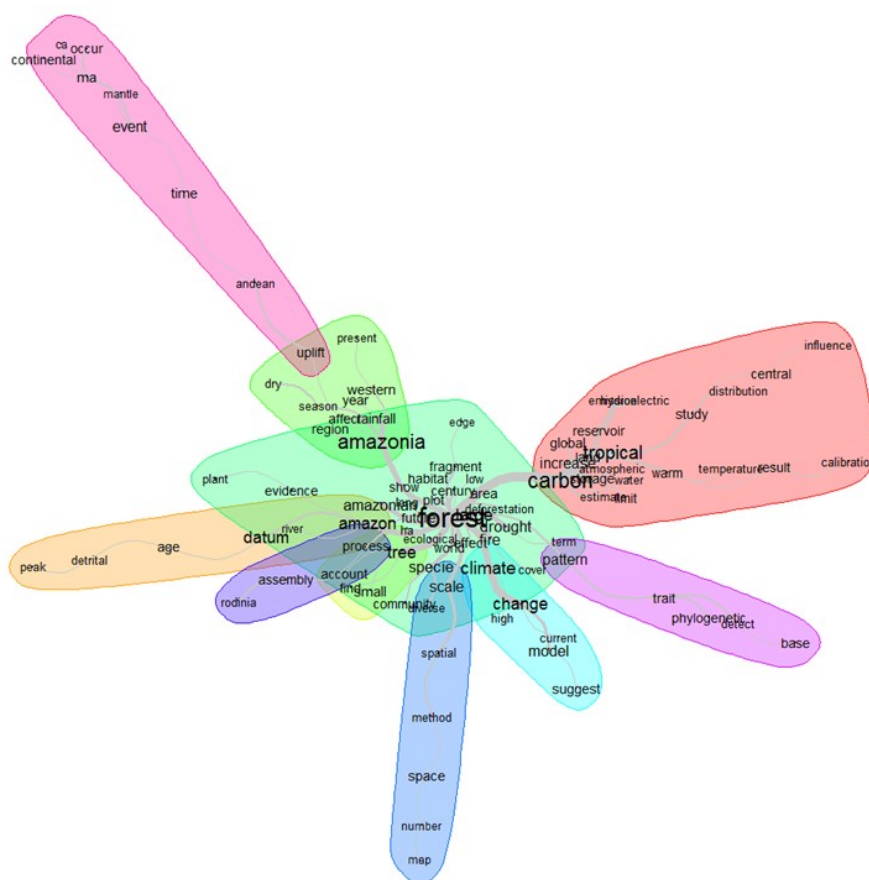
A primeira análise do Iramuteq foi a nuvem de palavras que organiza e agrupa os termos graficamente em função da sua frequência. A visualização gráfica permite a

A nuvem de palavras dos artigos de Amazônia na figura 1 destaca a palavra *forest* (frequência de 49 vezes) no centro da nuvem e mostra uma atração semântica entre os demais termos, os quais remetem ao universo do ecossistema amazônico.

A segunda foi a análise de similitude ou de semelhança é baseada na teoria dos grafos que representa a ligação entre as palavras do corpus textual. “A Análise de Similaridade (ADS) possibilita representar graficamente a estrutura de um corpus, distinguindo também as partes comuns e as especificidades de um corpus”. (MARCHAND; RATINAUD, 2012)

Na análise de similitude dos resumos de Amazônia na figura 2, a palavra *forest* possui uma posição nuclear e tem forte relação com termos *Amazon, tree, climate, change, carbono, tropical, Amazonia* e *datum*. Apenas um conjunto está mais afastado do aglomerado central, unido pelo termo *uplift* ao conjunto de *Amazonia*. Também, notou-se que existe um movimento do centro para as extremidades por parte do conjunto de palavras.

Figura 2: Resultado da similitude dos artigos de Amazônia.



Fonte: Elaborado pelos autores com base no Iramuteq.

Já no resultado do termo *Amazon* na análise dos 20 artigos selecionados, observou-se que 14 artigos possuem palavras-chaves do autor (*author keywords*); 14 artigos têm palavras-chaves indicadas pela base de dados (*index keywords*); 1 artigo não possui palavras-chaves de ambos; e 9 artigos possuem palavras-chaves do autor e da base de dados.

Nos resumos dos 20 artigos selecionados, o termo *Amazon* apareceu em 18 artigos; o termo *Amazônia* apareceu em 1 artigo e não constam os termos em 2 artigos. Além disso, identificou-se que dos 20 artigos selecionados 16 artigos têm relação com os serviços da *Amazon.com* e 4 artigos se referem a *Amazônia*.

Nas palavras-chaves dos autores dos 14 artigos, o termo *Amazon* apenas apareceu em 1 artigo. Nas palavras-chaves indicadas pela base de dados dos 14 artigos, o termo *Amazon* apareceu em 2 artigos e o termo *Amazônia* apareceu em outros 2 artigos.

A análise comparativa das palavras-chaves de autores (*author keywords*) versus as indicadas pela base de dados (*index keywords*) mostrou uma reciprocidade em termos de

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

quantidade e na forma de representação dos conceitos originais propostos pelos autores. Além disso, percebeu-se a existência de uma correspondência dos conceitos das palavras-chaves com o universo da tecnologia em geral. Tal fato talvez contribua para a maior probabilidade de que as palavras-chaves indicadas pelo autor sejam aderidas e replicadas no conjunto das palavras-chaves da base de dados para fins de recuperação.

Para melhor compreensão, realizou-se a nuvem de palavras bem como análise de similitude dos resumos dos 20 artigos mais citados na busca do termo *Amazon*.

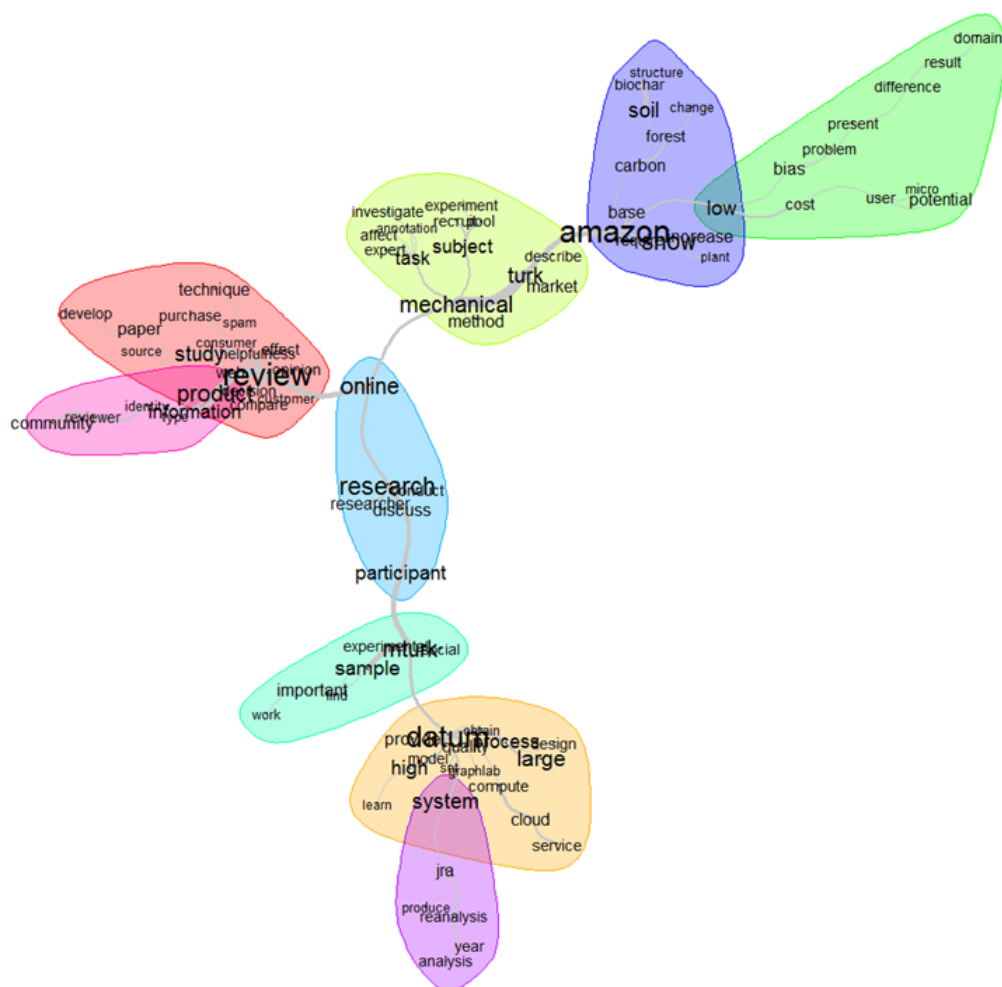
Figura 3: Nuvem de palavras dos artigos de *Amazon*.



Fonte: Elaborado pelos autores com base no Iramuteg.

Na nuvem de palavras dos resumos de *Amazon*, a centralidade da palavra review (frequência de 22 vezes) não demonstra a convergência dos demais termos. Ao contrário, observa-se uma maior heteronímia dos conceitos apresentados pelos termos.

Figura 4: Resultado da similitude dos artigos de *Amazon*.



Fonte: Elaborado pelos autores com base no Iramuteq.

Na análise similitude dos resumos de *Amazon* figura 3, a palavra *review* não possui uma posição nuclear como no caso da palavra *forest* na figura 2. Também, observou-se que os termos *Amazon*, *mechanical*, *online research*, *mturk*, *datum* e *system* compõe conjuntos semânticos aparentemente isolados. E o termo *online* está em uma posição que liga as extremidades dos agrupamentos de *review*, *mechanical* e *research*. Ademais, percebeu-se que a semântica do conjunto se refere ao universo da tecnologia como na nuvem de palavras na figura 4.

Na quarta fase, o resultado da análise do corpus textual demonstrou que as diferenças quantitativas refletem nas distancias semânticas entre os termos, conceitos e palavras-chaves dos nomes *Amazon* e *Amazônia*. É interessante mencionar que as análises do Iramuteq não se restringem apenas a análise estatística, pois confirmaram as diferentes configurações semânticas dos domínios de *Amazon* e *Amazônia*.

Destaca-se também o fato de que se de um lado os documentos recuperados são acessíveis e encontráveis na base, as distâncias semânticas podem comprometer o princípio de interoperabilidade entre termos e palavras-chaves dificultando a operacionalização do critério de reutilização de objetos de dados aos quais os termos se referem.

6 CONSIDERAÇÕES FINAIS

A utilização da análise de domínio ao lado da aplicação de técnicas quantitativas do programa Iramuteq serviu para demonstrar a aproximação dos resultados do presente estudo com as proposições dos autores Hjørland (2002, 2007), Jonh Sowa (1984) e Marisa Bräscher (2014); que apontam para a importância dos processos de significação para o entendimento dos Sistemas de Organização do Conhecimento (SOC) enquanto sistemas de relações semânticas entre conceitos e termos que se desenvolvem em um determinado contexto histórico e social.

O estudo comparativo dos termos *Amazon* e *Amazônia* assinala que a inseparável relação semântica com a selva amazônica” se reflete na (re)classificação do termo *Amazon* na sua extensão semântica com a tecnologia da empresa *Amazon*. O que pode explicar, em parte, uma maior correspondência das palavras-chaves dos autores (*author keywords*) com as palavras-chaves da base de dados (*index keywords*).

A nuvem de palavras e análise de similitude realizadas através do software Iramuteq permitiu uma percepção ampliada do *corpo textual* composto dos 20 artigos selecionados de modo a alcançar uma visão crítica dos processos de significação e representação de termos, conceitos e palavras-chaves para a construção de metadados.

A partir dos resultados dos estudos comparativos, os termos *Amazon* e *Amazônia*, no domínio da Scopus refletem configurações de termos e palavras que parecem indicar a pluralidade discursiva da biodiversidade da floresta Amazônia agora digitalizada pelos mais diversos instrumentos em especial radares e satélites que alimentam sistemas de geolocalização dos seus diferentes ecossistemas, o que pode significar uma relação de disputa e de biopoder:

A Amazônia, em particular, constitui um cenário territorial de suma importância, no que se refere aos desdobramentos práticos dos desafios e impasses hoje colocados internacionalmente em torno da conservação e do uso sustentável da biodiversidade, expressos na Convenção sobre diversidade biológica CDB, podendo ser vista, deste modo, como um campo

de ação “avançada” dos conflitos aí existentes, bem como de suas possíveis soluções, onde se mesclam e se articulam os diferentes níveis do global ao local. (ALBAGLI, 1998, p. 6).

Os termos de buscas e os desvios semânticos analisados podem refletir um conflito de representação política e resistência de (des)classificação entre os diversos agentes. Assim, podemos inferir que o grande volume da produção científica relacionado com a tecnologia na base de dados contribui para o aperfeiçoamento dos algoritmos na recuperação de documentos de modo eficiente e eficaz, ou, ao contrário, os temas com menor volume de dados armazenados, podem implicar em situações de ambiguidade e maior dificuldade de recuperação precisa de informação.

O domínio comercial do nome *Amazon* reconhecido recentemente à empresa *Amazon.com* indica a necessidade de estudos de classificação dos termos, bem como os conceitos em bases de dados nacionais e internacionais que levem também em conta a complexidade da floresta amazônica no contexto político, social e econômico do Antropoceno:

Em relação as suas populações, entre 180 grupos étnicos e com mais de 100 línguas, a região da Amazônia sempre esteve em relação de disputa com os interesses das políticas governamentais entre demarcações de terras e exploração de seus recursos materiais (ISSBERNER; LÉNA, 2019, p.62). (tradução nossa).

Os termos *Amazon* e *Amazônia* parecem indicar igualmente uma disputa entre espaços materiais geolocalizados e os sistemas simbólicos de representação no espaço da internet/web que precisa ser melhor estudada.

A aplicação dos princípios FAIR apresentada no estudo remete à sua importância para garantir não apenas uma recuperação eficaz de documentos, mas pode sobretudo ser um instrumento de reavaliação de políticas de indexação e representação em grandes bases de dados.

O presente experimento também demonstrou que a aplicação do software Iramuteq em relação aos princípios FAIR juntamente com a análise de domínio permitiram uma reflexão robusta sobre os resultados dos termos de busca na Scopus.

A utilização dos diversos tipos de modelagens, visualização e simulação de padrões complexos de estruturação de dados e informações com base de cálculo estatísticos não exclui a abordagem de caráter qualitativo para fins de análise crítica dos resultados.

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

Para futuros trabalhos, pretendemos ampliar a amostra envolvendo termos e diferentes bases de dados, não apenas para contribuir para os aspectos de recuperação da informação como também para entendermos os processos de significação, controle e construção de metadados utilizados em estratégias de classificação e indexação.

Agradecimentos : Agradecimentos aos professores Rosali Fernandez de Souza e Jorge Biolchini da cadeira Estrutura e Fluxo da Informação, bem como aos professores Gustavo Saldanha e Luana Salles da cadeira de Organização de Domínios de Conhecimento do PPGCI – IBICT/UFRJ, pelos conceitos adquiridos e seminários realizados nas disciplinas. Além do apoio da mestranda Priscila Ramos Carvalho.

REFERÊNCIAS

ABBOTT, Mark. R. **A new path for science?** *In*: The Fourth Paradigm Data-Intensive Scientific. Microsoft Corporation Microsoft Research. 2. ed. version 1.1, USA. October 2009.

ALBAGLI, Sarita. **Geopolítica da biodiversidade**. Brasília: Ibama, 1998

BARITÉ, Mário. **Organización del Conocimiento**: un nuevo marco teórico conceptual en Bibliotecología y Documentación. *In*: Kester Carrara ed. Educação, Universidade e pesquisa. Marília: Unesp Marília, 2001.

BRACHER, Marisa. Semantic Relations in Knowledge Organization Systems. **Knowl. Org.**, 41, n.2, 2014.

CAMARGO, Brígido Vizeu; JUSTO, Ana Maria. IRAMUTEQ: um software gratuito para análise de dados textuais. **Temas em Psicologia**, Ribeirão Preto, v. 21, n. 2, 2013, p. 513-518.

DAHLBERG, Ingetraut. knowledge Organization: its scope and Possibilities. **Knowledge Organization**, v. 20, 1993.

DODEBEI, Vera; GUIMARÃES, José A. C. **Prefácio**. *In*: Complexidade e organização do conhecimento: desafios de nosso século. Rio de Janeiro: ISKO-Brasil; Marília: FUNDEPE, 2013. (e-Book ISBN: 978-85-98176-51-2).

JEFFERYA, Keith G; ASSERSONB, Anne; HOUSSOSC, Nikos; BRASSED, Valerie; JÖRGE, Brigitte. From Open Data to Data-Intensive Science through CERIF. **Procedia Computer Science**, 33, p. 191-198, 2014.

HODGE, Gail. **Systems of knowledge organization for Digital libraries**. Beyond Traditional Authority, 2000.

HJØRLAND, Birger. **Semantic and Knowledge Organization**. Arist 41, 2007.

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

HJØRLAND, Birger. Domain analysis in Information science. Eleven approaches – traditional as well as Innovative. **Journal of Documentation**, v. 58, n. 4, p. 447-8, 2002.

GUTIÉRREZ, A. G. Desclassification in Knowledge organization: a post-colonial essay. **Transinformation**. Campinas, v. 23, n. 1, p. 5-14, 2011.

ISSBERNER, Liz-Rejane; LÉNA, Philippe. **Brazil in the Anthropocene**. Conflicts between predatory development and environmental policies, Edited by Liz-Rejane Issberner, Philippe Léna Routledge, 2019, 368 p.

LYNCH, Clifford. A. The Shape of the Scientific Article in the Developing Cyberinfrastructure. **CT Watch**, v. 3, n. 3, p. 5-11, aug. 2007.

MACIAS-CHAPULA, Cesar A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ci. Inf. [online]**, v. 27, n. 2, 1998.

MARCHAND, Pascal; RATINAUD, Pierre. L'analyse de similitude appliqué aux corpus textuelles: les primaires socialistes por l' election présidentielle française. In: JOURNÉES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNÉES TEXTUELLES, 11., JADT, 2012. **Actas [...]**. 2012. p.687-699

MOENS, Marie-Francine. **Automatic indexing and abstracting of documents texts**. Kluwer Academic Publishers, 2002

MONTEIRO, M. Construindo imagens e territórios: pensando a visualidade e a materialidade do sensoriamento remoto. **História, Ciências, Saúde Manguinhos**, RJ., v. 22, n. 2, abr./jun., p.577-591. 2015

SOUZA, Rosali Fernandez de. Áreas do conhecimento. **DataGramZero**: Revista de Ciência da Informação, v.5, n. 2, 2004.

SOWA, John F. **Conceptual structures**: information processing in mind and machine. Massachussetts: Addison-Wesley, 1984

THE ROYAL SOCIETY. **Science as an open enterprise**. London: The Royal Society Science Policy Centre, 2012. Disponível em: <https://royalsociety.org/~media/policy/projects/sape/2012-06-20-saoe.pdf>. Acesso em: 23 set. 2019.

ZENG, Márcia L. Knowledge Organization Systems (KOS). **Knowledge Organization**, v. 35, n. 2-3, p.160-182, jan. 2008.

WILKINSON, M. D. et al. **The FAIR Guiding Principles for scientific data management and stewardship**. Scientific Data Journal, 2016. Disponível em: <https://FAIRdomhub.org/publications/292>. Acesso em: 07 jul. 2019.