

Programa de Pós-Graduação em Ciência da Informação

Isaque Katahira

**ANÁLISE DE LOGS DE BUSCA COMO FONTE INFORMACIONAL PARA
ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO EM REPOSITÓRIOS
INSTITUCIONAIS**

**Marília-SP
2022**

Programa de Pós-Graduação em Ciência da Informação

Isaque Katahira

**ANÁLISE DE LOGS DE BUSCA COMO FONTE INFORMACIONAL PARA
ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO EM REPOSITÓRIOS
INSTITUCIONAIS**

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Universidade Estadual Paulista “Júlio de Mesquita Filho” - *Campus* de Marília, como requisito para a obtenção do título de Doutor em Ciência da Informação.

Orientadora: Prof.^a Dr.^a Mariângela Spotti Lopes Fujita.

Coorientador: Prof. Dr. Isidoro Gil Leiva.

Área de concentração: Informação, Tecnologia e Conhecimento.

Linha de pesquisa: Produção e Organização da Informação.

K19a

Katahira, Isaque

ANÁLISE DE LOGS DE BUSCA COMO FONTE INFORMACIONAL
PARA ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO EM
REPOSITÓRIOS INSTITUCIONAIS / Isaque Katahira. -- Marília, 2022
165 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de
Filosofia e Ciências, Marília

Orientadora: Mariângela Spotti Lopes Fujita

Coorientadora: Isidoro Gil Leiva

1. Análise de logs. 2. Repositórios institucionais. 3. Análise de log de
transações. 4. Vocabulário controlado. 5. Logs de buscas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Filosofia e
Ciências, Marília. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

ISAQUE KATAHIRA

**ANÁLISE DE LOGS DE BUSCA COMO FONTE INFORMACIONAL PARA
ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO EM REPOSITÓRIOS
INSTITUCIONAIS**

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Universidade Estadual Paulista “Júlio de Mesquita Filho” - *Campus* de Marília, como requisito para a obtenção do título de Doutor em Ciência da Informação.

BANCA EXAMINADORA

Prof.^a Dr.^a Mariângela Spotti Lopes Fujita (Orientadora)

Docente permanente do Programa de Pós-Graduação em Ciência da Informação (PPGCI), Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – UNESP *Campus* de Marília.

Prof.^a Dr.^a Flávia Maria Bastos

Membro Titular Interno

Coordenadoria Geral de Bibliotecas – Reitoria da Unesp.

Prof. Dr. Walter Moreira

Membro Titular Interno

Docente permanente do Programa de Pós-Graduação em Ciência da Informação (PPGCI), Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – UNESP *Campus* de Marília.

Prof. Dr. Rogério Aparecido Sá Ramalho

Membro Titular Externo

Departamento de Ciência da Informação da Universidade Federal de São Carlos.

Prof. Dr. Renato Fernandes Correa

Membro Titular Externo

Departamento de Ciência da Informação e ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco.

Membros suplentes:

Prof. Dr. Edberto Ferneda

Membro Suplente Interno

Docente permanente do Programa de Pós-Graduação em Ciência da Informação (PPGCI), Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – UNESP *Campus* de Marília.

Prof.^a Dr.^a Maria Luiza de Almeida Campos

Membro Suplente Externo

Docente permanente do Programa de Pós-Graduação em Ciência da Informação (PPGCI), Universidade Federal da Bahia.

Prof. Dr. José Fernando Modesto da Silva

Membro Suplente Externo

Departamento de Informação e cultura, Universidade de São Paulo, Escola de Comunicações e Artes.

Data da defesa: **09/08/2022**

Local: **Universidade Estadual Paulista**

Faculdade de Filosofia e Ciências

UNESP – *Campus* de Marília

ANÁLISE DE LOGS DE BUSCA COMO FONTE INFORMACIONAL PARA ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO EM REPOSITÓRIOS INSTITUCIONAIS

RESUMO

Os Repositórios Institucionais (RI) têm papel fundamental na divulgação do conhecimento, especialmente das publicações acadêmicas. Executados em diversas plataformas de software, com grande diversidade de sistemas de instalação, configuração e suporte, os RI remodelaram as formas de armazenamento, organização e recuperação de materiais, trazendo maior agilidade para todos esses processos. No entanto, disponibilizar recursos *on-line* para recuperação direta dos usuários é desafiador, na medida em que se faz necessário combinar o processo técnico realizado internamente nas bibliotecas com as necessidades dos usuários externos. Nessa perspectiva, a falta de articulação entre os vocabulários utilizados durante a indexação e os vocabulários utilizados pelos usuários tende a provocar ruídos e fracassos nos processos de busca e recuperação. Em vista disso, esta pesquisa teve como objetivo realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais. Para tanto, realizou-se pesquisa qualitativa e quantitativa e seu desenvolvimento se realizou em duas etapas: pesquisa exploratória e pesquisa descritiva. A pesquisa exploratória utilizou levantamento bibliográfico e análise documental e a pesquisa descritiva focalizou o trabalho com os dados coletados, para realizar análises as quais subsidiam o desenvolvimento de uma metodologia empírica para a análise de logs, baseada no modelo teórico de Jansen (2008). Assim, o processo descritivo e analítico resultou na proposta metodológica para atualização de vocabulário controlado por meio da análise de logs de busca, implementada na ferramenta computacional desenvolvida nesta Tese: a RILogUser (*Institutional Repository for User Log*), disponível em: <https://github.com/ikatahira/Doutorado>, uma ferramenta capaz de, a partir de um grande volume de dados, propiciar a seleção de dados relevantes e a comparação de vocabulários empregados por diferentes agentes de maneira produtiva. Entre as funcionalidades e características dessa ferramenta, destacam-se: (a) apresentação de uma visão sistemática do processo de análise de logs de buscas dos usuários; (b) possibilidade de integração com outras listagens de palavras, possibilitando a aplicação em RI diferentes; (c) apresentação de forma detalhada das etapas com exemplificações e seus respectivos produtos; (d) possibilidade de criação de vários vocabulários (baseado nos logs de buscas, palavras-chave dos autores e termos do Tesauro Unesp). A análise da lista de palavras oriundas dos logs de busca dos usuários, comparada à lista de palavras-chave atribuídas pelos autores aos documentos indexados no RI Unesp, viabilizou a obtenção de uma lista de expressões de busca contendo palavras simples, palavras compostas e frases, as quais compõem a lista de 5.826 palavras exclusivas às pesquisas dos usuários e 760 palavras comuns aos logs de pesquisas dos usuários e palavras-chave dos autores, ambas as listagens candidatas à atualização do vocabulário controlado

Tesouro Unesp. Dessa forma, a análise de logs se apresentou como fonte confiável de informações, as quais, gerenciadas por estratégias computacionais, têm amplo potencial para otimizar instrumentos de representação, a fim de beneficiar os processos de busca e, conseqüentemente, de recuperação de informações. Os resultados experimentais, obtidos a partir da metodologia desenvolvida, se mostrou eficaz por viabilizar uma grande depuração de expressões de busca de usuários e palavras-chave dos autores, que produziu um *corpus* terminológico consistente proveniente diretamente da linguagem utilizada no RI, como fonte informacional para atualização de vocabulário.

Palavras-chave: Análise de logs. Repositórios institucionais. Análise de log de transações. Vocabulário controlado. Logs de buscas.

ANALYSIS OF SEARCH LOGS AS A SOURCE OF INFORMATION FOR CONTROLLED VOCABULARY UPDATE IN REPOSITORY INSTITUTIONAL

ABSTRACT

Institutional Repositories (IR) play a fundamental role in the dissemination of knowledge, especially academic publications. Executed on several software platforms, with a great diversity of installation, configuration and support systems, the IR remodeled the ways of storing, organizing and recovering materials, bringing greater agility to all these processes. However, providing online resources for direct user retrieval is challenging, as it is necessary to combine the technical process carried out internally in libraries with the needs of external users. From this perspective, the lack of articulation between the vocabularies used during indexing and the vocabularies used by users tends to cause noise and failures in the search and retrieval processes. In view of this, this research aimed to carry out a theoretical and methodological study on the analysis of logs, their functions and their use to update controlled vocabularies for institutional repositories. The research is characterized as qualitative and quantitative and its development was carried out in two stages: exploratory research and descriptive research. The exploratory research used a bibliographic survey and document analysis and the descriptive research focused on the work with the collected data, to carry out quantitative and qualitative analyzes which subsidize the development of an empirical methodology for the analysis of logs. This descriptive and analytical process resulted in the methodological proposal for updating controlled vocabulary through the analysis of search logs, implemented in the computational tool developed in this Thesis: the RILogUser (Institutional Repository for User Log), based on the theoretical model of Jansen (2008). The descriptive and analytical process resulted in the methodological proposal for updating controlled vocabulary through the analysis of search logs, implemented in the computational tool developed in this Thesis: the RILogUser (Institutional Repository for User Log), available at: <https://github.com/ikatahira/Doutorado>, a tool capable of, from a large volume of data, providing the selection of relevant data and the comparison of vocabularies employed by different agents in a productive way. Among the features and characteristics of this tool, the following stand out: (a) presentation of a systematic view of the process of analyzing user search logs; (b) possibility of integration with other word lists, enabling application in different IRs; (c) detailed presentation of the steps with examples and their respective products; (d) possibility of creating several vocabularies (based on search logs, authors' keywords and terms from the Tesouro Unesp). The analysis of the list of words from the users' search logs, compared to the list of keywords attributed by the authors to the documents indexed in RI Unesp, made it possible to obtain a list of search expressions containing simple words, compound words and phrases, which make up the list of 5,826 words exclusive to user searches and 760 words common to user search logs and authors' keywords, both candidate lists for updating the controlled vocabulary Tesouro Unesp. The analysis of logs was presented as a reliable source of information, which, managed by computational strategies, have ample potential to optimize representation instruments, in order to benefit the search processes and, consequently, information retrieval. The experimental results, obtained from the methodology developed, proved to be effective as it enabled a great debugging of user search expressions and authors' keywords, which produced a consistent terminological corpus coming directly from the language used in the IR, as an informational source for vocabulary update.

Keywords: Log analysis. Institutional repositories. Transaction log analysis. Controlled vocabulary. Search logs.

ANÁLISIS DE REGISTROS DE BÚSQUEDA COMO FUENTE DE INFORMACIÓN PARA ACTUALIZACIÓN DE VOCABULARIO CONTROLADO EN REPOSITORIO INSTITUCIONAL

RESUMEN

Los Repositorios Institucionales (RI) juegan un papel fundamental en la difusión del conocimiento, especialmente de las publicaciones académicas. Ejecutado en varias plataformas de software, con una gran diversidad de sistemas de instalación, configuración y soporte, el IR remodeló las formas de almacenamiento, organización y recuperación de materiales, aportando mayor agilidad a todos estos procesos. Sin embargo, proporcionar recursos en línea para la recuperación directa del usuario es un desafío, ya que es necesario combinar el proceso técnico que se lleva a cabo internamente en las bibliotecas con las necesidades de los usuarios externos. Desde esta perspectiva, la falta de articulación entre los vocabularios utilizados durante la indexación y los vocabularios utilizados por los usuarios tiende a generar ruido y fallas en los procesos de búsqueda y recuperación. Ante ello, esta investigación tuvo como objetivo realizar un estudio teórico y metodológico sobre el análisis de bitácoras, sus funciones y su uso para la actualización de vocabularios controlados para repositorios institucionales. La investigación se caracteriza por ser cualitativa y cuantitativa y su desarrollo se llevó a cabo en dos etapas: investigación exploratoria e investigación descriptiva. La investigación exploratoria utilizó un levantamiento bibliográfico y análisis de documentos y la investigación descriptiva se centró en el trabajo con los datos recolectados, para realizar análisis cuantitativos y cualitativos que subsidian el desarrollo de una metodología empírica para el análisis de bitácoras. Este proceso descriptivo y analítico dio como resultado la propuesta metodológica de actualización de vocabulario controlado a través del análisis de registros de búsqueda, implementada en la herramienta computacional desarrollada en esta Tesis: el RILogUser (Repositorio Institucional de Registro de Usuario), basada en el modelo teórico de Jansen (2008). El proceso descriptivo y analítico dio como resultado la propuesta metodológica de actualización de vocabulario controlado a través del análisis de registros de búsqueda, implementada en la herramienta computacional desarrollada en esta Tesis: el RILogUser (Repositorio Institucional de Registro de Usuarios), disponible en: <https://github.com/ikatahira/Doutorado>, una herramienta capaz de, a partir de un gran volumen de datos, proporcionar la selección de datos relevantes y la comparación de vocabularios empleados por diferentes agentes de manera productiva. Entre las funcionalidades y características de esta herramienta se destacan: (a) presentación de una visión sistemática del proceso de análisis de los registros de búsqueda de los usuarios; (b) posibilidad de integración con otras listas de palabras, posibilitando su aplicación en diferentes RI; (c) presentación detallada de los pasos con ejemplos y sus respectivos productos; (d) posibilidad de crear varios vocabularios (basados en registros de búsqueda, palabras clave de los autores y términos del Tesauro Unesp). El análisis de la lista de palabras de los registros de búsqueda de los usuarios, comparada con la lista de palabras clave atribuidas por los autores a los documentos indexados en RI Unesp, permitió obtener una lista de expresiones de búsqueda que contienen palabras simples, palabras compuestas y frases. , que componen la lista de 5.826 palabras exclusivas de búsqueda de usuarios y 760 palabras comunes a los registros de búsqueda de usuarios y palabras clave de los autores, ambas listas candidatas para la actualización del vocabulario controlado Tesauro Unesp. El análisis de bitácoras se presentó como una fuente confiable de información, las cuales, manejadas por estrategias computacionales, tienen amplio potencial para optimizar instrumentos de representación, en beneficio de los procesos de búsqueda y, consecuentemente, de recuperación de información. Los resultados experimentales, obtenidos de la metodología desarrollada, demostraron ser efectivos ya que permitieron una gran depuración de las expresiones de búsqueda de los usuarios y las palabras clave de los autores, lo que produjo un corpus terminológico consistente proveniente directamente del lenguaje utilizado en el RI, como fuente de información para actualización de vocabulario.

Palabras clave: Análisis de registro. Repositorios institucionales. Análisis de registro de transacciones. Vocabulario controlado. Registros de búsqueda.

LISTA DE ILUSTRAÇÕES

Figura 1. Publicações selecionadas quanto aos vínculos por subáreas de estudo	32
Figura 2. Tipos de documentos disponibilizados no RI Unesp	64
Figura 3. Metadados do RI Unesp	65
Figura 4. Processo técnico documental de entrada e saída de informações	67
Figura 5. Busca por termos no Tesauro Unesp	76
Figura 6. Caracterização dos termos do Tesauro Unesp	76
Figura 7. Formatos dos metadados do RI Unesp	77
Figura 8. Modelo de recuperação da informação	78
Figura 9. Exemplo de potenciais perdas na recuperação por divergências de grafia	82
Figura 10. Página inicial do RI Unesp	84
Figura 11. Página de busca avançada no RI Unesp	85
Figura 12. Fluxo documental no RI Unesp: entrada, armazenamento e recuperação	90
Figura 13. Duplicidade de palavras-chave dos autores no RI Unesp	91
Figura 14. Seleção de estudos para revisão de literatura	95
Figura 15. Análise de coocorrência de palavras-chave dos 39 documentos selecionados	95
Figura 16. Etapas para análise dos logs de busca extraídos do RI Unesp	101
Figura 17. Coleta de logs extraídos contendo <i>id</i>	102
Figura 18. Código para seleção de dados dos logs	102
Figura 19. Código para edição de critérios de exclusão	103
Figura 20. Exclusão de logs conforme critérios estabelecidos	103
Figura 21. Código para exclusão de palavras duplicadas nos logs	104
Figura 22. Exclusão de palavras duplicadas	105
Figura 23. Fragmento do código do <i>web crawler</i> para extração de palavras-chave	105
Figura 24. Coleta de palavras-chave no RI Unesp	106
Figura 25. Exemplos da localização de erros na lista por tema	108
Figura 26. Fragmento do código para extração de termos, via API, do Tesauro Unesp	108
Figura 27. Coleta de termos do Tesauro Unesp	109
Figura 28. Fragmento do código do módulo desenvolvido para comparação das listagens	112
Figura 29. Página da RILogUser para extração de produtos: módulo <i>on-line</i>	112
Figura 30. Página da RILogUser para extração de produtos: módulo <i>off-line</i>	113
Figura 31. Diagrama das listas geradas com representação dos vocabulários	114
Figura 32. Relação entre etapas metodológicas de TLA e apresentação dos resultados	115
Figura 33. Principais bibliotecas utilizadas para manipulação das listagens	116

Figura 34. Bibliotecas utilizadas para a coleta de dados	117
Figura 35. Resultados das listagens de palavras por fonte de coleta de dados	118
Figura 36. Quantitativos da 1ª etapa de seleção dos logs de busca dos usuários	118
Figura 37. Fragmento dos códigos responsáveis pela limpeza dos logs de busca	121
Figura 38. Exemplos de tratamento das duplicações de palavras-chave dos autores	123
Figura 39. Fragmento do código para eliminar duplicações	124
Figura 40. Fragmento do código para identificação de coocorrência de palavras-chave dos autores	125
Figura 41. Fragmento do código para acesso à API do Tesouro Unesp e coleta dos termos	129
Figura 42. Código para identificação dos vocabulários	130
Figura 43. Diagrama de produtos das comparações e identificação de conjuntos comuns	130
Figura 44. Tela inicial da ferramenta RILogUser	133
Figura 45. Tela secundária para carregamento e análise de listas de vocabulários	134
Figura 46. Página de resultados para palavras exclusivas aos logs	135
Figura 47. Código para carregamento de registros	136
Figura 48. Código para carregamento de registros de modo comparativo	136
Figura 49. Código para tratamento de acentos e conversão das letras para maiúsculas	137
Figura 50. Exemplo de código para composição do vocabulário coincidente	137
Figura 51. Código para composição dos vocabulários em ordem alfabética	138
Figura 52. Código para composição das intersecções dos vocabulários	138
Figura 53. Código para referências dos objetos nas classes dos vocabulários	139
Figura 54. Código para exclusão das cidades brasileiras dos vocabulários	139
Figura 55. Tela de <i>download</i> do vocabulário selecionado	140
Figura 56. Tela para limpeza e depuração dos dados	140
Figura 57. Códigos para limpeza e depuração dos dados	141
Figura 58. Códigos para retirada de caracteres inválidos	141
Figura 59. Métricas das listagens dos vocabulários	142
Figura 60. Limpeza de palavras inválidas dos vocabulários	142
Figura 61. Processo de verificação textual das palavras dos vocabulários	143
Figura 62. Resultado do processo de preparação dos dados das listagens dos vocabulários	144
Figura 63. Verificação dos idiomas nos vocabulários criados	145
Figura 64. Seleção dos idiomas nos vocabulários criados	146
Figura 65. Criação dos arquivos separados por idiomas dos vocabulários	146
Figura 66. Resultados completos obtidos pela RILogUser em processamento <i>off-line</i>	147

LISTA DE QUADROS

Quadro 1. Organização estrutural da Tese	26
Quadro 2. Sistematização dos trabalhos selecionados	29
Quadro 3. Tipos de vocabulários controlados	72
Quadro 4. Exemplo de armazenagem dos logs de buscas dos usuários no RI Unesp	87
Quadro 5. Estratégias de busca para revisão de literatura	94
Quadro 6. Exemplo de erros detectados na listagem de palavras-chave dos autores	106
Quadro 7. Inconsistências de palavras-chave na recuperação de documentos	107
Quadro 8. Exemplos de depuração de logs de busca dos usuários	119

LISTA DE TABELAS

Tabela 1. Quantitativos da 2ª etapa de seleção dos logs de busca dos usuários	119
Tabela 2. Quantitativos da 3ª etapa de limpeza com aplicação de rotinas automatizadas para depuração de logs de pesquisa	120
Tabela 3. Coocorrência de palavras-chave dos autores	125
Tabela 4. Palavras-chave dos autores que recuperaram mais de 1000 documentos	126
Tabela 5. Listagem dos termos do Tesauro Unesp	128
Tabela 6. Quantitativos dos vocabulários das listagens	129

LISTA DE ABREVIATURAS E SIGLAS

AFDS	<i>Automated Forensic Diagnosis System</i>
Alice	Acesso Livre à Informação Científica da Embrapa
API	<i>Application Programming Interface</i>
BDH	Biblioteca Digital Hispânica
C2C	<i>Consumer-to-Consumer</i>
CGB-Unesp	Coordenadoria Geral de Bibliotecas da Unesp
CPT	Teste de Desempenho Controlado
Cruesp	Conselho de Reitores das Universidades Estaduais Paulistas
CSBIB	Bibliografias de Ciência da Computação de Karlsruhe
CSTR	Relatórios Técnicos de Ciência da Computação da Biblioteca Digital da Nova Zelândia
CSV	Valores Separados por Vírgulas
DNS	<i>Domain Name System</i>
GRI-Unesp	Grupo Gestor da Política do Repositório Institucional Unesp
HEC	Biblioteca Digital Comissão de Educação Superior
HTML	<i>HyperText Markup Language</i>
IBGE	Instituto Brasileiro de Geografia e Estatística
IEDCYT	Instituto de Estudos Documentários em Ciência e Tecnologia
IP	<i>Internet Protocol</i>
ISO	<i>International Organization for Standardization</i>
JSP	<i>Java Server Pages</i>
LCARB	Lista de Cabeçalhos de Assuntos da Rede Bibliodata
LCSH	<i>Library Congress Subject Headings</i>
LGPD	Lei Geral de Proteção de Dados Pessoais
MARC	<i>Machine Readable Cataloging</i>
MeSH	<i>Medical Subject Headings</i>
NDSL	<i>National Science and Technology Information Center</i>
NISO	<i>National Information Standards Organization</i>
NLTK	<i>Natural Language Toolkit</i>
OA	<i>Open Access</i>
OAI	<i>Open Archives Initiative</i>
OKAPI	Catálogo experimental executado em um site da Biblioteca Britânica

OPAC	Catálogo de Acesso Público <i>On-line</i>
PLN	Processamento de Linguagem Natural
PPGCI	Programa de Pós-Graduação em Ciência da Informação
RDF	<i>Resource Description Framework</i>
RI	Repositório Institucional
RILM	<i>Répertoire International de Littérature Musicale</i>
RVBI	Rede Virtual de Bibliotecas
SciELO	<i>Scientific Eletronic Library Online</i>
TBN	Terminologia da Biblioteca Nacional
TLA	<i>Transaction Log Analysis</i>
UET	Universidade de Engenharia e Tecnologia
UFRN	Universidade Federal do Rio Grande do Norte
Unesp	Universidade Estadual Paulista Júlio de Mesquita Filho
Unilab	Repositório da Universidade da Integração Internacional da Lusofonia Afro- Brasileira
URI	<i>Uniform Resource Identifier</i>
URL	<i>Uniform Resource Locator</i>
VCB	Vocabulário Controlado Básico
XML	<i>Extensible Markup Language</i>
WOS	<i>Web of Science</i>

SUMÁRIO

1 INTRODUÇÃO	17
1.1 Delimitação do problema e Tese	23
1.1.1 Objetivos	24
1.1.2 Justificativa	25
1.2 Sistemática da Tese	26
2 REVISÃO DA LITERATURA	29
2.1 Histórico e panorama da aplicação da TLA: artigos de revisão	33
2.2 TLA aplicada a catálogos de acesso público <i>on-line</i> em bibliotecas	38
2.3 TLA aplicada a sites comerciais e buscadores da <i>web</i>	45
2.4 Modelos propostos de TLA e níveis de análise	47
2.5 Aplicação da TLA para vocabulários controlados	54
2.6 Síntese dos resultados da revisão da literatura	56
3 FUNDAMENTOS TEÓRICOS	58
3.1 Constituição dos Repositórios Institucionais	58
3.1.1 Estrutura e funcionamento do RI Unesp	61
3.2 Indexação	66
3.2.1 Indexação nos Repositórios Institucionais brasileiros	69
3.3 Vocabulário controlado	71
3.3.1 Tesouro Unesp	74
3.4 Sistemas de recuperação de informação	77
3.4.1 Estratégias de busca	79
3.4.2 Recuperação de informações no RI Unesp	81
3.4.3 Estratégias de busca no RI Unesp	84
3.5 Logs de busca de usuários e Processamento de Linguagem Natural	86
3.5.1 Logs de buscas de usuários no RI Unesp	86
3.5.2 Processamento de Linguagem Natural	88

3.6 Síntese dos resultados dos fundamentos teóricos	89
4 METODOLOGIA	93
4.1 Caracterização da pesquisa	93
4.1.1 Pesquisa exploratória com pesquisa bibliográfica e documental	93
4.1.2 Pesquisa descritiva	97
5 PROPOSTA PARA ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO POR MEIO DA ANÁLISE DE LOGS DE BUSCA	100
5.1 Coleta de dados	109
5.2 Preparação de dados	111
5.3 Análise de dados	111
6 FERRAMENTA DESENVOLVIDA PARA ANÁLISE DE VOCABULÁRIOS E RESULTADOS	115
6.1 RILogUser: módulo <i>off-line</i>	116
6.1.1 Palavras de buscas dos usuários extraídos dos logs	118
6.1.2 Palavras-chave empregadas pelos autores	122
6.1.3 Lista de termos do Tesouro Unesp	128
6.2 RILogUser: módulo <i>on-line</i>	133
7 CONSIDERAÇÕES FINAIS	148
7.1 Limitações da pesquisa	150
7.2 Trabalhos futuros	150
REFERÊNCIAS	152

1 INTRODUÇÃO

Esta Tese tem como temática a otimização da representação da informação pela coleta, preparação e análise de logs, a fim de propor a atualização de vocabulários controlados de repositórios institucionais, elaboradas a partir das palavras buscadas pelos usuários e palavras-chave de autores. Como experimento é verificado o caso do Repositório Institucional Unesp e como parâmetro de vocabulário controlado, adotar-se-á o Tesauro Unesp. Desse modo, por abranger práticas computacionais, insere-se na área da Ciência da Informação, mais especificamente na linha de Produção e Organização da Informação.

Os repositórios institucionais (RI) têm papel fundamental na divulgação do conhecimento, especialmente das publicações acadêmicas. Executados em diversas plataformas de software, com grande diversidade de sistemas de instalação, configuração e suporte, os RI remodelaram as formas de armazenamento, organização e recuperação de materiais, trazendo maior agilidade para todos esses processos. Todavia, para se efetivarem como espaços interativos e disseminadores do conhecimento produzido pelas instituições que representam e não meros depósitos, devem aprimorar permanentemente as formas de acesso e meios de busca disponibilizados aos potenciais usuários.

Por sua importância crescente, os RI, além de se constituírem como fonte essencial de informações na contemporaneidade, também se tornaram objeto de estudo, posto que viabilizam não apenas a realização de pesquisas, mas também o conhecimento de todo o percurso de busca realizado pelo usuário. Nesse sentido, o rastreamento de *downloads* de arquivos como uma métrica do sucesso do RI, por exemplo, pode auxiliar de maneira significativa ao fornecer dados consistentes e precisos para verificação do tráfego do usuário (OBRIEN, 2017).

Os RI têm ganhado destaque igualmente pela diversidade de materiais disponibilizados. Em um só lugar, o pesquisador pode encontrar desde gêneros tradicionais como resumos, resenhas, artigos, dissertações e teses, até *podcasts*, *data papers* e planos de gestão de dados (UNESP, 2020). Martínez-Comeche (2017) destaca que o acesso remoto a um crescente número de coleções digitalizadas e aos metadados incorporados a seus documentos, favorece em grande medida a difusão de seus recursos. No entanto, disponibilizar recursos *on-line* para recuperação direta dos usuários é desafiador, na medida em que se faz necessário combinar o processo técnico realizado internamente nas bibliotecas com as necessidades dos usuários externos.

Nesse sentido, Fujita (2017, p. 1) afirma que “a falta de controle de vocabulário em sistemas de armazenamento, tratamento e recuperação de recursos de informação gera vários tipos de inconsistências determinantes para a imprecisão e irrelevância em qualquer sistema de busca”. Nessa perspectiva, a falta de articulação entre os vocabulários utilizados, durante a indexação e a recuperação, tende a provocar ruídos e fracassos nos processos de busca, “[...] principalmente, a recuperação de grande quantidade de itens sem compatibilidade com o assunto desejado” (FUJITA, 2017, p. 2).

De acordo com a Norma ANSI/NISO (*National Information Standards Organization*) Z39.19-2005 (R2010) (NISO, 2010), a indexação é considerada um processo de atribuição de termos para descrever os conceitos de um documento, envolvendo termos de um ou mais vocabulários controlados ou outras fontes utilizadas para descrever o assunto. O controle de vocabulário se relaciona a um processo pautado nos princípios teórico-metodológicos da terminologia, buscando eliminar ambiguidades. Conforme Smit (2018, p. 52), o controle de vocabulário requer “[...] tanto padronizar a forma gramatical dos termos quanto identificar termos sinônimos ou quase-sinônimos e antônimos”.

A proposição de padronização é uma tentativa de minimizar a incompatibilidade entre linguagem indexada e linguagem buscada, uma vez que os gestores dos repositórios veem a necessidade de disponibilizar aos usuários várias ferramentas que os ajudem em suas buscas, de modo a aprimorar os mecanismos e resultados de pesquisa. Porém, para cumprir essa tarefa com êxito, é fundamental entender como os usuários agem na interação com os RI, pois, assim, as atualizações dos sistemas serão capazes de eliminar ou minimizar os obstáculos que os usuários encontram ao tentar acessar o material desejado, e poderão melhorar a operação e a eficiência dos sistemas de recuperação.

A interatividade é uma das principais características a serem consideradas para o sucesso de um RI. Nessa perspectiva, o desenvolvimento desta investigação justifica-se diante da importância de garantir que os usuários tenham êxito em suas buscas, posto que o processo comunicativo só se completa quando a mensagem produzida pelo autor chega ao leitor.

Relembrando os seis elementos básicos para que a comunicação se efetive (JAKOBSON, 2010), sabemos que é preciso um emissor disposto a produzir uma mensagem direcionada a um receptor, utilizando um código comum a ambos, em um canal viável, em um contexto/referente compartilhado. Sendo assim, podemos compreender os RI como um grande canal que necessita ser constantemente testado, sob o risco de que muitas comunicações acabem silenciadas pelo esquecimento oriundo da falta de acesso.

Na Ciência da Informação, para que esses testes sejam eficazes, *a priori*, duas instâncias precisam ser analisadas, a saber: o autor da indexação e o autor da busca. A viabilidade do canal será proporcional ao grau de compatibilidade desses dois agentes, pois, se o código utilizado por eles for demasiadamente diverso, a comunicação não se efetivará e o RI não cumprirá plenamente seu papel de difusor do conhecimento.

Nesse sentido, os caminhos percorridos pelos indexadores são mais previsíveis que os caminhos traçados pelos usuários, haja vista, inclusive, as diversas ferramentas já desenvolvidas para a indexação automática de materiais, considerando padrões estruturais, vocabulários controlados, tesouros e ontologias. A indexação, como procedimento técnico, é mais previsível por adotar padrões, ao passo que, o comportamento do usuário, tende a ser mais instintivo, por ser orientado pela subjetividade e repertório intelectual individual (FUJITA, ALMEIDA, 2020).

Com a compreensão das variáveis supracitadas, a *Transaction Log Analysis* (TLA) (AVERY; TRACY, 2014; BLECIC *et al.*, 1998; HERRERÍAS; GÓMEZ, 2010; LAIPELT, 2015a; VÁLLEZ *et al.*, 2015; WU; FAN, 2018), análise de logs de transações dos usuários, método adotado nesta Tese, torna-se uma ferramenta promissora, visto que permite ao gestor conhecer diretamente o percurso investigativo feito pelo usuário, a fim de propor melhorias para suas experiências de pesquisa.

A experiência do usuário tem sido tema de estudos em diversas áreas: no meio empresarial (FREITAS; FREGADOLLI; FREITAS, 2020), educacional (LISBOA; BARROSO; ROCHA, 2020), ambiental (BRITO; SOUSA, 2018), saúde (WHITE; HORVITZ, 2014) entre outros. Sendo assim, os recursos disponibilizados pelos meios digitais, nomeadamente os registros de logs, fornecem as “pegadas” necessárias para que possamos refazer, como pesquisadores, os caminhos dos usuários e, assim, garantir que esses caminhos não terminem sem chegar ao destino, isto é, sem encontrar os materiais desejados.

Conforme Jansen (2006, p. 409, tradução nossa), os registros de transações obtidos pela coleta de logs constituem:

[...] um método discreto de coletar quantidades significativas de dados de pesquisa em um número considerável de usuários do sistema. No entanto, tem havido um número limitado de pesquisadores que exploraram a metodologia de log de transações para estudar a pesquisa na *web*.

Mari Vállez *et al.* (2015) utilizam a análise de logs para propor um modelo semiautomático para atualização de vocabulários controlados, de modo a viabilizar um

comparativo entre as palavras-chave utilizadas pelos usuários durante o processo de busca e o vocabulário utilizado no processo de indexação para atualização deste. Os autores destacam a viabilidade da proposta em ambientes de rede controlada como repositórios, além de ressaltarem que, ao levar em consideração a perspectiva dos usuários, é possível propor termos para inclusão em vocabulários controlados, de forma a otimizar o acesso às informações.

No Brasil, Laipelt (2015b) propôs a análise de lexias contidas nos logs para identificar a linguagem do usuário e contribuir para a atualização do tesauro e do catálogo de autoridades no Portal LexMI, da área do Direito do Trabalho e Previdenciário. Nesse sentido, os logs são utilizados como pegadas digitais dos usuários no sistema para proporcionar coletas de termos e descritores, de modo a automatizar o tesauro e a gestão de controle de catálogo de autoridades como estratégia para otimizar a recuperação de informação pelo usuário.

Laipelt destacou que a análise de logs ainda era pouco utilizada pelos pesquisadores no Brasil, visto que, em levantamento realizado por ela em 2012, encontrara apenas o trabalho fruto de especialização realizada por Pontes (2006) e mais quatro trabalhos de conclusão de curso desenvolvidos por alunos do curso de Biblioteconomia da Universidade Federal do Rio Grande do Sul (LAIPELT, 2015b). Como causa para essa baixa utilização, a autora identifica a configuração de muitos softwares para exclusão automática dos logs, a fim de que não ocupem espaço nos servidores e, além disso, há ainda as dificuldades de acesso a tais registros, mesmo quando armazenados. Apesar disso, a autora defende a análise de logs como ferramenta de investigação direta das expressões utilizadas pelos usuários.

Na área da saúde, o uso dos logs é adotado por White e Horvitz (2014) para verificar o comportamento dos usuários em pesquisas médicas. Os autores utilizam duas metodologias complementares: a análise de logs de pesquisa para estudo do comportamento dos usuários e a aplicação de questionário fechado aos mesmos usuários, a fim de ampliar a compreensão das motivações associadas a tomadas de decisões de cuidados de saúde. Assim, os autores buscaram, com o questionário, confirmar e analisar tendências já observadas nos registros dos logs de busca.

Sobre as diferenças entre o vocabulário controlado utilizado para a indexação e a linguagem natural empregada pelos usuários durante as buscas, Martínez-Comeche (2017) expõe que o tratamento documental envolve uma organização e disposição dos recursos que normalmente diferem da estrutura do conhecimento que o usuário possui, podendo, às vezes, dificultar a tarefa de localizar e acessar a documentação desejada.

Cumprir destacar que a padronização terminológica facilitaria tanto a indexação quanto a recuperação do conhecimento; porém, a harmonização entre a linguagem do indexador, ou

mesmo dos materiais indexados, os vocabulários controlados e a linguagem do usuário não acontece de forma natural. Diante dessa dificuldade, como destacam Carlan e Medeiros (2011, p. 54), os vocabulários controlados são semanticamente estruturados para minimizar incompatibilidades de vocabulário, incluindo “[...] a eliminação de ambiguidades, o controle de sinônimos ou equivalentes e estabelecimento de relacionamentos semânticos entre conceitos [...]”, de modo a permitir uma representação documental adequada, tendo como critério, para validação, a eficiência para a recuperação.

Nesse cenário, para verificarmos se os vocabulários controlados dos RI estão sendo eficientes, a análise de logs se configura como estratégia válida por permitir o monitoramento da experiência de busca dos usuários. A propósito, Martínez-Comeche (2017) destaca a análise de arquivos de log como um procedimento para entender as estratégias de busca de usuários, pois registram as interações entre as pessoas e os sistemas de informação.

Ao falarmos em representação da informação, é imperativo destacarmos identicamente a importância dos vocabulários controlados, dada sua função na indexação e recuperação de documentos. Imagine um centro de trabalho com três indexadores que **não** utilizam um vocabulário controlado (tesauro, lista de descritores etc.) para indexar os documentos. O indexador 1, para um documento que trata de *pen drive*, insere-o ao sistema com o termo de indexação “*pen drive*”. O indexador 2, um mês depois, para outro documento de *pen drive*, atribui o termo de indexação “memória” e, um ano depois, o indexador 3, para um documento que trata de memória portátil, atribui o termo de indexação “usb”.

Assim, quando um usuário vai pesquisar o assunto no banco de dados onde trabalham os três indexadores supracitados, existem várias possibilidades: a) se o usuário usa o termo “*pen drive*”, ele recupera um documento; b) se o usuário usa o termo “memória”, ele recupera um documento e c) se o usuário usa o termo “usb”, ele recupera um documento. Porém, essa recuperação não é eficiente, já que há, na verdade, três documentos que tratam do mesmo assunto, todavia, a recuperação é parcial, conforme a correspondência entre o termo de busca e termo indexado, exigindo correspondência literal, sem consideração de relações semânticas de sinonímia. Dessa forma, o usuário acaba perdendo informação que seria de seu interesse.

Em um outro contexto de indexação, imagine um centro de trabalho com três indexadores que adotam um vocabulário controlado para indexar documentos e, posteriormente, para pesquisar no banco de dados. Em relação ao termo *pen drive*, o tesauro indica o seguinte:

PEN DRIVE
 pen drive USE *PEN DRIVE*

 memória USE *PEN DRIVE*
 memória portátil USE *PEN DRIVE*
 ...
 USB USE *PEN DRIVE*

O indexador 1, para um documento que trata sobre usb, de acordo com o tesauro, atribui o termo de indexação “*pen drive*”. Quando o indexador 2 analisa um documento que trata de um *pen drive*, de acordo com o tesauro, igualmente o atribui como o termo de indexação “*pen drive*”, assim como o indexador 3, que para outro documento em “memória portátil”, da mesma forma o atribui como “*pen drive*”. Quando um usuário precisa ler documentos que abordem sobre “*pen drive*”, o banco de dados oferece a possibilidade de utilizar o tesauro para fazer buscas. O usuário consulta o tesauro e descobre que deve usar o termo “*pen drive*”; o sistema oferece diretamente uma lista de termos em que aparece “*pen drive*”. Quando o usuário pesquisa com o termo “*pen drive*”, ele recupera três documentos.

Nesse exemplo simples, evidencia-se a importância da utilização de vocabulários controlados, tendo em vista que em um centro de trabalho com muitos indexadores, é necessário adotar vocabulários controlados, caso contrário a qualidade do sistema poderá ser comprometida por inconsistências na indexação dos documentos. A propósito, destacam-se quatro princípios importantes de controle de vocabulário que orientam seu *design* e desenvolvimento: I) eliminar ambiguidades; II) controlar sinônimos; III) estabelecer relações entre os termos, quando apropriado e IV) testar e validar termos.

Assim, pesquisas que ampliem a convergência entre a linguagem natural dos usuários e dos autores com os vocabulários controlados são importantes para viabilizarem o gerenciamento sustentável e a recuperação eficiente de documentos.

Como parâmetro comparativo para a lista de termos de busca dos usuários, conforme identificação nos logs, e lista de palavras-chave atribuídas pelos autores aos documentos indexados, utilizamos o Tesauro Unesp (FUJITA; MOREIRA, 2021), um vocabulário controlado que considera equivalência de termos, questões hierárquicas, de homografia e de associação entre termos, de modo que possam ser notados padrões de relacionamentos (NISO, 2010).

Frente à amplitude dos estudos relacionados à Recuperação da Informação e a necessidade eminente de controle de vocabulário, a seguir, apresenta-se a delimitação do

problema e a tese defendida nesta investigação.

1.1 Delimitação do problema e Tese

As reuniões do grupo de pesquisa “Representação temática da informação”, liderado pelas professoras Mariângela Spotti Lopes Fujita e Paula Regina Dal'Evedove, contribuíram para o desenvolvimento de reflexões sobre aspectos da recuperação de informações e as estratégias de busca de usuários, fato que colaborou com o propósito de compreender como os RI e o processo de indexação podem ser aperfeiçoados a partir da TLA.

Com base nas discussões feitas pelo grupo e na contextualização do problema, mediante fundamentação teórica exposta acima, esta pesquisa parte de três pressupostos principais pautados na interatividade; na atualização de vocabulário controlado e na análise de logs dos usuários.

Ao compreender a interatividade como princípio para a efetividade de um repositório, é possível utilizá-la como meio de mensurar o sucesso do sistema de recuperação adotado. Para pensar em recuperação da informação, deve-se dar um passo atrás para compreender a sua forma de organização e, com essa percepção, a necessidade de testes contínuos capazes de verificar o grau de compatibilidade da linguagem de organizadores (indexadores) e recuperadores (usuários) torna-se evidente.

Em adição, a partir da compreensão de que as formas de organizar o conhecimento precisam ser atualizadas para garantir a efetividade de pesquisas posteriores, há uma demanda crescente por vocabulários os quais se atualizem constantemente por meio de recursos computacionais capazes de acompanhar a velocidade das publicações e propiciar ambientes de recuperação precisa e intuitiva.

Sequencialmente, ao reconhecermos que a produção de materiais e as buscas crescentes por informações digitais exacerbam a necessidade de desenvolver vocabulários que usuários não especialistas possam utilizar de forma intuitiva, a análise de logs é uma forma de mapear as estratégias de busca de usuários dos sistemas de informação para subsidiar a revisão e atualização de vocabulários controlados, de modo a aprimorar a recuperação de informações pelos usuários.

A partir desses pressupostos, a relação entre as palavras de busca de assunto preenchidas pelos usuários e os metadados de assuntos preenchidos pelos autores ou sistemas de indexação automática precisa ser investigada. Nesse contexto, a análise de logs se apresenta como rica fonte de informações, as quais, gerenciadas por estratégias computacionais, podem

ser utilizadas para otimizar instrumentos de representação, a fim de também beneficiar os processos de busca e, consequentemente, de recuperação da informação.

Problema de pesquisa: como os logs resultantes das pesquisas feitas pelos usuários em um repositório institucional podem auxiliar na atualização de vocabulários controlados, no que concerne à padronização terminológica entre os processos de indexação e de busca de materiais?

A motivação da pesquisa está pautada no entendimento de que, para a efetivação do processo comunicativo em diferentes contextos, é requerido o compartilhamento do código, isto é, da língua utilizada para representar as ideias e os materiais elaborados. Com esse entendimento, é necessário considerar o percurso do material desde sua produção pelo autor até sua recuperação pelos potenciais leitores, o que nos leva ao reconhecimento da indexação como etapa fundamental para a representação documental, de modo que a recuperação será efetiva na medida em que a forma de indexação se aproximar, o máximo possível, da linguagem do usuário. Essa percepção motivou a investigação de um método capaz de monitorar as estratégias de busca de usuários de Repositórios Institucionais, de modo a propor melhorias na recuperação das informações desejadas. Nesse contexto, a extração de logs de buscas se configurara como estratégia viável e confiável para a compreensão das estratégias de busca de usuários.

A extração de logs identificados como base para a análise do vocabulário dos usuários de um repositório corrobora a utilização de ferramentas computacionais como instrumento cada vez mais comum e necessário para representação e recuperação de informações. Na literatura, vários autores já relataram suas opiniões sobre o emprego dessas ferramentas tanto na Ciência da Computação quanto em todas as áreas do conhecimento, posto que o aumento de dados produzidos diariamente requer aprimoramento veloz e contínuo dos métodos de processamento.

Desse modo, temos como **tese** de pesquisa a demonstração de que a análise de logs, a partir das palavras empregadas pelos usuários, pode viabilizar a **proposição** de atualizações de vocabulários, por exemplo, com a inclusão de novos descritores oriundos de palavras da linguagem natural empregada pelo usuário.

1.1.1 Objetivos

Realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais.

Para tanto, temos como objetivos específicos:

1. Estudar os logs de buscas de usuários do Repositório Institucional Unesp para proposta de atualização do vocabulário controlado Tesauro Unesp;

2. Analisar, a partir de logs de pesquisa feitas pelos usuários no RI Unesp, a viabilidade metodológica de aplicação de recursos computacionais para realização de rotinas automatizadas para depuração / limpeza de logs de pesquisa para composição de lista de palavras candidatas à inclusão no Tesouro Unesp, subsidiando metodologia específica de análise de logs;

3. Criar e investigar a aplicação experimental da análise de logs implementada por meio de ferramenta computacional desenvolvida durante a produção desta Tese, para a atualização de vocabulário controlado do Tesouro Unesp, com vistas à ampliação de seu uso no RI Unesp.

1.1.2 Justificativa

A proposta se justifica por demonstrar como recursos computacionais podem colaborar para a otimização dos processos de representação e recuperação de informações de modo a viabilizar avanços em diversos setores da sociedade. Para tanto, a investigação a partir de logs, além de permitir acesso a um grande volume de dados primários, por sua extensão e presença em qualquer sistema informatizado, tem potencial informativo exponencial.

No contexto dos repositórios institucionais, locais cada vez mais importantes para a divulgação do conhecimento, o estudo dos logs traz a originalidade e o valor do método proposto por viabilizar análises quali-quantitativas das palavras de busca dos usuários para auxiliar os gerentes de RI a determinar meios eficientes de promover a atualização de vocabulários controlados para representação e recuperação de documentos no RI Unesp.

Em perspectiva ampliada, a pesquisa apresenta relevância social ao contribuir para a evolução da ciência, uma vez que favorece o aprimoramento da organização e representação do conhecimento, a partir do estudo do vocabulário empregado por seu público-alvo, ou seja, os usuários que desejam consultar e recuperar materiais que atendam seus interesses de pesquisa.

Dado o exposto, é fundamental a busca por métodos que permitam acompanhar as estratégias de busca dos usuários para que seja possível aprimorar suas experiências no RI, de modo a garantir o acesso e a usabilidade das informações e conhecimentos produzidos e armazenados nas instituições de ensino e pesquisa.

Cumprido destacar que a compreensão das aplicações da tecnologia vem ao encontro da necessidade de aprimorar a recuperação do conhecimento produzido, posto que mesmo com a ampliação dos meios de divulgação acadêmica, nem sempre encontrar materiais relevantes é tarefa simples.

Justifica-se a proposição desta investigação, portanto, pela necessidade de buscar recursos que colaborem para o conhecimento do vocabulário usado pelos usuários no

Repositório Institucional Unesp, 5º maior repositório brasileiro e 24º maior repositório mundial¹. A partir dos dados extraídos nos logs, é possível analisar as estratégias de busca de usuários e propor melhorias para garantir a recuperação das informações intencionadas.

1.2 Sistemática da Tese

De modo a contribuir para a visualização da organização desta Tese, elaboramos o Quadro 1 com os tópicos estruturantes do desenvolvimento teórico e metodológico da pesquisa.

Quadro 1. Organização estrutural da Tese

DADOS INTRODUTÓRIOS
Problema: Como os logs resultantes das pesquisas feitas pelos usuários em um repositório institucional podem auxiliar na atualização de vocabulários controlados no que concerne à padronização terminológica entre os processos de indexação e de busca de materiais?
Tese: A análise de logs de buscas feitas por usuários pode viabilizar atualizações de vocabulários controlados que otimizem a recuperação de informações.
Proposição: Propomos realizar a coleta, preparação e análise de logs de busca dos usuários, de modo a otimizar a representação da informação pela utilização de estratégias computacionais capazes de colaborar para atualizações de vocabulários controlados a partir de palavras de busca empregadas pelos usuários.
Objetivo geral: Realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais.
ESTRUTURA E DESENVOLVIMENTO
1 INTRODUÇÃO
Objetivo específico 1: Estudar os logs de buscas de usuários do Repositório Institucional Unesp para proposta de atualização do vocabulário controlado Tesouro Unesp. Desenvolvido nas seções 2 e 3: REVISÃO DA LITERATURA e FUNDAMENTOS TEÓRICOS
Objetivo específico 2: Analisar, a partir de logs de pesquisa feitas pelos usuários no RI Unesp, a viabilidade metodológica de aplicação de recursos computacionais para realização de rotinas automatizadas para depuração / limpeza de logs de pesquisa para composição de lista de palavras candidatas à inclusão no Tesouro Unesp, subsidiando metodologia específica de análise de logs. Desenvolvido na seção 4: METODOLOGIA
Objetivo específico 3: Criar e investigar a aplicação experimental da análise de logs implementada por meio de ferramenta computacional desenvolvida durante a produção desta Tese para a atualização de vocabulário controlado do Tesouro Unesp com vistas à ampliação de seu uso no RI Unesp.

¹ Disponível em: <https://repositories.webometrics.info/en/transparent>. Acesso em: 17 fev. 2022.

Apresentado nas seções 5 e 6: ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO POR MEIO DA ANÁLISE DE LOGS DE BUSCA e FERRAMENTA DESENVOLVIDA PARA ANÁLISE DE VOCABULÁRIOS E RESULTADOS

7 CONSIDERAÇÕES FINAIS

Fonte: elaborado pelo autor.

Conforme síntese apresentada no Quadro 1, o presente trabalho está organizado em sete seções. Na primeira, há a introdução contextualizada que apresenta os objetivos, justificativa e a relevância da pesquisa, assim como os métodos escolhidos para o desenvolvimento.

Na segunda seção, descrevemos os resultados da revisão de literatura realizada nas bases *Scopus* e *Web of Science* (WOS) sobre as aplicações da TLA em sistemas de catálogo de acesso público *on-line*, em bibliotecas, sites comerciais, além de pesquisas de revisão, proposição de modelos de TLA e níveis de análise e a TLA como fonte de informações para a atualização de vocabulários.

Na terceira, são apresentados os fundamentos teóricos que subsidiam a investigação. Para tanto, é descrita a forma de constituição dos RI, com destaque para suas características estruturais e de funcionamento. Sequencialmente, descrevemos fundamentação relacionada à indexação e aos vocabulários controlados, bem como características fundamentais dos sistemas de recuperação de informação e dos logs de buscas de usuários com auxílio de ferramentas para o processamento de linguagem natural.

Na quarta seção, é descrita a caracterização metodológica da pesquisa dividida em duas etapas: pesquisa exploratória e pesquisa descritiva. A pesquisa exploratória pautada no levantamento bibliográfico e na análise documental e a pesquisa descritiva com foco nos dados coletados para realizar análises quantitativas e qualitativas, as quais subsidiam o desenvolvimento de uma metodologia empírica para a análise de logs.

Na quinta seção, é descrita a metodologia empregada para a atualização de vocabulário controlado por meio da análise de logs de busca, a qual organiza-se a partir das três etapas definidas pelo modelo teórico proposto por Jansen (2008): coleta, preparação e análise de logs e estrutura a ferramenta RILogUser, desenvolvida durante esta Tese para auxiliar no cumprimento do objetivo geral da pesquisa.

Na sexta seção, por fim, é descrita a ferramenta RILogUser, construída a fim de possibilitar a análise dos logs dos usuários do RI Unesp. São expostos os resultados obtidos pela análise dos logs de buscas dos usuários, além dos resultados da lista de palavras-chave

empregadas pelos autores para representação de suas publicações e da lista de termos do Tesouro Unesp, destacando palavras exclusivas e palavras comuns como candidatas à atualização do vocabulário controlado que compõe o Tesouro Unesp.

Por fim, na sétima seção, concluímos o texto com considerações sobre os resultados da pesquisa e reflexões para proposição de trabalhos futuros, os quais – a partir do reconhecimento dos logs como objeto de estudo e fonte informacional do vocabulário de busca dos usuários – possam colaborar para o processo de recuperação de materiais publicados em diferentes bases de dados. Sequencialmente, apresentamos a lista de referências citadas ao longo desta Tese.

2 REVISÃO DA LITERATURA

No intuito de contextualizar a aplicação da TLA como estratégia de recuperação das palavras de busca dos usuários, é apresentada, nesta seção, a revisão da literatura feita por meio da identificação e compilação sistemática das principais referências sobre o tema. Destacamos que os estudos teóricos e metodológicos da TLA, suas funções e sua utilização para atualização ou criação de vocabulários controlados para RI, embora não sejam recentes, ainda têm muito a crescer.

Nesse sentido, é sistematizada no Quadro 2 a listagem dos títulos recuperados na revisão de literatura, conforme critérios que serão descritos na seção 4, subitem 4.1.1, além dos atributos: *Id* (identificação do trabalho); autor; área e ano de publicação:

Quadro 2. Sistematização dos trabalhos selecionados

<i>Id</i>	Título	Autor	Área	Ano
1	<i>Evaluation of user reactions to a prototype on-line information retrieval system</i>	Meister, D., Sullivan, D.J.	Engenharia	1967
2	<i>Transaction log analysis online catalogs</i>	Tolle, J. E.	Ciência da Computação	1983
3	<i>The use of computer-monitored data in information science and communication research</i>	Rice, R.E., Borgman, C.L.	Engenharia	1983
4	<i>Improving Okapi: Transaction Log Analysis</i>	Jones, R.	Ciência da Computação; Ciência Sociais.	1986
5	<i>Monitoring user search success through transaction log analysis: The Wolf PAC example</i>	Zink, S.D.	Ciência Sociais	1991
6	<i>Applying the results of transaction log analysis</i>	Sandore, B.	Ciência da Computação; Ciência Sociais	1993
7	<i>Research Methodologies and Transaction Log Analysis: Issues, Questions, and a Proposed Model</i>	Kaske, N.K.	Ciência da Computação; Ciência Sociais	1993
8	<i>The history and development of transaction log analysis</i>	Peters, T.A.	Ciência da Computação; Ciência Sociais	1993
9	<i>A manifesto: Regarding the future of transaction log analysis</i>	Sandore, B., Flaherty, P., Kaske, N. K., Kurth, M., Peters, T.	Ciência da Computação; Ciência Sociais	1993

10	<i>The limits and limitations of transaction log analysis</i>	Kurth, M.	Ciência da Computação; Ciência Sociais	1993
11	<i>Using transaction log analysis to improve OPAC retrieval results</i>	Blecic, D.D., Bangalore, N.S., Dorsch, J.L., Handerson, C.L., Koenig, M.H., Weller, A.C.	Ciência Sociais	1998
12	<i>A transaction log analysis of a digital library</i>	Jones, S., Cunningham, S.J., McNab, R., Boddie, S.	Ciência Sociais	2000
13	<i>A comparative transaction log analysis of two computing collections</i>	Mahoui, M., Cunningham, S.J.	Ciência da Computação; Matemática	2000
14	<i>Users' perceptions of the Web as revealed by transaction log analysis</i>	Moukdad, H., Large, A.	Ciência da Computação; Ciência Sociais	2001
15	<i>Search behavior in a research-oriented digital library</i>	Mahoui, M., Cunningham, S.J.	Ciência da Computação; Matemática	2001
16	<i>Search log analysis: What it is, what's been done, how to do it</i>	Jansen, B.J.	Ciência da Computação; Ciência Sociais	2006
17	<i>Query log analysis in biomedicine</i>	Bernstam, E.V., Herskovic, J.R., Hersh, W.R.	Ciência da Computação	2008
18	<i>Handbook of research on web log analysis</i>	Jansen, B.J., Spink, A., Taksa, I.	Ciência da Computação	2008
19	<i>The methodology of search Log Analysis</i>	Jansen, B.J.	Ciência da Computação	2008
20	<i>Web log analysis: Diversity of research methodologies</i>	Taksa, I., Spink, A., Jansen, B.J.	Ciência da Computação	2008
21	<i>Research and methodological foundations of transaction log analysis</i>	Jansen, B.J., Taksa, I., Spink, A.	Ciência da Computação	2008
22	<i>Historic perspective of log analysis</i>	Penniman, W.D.	Ciência da Computação	2008
23	<i>Topic analysis and identification of queries</i>	Ozmutlu, S., Ozmutlu, H.C., Spink, A.	Ciência da Computação	2008
24	<i>Exploration of interdisciplinarity in nanotechnology queries</i>	Shiri, A.	Ciência da Computação; Ciência da Decisão; Matemática; Ciências Sociais	2009
25	<i>Log analysis towards an Automated Forensic Diagnosis System</i>	Herrerías, J., Gómez, R.	Ciência da Computação; Engenharia	2010
26	<i>Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis</i>	Huurnink, B., Hollink, L., Van Heuvel, W.D., De Rijke, M.	Ciência da Computação	2010

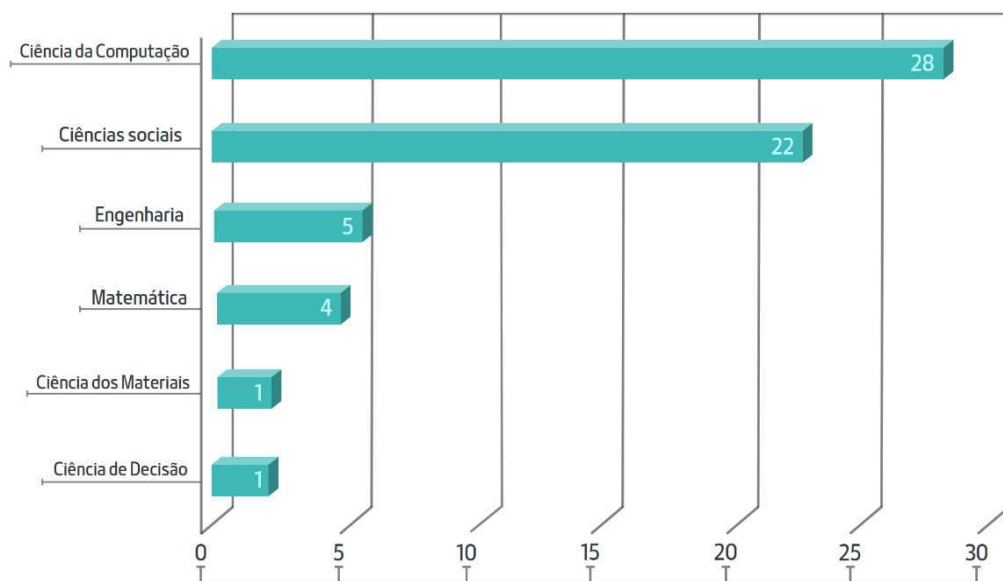
27	<i>Understanding science and technology information users through transaction log analysis</i>	Park, M., Lee, T.	Ciência da Computação; Ciências Sociais	2013
28	<i>Using transaction log analysis to assess student search behavior in the library instruction classroom</i>	Avery, S., Tracy, D.G.	Ciências Sociais	2014
29	<i>MR-CLOPE: A MapReduce based transactional clustering algorithm for DNS query log analysis</i>	Li, Y.-F., Le, J.-J., Wang, M., Zhang, B., Liu, L.-X.	Ciência dos Materiais; Engenharia	2015
30	<i>Updating controlled vocabularies by analysing query logs</i>	Vállez, M., Pedraza-Jiménez, R., Codina, L., Blanco, S., Rovira, C.	Ciência da Computação; Ciências Sociais	2015
31	<i>Examining task relationships in multitasking consumer search sessions: A query log analysis</i>	Zhou, X., Zhang, P., Wang, J.	Ciência da Computação; Ciências Sociais	2016
32	<i>A query log analysis of dataset search</i>	Kacprzak, E., Koesten, L.M., Ibáñez, L.-D., Simperl, E., Tennison, J.	Ciência da Computação; Matemática	2017
33	<i>Determinación de grupos de usuarios de bibliotecas digitales mediante el análisis de ficheros log</i>	Martínez-Comeche, J.-A.	Ciências Sociais	2017
34	<i>Design and implementation of real-time log analysis system of map world platform</i>	Zhang, H., Huang, W., Yang, J.	Ciência da Computação; Ciências Sociais	2018
35	<i>Music literature indexing: comparing users' free-text queries and controlled vocabularies</i>	Wu, S., Fan, Y.	Ciências Sociais	2018
36	<i>Transaction log analysis within a bento discovery system</i>	Mischo, W.H., Norman, M.A., Schlembach, M.C.	Engenharia	2019
37	<i>Use patterns of e-journals among the science community: a transaction log analysis</i>	Rafique, A., Ameen, K., Arshad, A.	Ciência da Computação; Ciências Sociais	2019
38	<i>Known-item searches and search tactics in library search systems: Results from four transaction log analysis studies</i>	Schultheiß, S., Linhart, A., Behnert, C., Rulik, I., Lewandowski, D.	Ciências Sociais	2020
39	<i>A transaction log analysis of EBSCO Discovery Service using Google Analytics: the methodology</i>	Fischer, R.K., Iglesias, A., Daugherty, A.L., Jiang, Z.	Computer Science; Ciências Sociais	2021

Fonte: elaborado pelo autor a partir de categorização feita pela Scopus.

O Quadro 2 explicita o vínculo das publicações selecionadas a seis áreas de pesquisa atribuídas pela base *Scopus* para classificar os documentos. Ressaltamos que a base permite que um mesmo documento pertença a mais de uma área de pesquisa, adicionalmente, a análise das

áreas de aplicação da TLA demonstra que o uso dos logs para recuperação e análise pode ser explorado em diversas áreas. Sublinha-se que, dos 39 documentos selecionados para leitura completa, 20 estão vinculados a mais de uma área de estudo, com destaque para a simultaneidade de vínculo com Ciências Sociais e Ciência da Computação, recorrente em 15 publicações. Para melhor visualização das vinculações das pesquisas, a Figura 1 traz os quantitativos por área:

Figura 1. Publicações selecionadas quanto aos vínculos por subáreas de estudo



Fonte: elaborada pelo autor a partir de categorização feita pela Scopus.

Na Figura 1, é possível verificar que a área da **Ciência da Computação** se destaca com 28 trabalhos, seguida da **Ciências Sociais** com 22 trabalhos. Há também publicações vinculadas às Engenharias, à Matemática, à Ciência de Decisão e à Ciência dos Materiais.

Analiticamente, a combinação entre as áreas de Ciência da Computação e Ciências Sociais, com a intersecção das áreas em 15 trabalhos, reforça o entendimento da importância de pesquisas interdisciplinares que associem conhecimentos terminológicos e bibliométricos das Ciências Sociais (sobretudo da biblioteconomia) e saberes computacionais da Ciência da Computação.

Essa combinação reforça o reconhecimento da construção de pesquisas científicas em redes de organização e repercute também na recuperação interconectada de materiais. Portanto, a articulação entre as áreas evidencia a contribuição aos estudos concernentes às experiências

de busca dos usuários em bases de dados, com o objetivo de propor recursos para melhoria contínua de sistemas de recuperação da informação.

Assim, a revisão de literatura sobre aplicação da TLA está organizada em de cinco categorias, as quais correspondem às subseções: 3.1 Histórico e panorama da aplicação da TLA: artigos de revisão, composto por seis artigos; 3.2 TLA aplicada a catálogos de acesso público *on-line* em bibliotecas, com catorze artigos; 3.3 TLA aplicada a sites comerciais e buscadores da *web*, com 3 artigos; 3.4 Modelos propostos de TLA e níveis de análise, com catorze artigos e 3.5 Aplicação da TLA para vocabulários controlados, composta por somente dois artigos. Por fim, a subseção 3.6 apresenta a síntese dos resultados da revisão da literatura, destacando-se o uso da TLA como estratégia de recuperação do percurso e/ou das palavras de busca utilizadas por usuários na interação com diversos sistemas.

Sequencialmente, iniciando a apresentação dos documentos selecionados, a primeira subseção traz a descrição dos seis artigos de revisão que permitem delinear um panorama da aplicação da TLA.

2.1 Histórico e panorama da aplicação da TLA: artigos de revisão

Foi possível identificar seis artigos centrados na revisão de publicações com foco na TLA: Peters (1993); Jansen (2006); Jansen, Taksa e Spink (2008); Jansen (2008); Penniman (2008); Taksa, Spink e Jansen (2008), os últimos quatro integrantes do livro de Jansen, Spink e Taksa (2008). Os materiais permitiram a construção de um panorama dos estudos realizados desde meados da década de 1960.

Peters (1993), com o objetivo de apresentar uma visão histórica dos usos da TLA na Biblioteconomia e na Ciência da Informação, organizou revisão bibliográfica sobre o tema contemplando vinte e cinco anos de publicações, nos quais o autor destacou três fases: a primeira de 1960 a 1970; a segunda do final dos anos de 1970 a meados 1980 e a terceira a partir do final de 1980.

Na primeira fase, os estudos estiveram centrados na avaliação do desempenho dos sistemas e não no comportamento e desempenho do usuário. Já o segundo momento trouxe a análise de logs de transações aplicada pela primeira vez ao estudo de sistemas de catálogos *on-line*, momento em que ficou claro o interesse dos pesquisadores em como os sistemas estavam sendo usados (opções de pesquisa, tempo de sessão, erros recorrentes). A propósito da terceira fase, Peters sublinha a diversificação, visto que alguns pesquisadores optaram por focalizar

tópicos específicos, como pesquisas por assunto, destacaram-se também replicações de TLA, estudos exploratórios e análises de sistemas para conhecimento do comportamento do usuário.

Por fim, como ponto de encontro, Peters (1993, p. 43, tradução nossa) destaca que “[...] a maioria das análises de log de transações tem focado no uso real de sistemas digitais por usuários públicos [...]” e que poucas análises de TLA foram estudadas considerando o uso de um sistema por um grupo experimental selecionado sobre um período, ainda que esse tipo de análise pudesse trazer uma visão importante sobre como os usuários aprendem utilizando um sistema:

Se um sistema oferece mais de uma interface do usuário, seria extremamente útil aprender como e quando os usuários migram de uma interface para outra, se é que o fazem. Estudos de uso de repositório pela equipe de sistemas têm se concentrado no uso de serviços técnicos. Estudos de uso de repositórios por provedores de serviços de referência ainda precisam ser realizados. Outros métodos de observação provavelmente seriam necessários para agrupar pesquisas de sessões em terminais de acesso público e de sessões conduzidas no balcão de referência (PETERS, 1993, p. 42, tradução nossa).

Assim, o avanço cumulativo da TLA como metodologia fica evidenciado pela revisão realizada por Peters (1993), bem como a necessidade de refinamento das investigações para que seu potencial seja explorado a fim de captar dados sobre as interações humanas com os sistemas de armazenamento, de modo que a recuperação de informações seja efetivada de maneira eficiente.

Em 2006, Jansen destacou a análise de logs como meio de conhecer o comportamento de busca dos usuários, através de um “[...] método discreto de coletar quantidades significativas de dados de pesquisa em um número considerável de usuários do sistema” (JANSEN, 2006, p. 409, tradução nossa). Assim, os logs permitiriam conhecer as formas com que os usuários interagem com os mecanismos de busca oferecidos por diferentes sistemas e, dessa forma, identificar as fragilidades dos sistemas que acabam dificultando ou mesmo impedindo a recuperação das informações desejadas pelo usuário.

Jansen (2006) destacou que, apesar do potencial desses dados, poucos estudos tinham sido feitos, por isso se propôs a sintetizar três etapas para exploração: coleta, preparação e análise dos logs, de forma a trazer à tona informações valiosas para a compreensão dos processos de pesquisa *on-line* a partir das ações dos usuários. Em suma, na etapa de coleta, é definido o intervalo de tempo para obtenção dos dados de interação do usuário com o sistema e são coletados os registros relacionados. Em sua proposta, de modo amostral Jansen (2006) coletou o IP, data, hora e as palavras de consulta fornecidas pelos usuários.

Sequencialmente, a preparação consiste na limpeza dos dados, de forma a viabilizar uma análise mais ágil e confiável, posto que durante a limpeza são eliminados dados inconsistentes e ruídos oriundos de registros gerados pelo próprio sistema sem a interação de usuários humanos. Para essa etapa, Jansen (2006) importou os dados dos logs coletados para um banco de dados relacional e atribuiu a cada registro uma chave primária, limpando os dados (ou seja, eliminando inconsistências) e calculando as métricas de interação padrão que serviriam de base para análises.

Finalmente, durante a análise dos logs, o autor buscou caminhos para o uso produtivo das informações obtidas, de maneira a propor melhorias no sistema a partir dos padrões de comportamento dos usuários observáveis pelos registros de logs. Logo, foram sublinhadas as ocorrências de consultas repetidas, palavras e pares de palavras mais usadas, ou seja, a frequência e o grau de associação entre as palavras que auxiliam na compreensão da complexidade da sessão, da estrutura da consulta, bem como as modificações da consulta definidas pelos usuários e os relacionamentos de palavras mais utilizadas.

Destaca-se que Jansen (2006) apresentou um aplicativo de código aberto focado nas interações típicas dos usuários, fornecendo assim a granularidade de dados necessária para análises produtivas, sem registrar grandes quantidades de dados que tornaria o processo de análise lento. O autor chega a disponibilizar o link para acesso ao aplicativo criado, contudo, o mesmo não está mais em operação, não tendo sido encontrada atualização ativa.

Em 2008, a publicação da obra *Handbook of Research on Web Log Analysis*, organizada por Bernard J. Jansen, Amanda Spink e Isak Taksa, contribuiu de forma significativa para a construção de estudos críticos sobre os usos da análise de logs. A obra funciona como uma espécie de manual de pesquisa sobre análise de logs da *web* para análise do comportamento de busca dos usuários. O livro, em seus 25 capítulos divididos em cinco seções, propõe novas direções de pesquisa e novas aplicações do conhecimento existente.

Na primeira seção, há cinco artigos com ênfase na perspectiva histórica de análise de log da *web*, examinando pesquisas como um método complementar para TLA, além da investigação de questões relacionadas à privacidade e à medição de tráfego. Já na segunda seção, composta por cinco artigos, o foco se desloca para a observação de metodologias e métricas a fim de revisitar fundamentos, tendências e limitações das metodologias disponíveis para examinar a granularidade, a validade dos dados de logs e recomendá-los como material para futuros estudos. A terceira seção, com seus cinco artigos, centraliza a análise no comportamento dos usuários durante várias atividades na *web*, sugerindo direções para identificar, encontrar significado e rastrear seu comportamento de busca. A quarta seção,

também composta por cinco artigos, examina a classificação e a identificação de consultas em motores de busca por meio da TLA em diversos contextos. Por fim, a quinta seção traz com seus cinco artigos uma análise contextual e especializada ao propor um novo modelo teórico para avaliar sites de conectores que facilitam as redes sociais *on-line* por meio da extração de informações de textos, exploração do uso da netnografia em estudo da comunicação mediada por computador (JANSEN; SPINK; TAKSA, 2008).

Pela importância da obra organizada por Jansen, Spink e Taksa (2008), a seguir destacamos as contribuições de quatro dos trabalhos compilados quanto à revisão bibliográfica do tema: Jansen, Taksa e Spink (2008); Penniman (2008); Jansen (2008) e Taksa, Spink e Jansen (2008).

Ao destacarem que por meio da TLA o pesquisador é capaz de coletar os dados de modo discreto, Jansen, Taksa e Spink (2008) realizam revisão bibliográfica sobre a TLA de forma a ponderar sobre as potencialidades e limitações do método. Os autores enfatizam que a TLA geralmente se concentra nos comportamentos de interação que ocorrem entre os usuários, o sistema e as informações. Nesse sentido, ponderam sobre vantagens e desvantagens do método: de um lado estão a eficiência, a possibilidade de trabalhar-se com um volume significativo de dados com baixo custo relativo; do outro lado da balança há as dificuldades para preparação, limpeza e análise, especialmente de grandes volumes de registros. Como empecilhos, tem-se quatro fatores principais: I) a menor versatilidade em relação a dados primários, dada a possível escolha a partir de questões de pesquisas particulares; II) a menor riqueza dos dados, uma vez que os logs de transações não oferecem tantas informações como alguns outros métodos, podendo deixar lacunas conceituais; III) registros vagos, pois os vínculos conceituais supostamente mensurados podem ser frágeis; e IV) os usuários cientes que estão sendo gravados podem alterar suas ações e sem tal ciência poderia se configurar violação de privacidade (JANSEN; TAKSA; SPINK, 2008).

Penniman (2008) trouxe uma perspectiva histórica da análise de logs da *web*, descrevendo técnicas e estudos desde as primeiras investigações sobre o tema até sugestões de como o conhecimento construído pode ser aplicado a estudos atuais e futuros. O autor sublinha a necessidade de abordagens adicionais mais qualitativas que viabilizem compreensão profunda do comportamento do usuário, bem como de suas necessidades em relação aos sistemas, a fim de aprimorar os sistemas de recuperação de informação:

A intenção de analisar logs de transações sempre foi dupla: primeiro descrever o que os usuários estão realmente fazendo ao interagirem com um sistema e, segundo, usar esse entendimento para prever quais devem ser as próximas

ações que eles podem tomar para tornar o sistema mais efetivo (ou para corrigir uma dificuldade que encontraram). Os registros de transações permanecem promissores em ambos os casos (PENNIMAN, 2008, p. 19, tradução nossa).

Nas palavras de Penniman (2008) fica clara a relevância da TLA como método preciso e discreto que pode fornecer subsídios confiáveis para a identificação de como ocorrem as interações dos usuários com os sistemas, a fim de que os processos de recuperação possam ser melhorados e, assim, as necessidades de acesso a informações sejam atendidas.

Jansen (2008) apresenta revisão de literatura da análise de logs em pesquisas na *web*, ressaltando o detalhamento e sistematização de cada comando de pesquisa ou consulta feita pelos usuários. O autor pondera que, mesmo com literatura abundante sobre TLA, há poucos manuscritos publicados sobre como realmente conduzi-la.

Em complemento, Taksa, Spink e Jansen (2008) destacaram a natureza interdisciplinar da TLA associada à variedade de aplicações *web* que podem utilizá-la como recurso para coleta e investigação do comportamento de busca dos usuários. Ademais, os autores descrevem metalinguisticamente a observação de abordagens metodológicas recorrentes nos capítulos da própria obra a qual integram e são elas: I) investigação da estrutura conceitual; II) fenomenologia/etnometodologia como metodologias interpretativas que examinam o comportamento dos usuários e suas interações; III) análise de conteúdo, metodologia replicável, utilizada, sobretudo, para investigação quantitativa de grandes conjuntos de dados; IV) etnografia para foco qualitativo em que o pesquisador observa membros de um grupo escolhido em um ambiente natural por um longo período de tempo; V) método histórico, isto é, abordagem diacrônica que coleta e examina fatos sobre eventos, pessoas e o ambiente do passado; VI) análise do discurso tratada como método de avaliação do argumento científico; e VII) estudo de caso, compreendido como estudo abrangente de um único assunto, influenciado por uma seleção adequada da unidade de análise (TAKSA; SPINK; JANSEN, 2008).

A leitura dos trabalhos que compuseram essa categoria de publicações de revisões de literatura sobre TLA demonstrou como os estudos na área vêm sendo conduzidos ao longo das últimas cinco décadas, evidenciando pontos favoráveis e desfavoráveis de sua aplicação. Ainda que os autores tenham descrito avanços inegáveis, suas pesquisas também convergem no sentido de que há muito potencial a ser explorado a fim de captar dados sobre as interações humanas com os sistemas e, assim, melhorar a experiência dos usuários.

2.2 TLA aplicada a catálogos de acesso público *on-line* em bibliotecas

Nessa segunda categoria, composta por pesquisas centradas na aplicação da TLA para propor melhorias aos catálogos de acesso público *on-line* em bibliotecas, foi possível identificar treze trabalhos, publicados entre 1983 e 2020.

Em 1983, Tolle coletou dados na Biblioteca do Congresso e dividiu suas investigações acerca dos catálogos de acesso público *on-line* (OPACs) em três partes principais: o uso real; a percepção do usuário sobre o uso do catálogo e, por fim, a aplicação das descobertas para aprimoramento do sistema. Para tanto, o autor buscou obter logs de transações legíveis dos catálogos *on-line* para posterior análise por desenvolvimento de padrões de pesquisa estocásticos e modelos matemáticos utilizando cadeias de Markov (CHING *et al.*, 2013) e matrizes de probabilidade de transição. Ao comparar o uso real e a percepção do usuário sobre o uso do sistema, Tolle buscou formas de redesenhar e criar sistemas melhores a partir de padrões observados.

A metodologia desenvolvida seguiu as seguintes etapas: I) obtenção dos logs; II) verificação da validade dos dados, reformatação e cópia dos logs; III) redução dos dados pela eliminação de informações irrelevantes para a análise; IV) desenvolvimento de uma taxonomia pelos códigos de atividade, que são mutuamente exclusivos e exaustivos; V) desenvolvimento de mapeamentos do sistema incluindo seus códigos e comandos; VI) divisão dos logs em sessões de usuário e análise para atribuição de códigos de estado; VII) análise subsequente dos dados usando probabilidade de transição de pesquisa matrizes, cadeias de Markov e modelos de distribuição de probabilidade; e, por fim, VIII) atualização da taxonomia e mapeamentos de estado, incluindo outros possíveis mapeamentos (TOLLE, 1983).

Tolle relatou que, embora houvesse o esforço significativo para o desenvolvimento de uma metodologia de análise de logs de transações em catálogos *on-line*, ainda havia um longo percurso de análise para conduzir, em especial, devido à heterogeneidade das fontes de dados, posto que cada sistema requer uma análise própria e independente (TOLLE, 1983).

Jones (1986) ressaltou que, antes de um trabalho mais extenso, quatro estudos dos logs de transações dos usuários foram feitos, a fim de fornecer uma base mais firme para a melhoria do OKAPI (catálogo experimental executado em um site da Biblioteca Britânica). O primeiro deles (Log1) olhou para 214 pesquisas de autor/título realizadas de 14 a 16 de outubro de 1985. O segundo (Log2) observou para 49 autores apenas das pesquisas feitas durante o mesmo período. As duas outras pesquisas estavam focadas no assunto “pesquisas”. Uma examinou 2.372 pesquisas de assunto feitas de novembro de 1984 a outubro de 1985 (Log3) para

evidências de erros ortográficos e de digitação; a outra (Log4) analisou 285 pesquisas de assunto, a fim de estimar seu índice de sucesso.

A partir dos dados coletados em todas essas investigações, o artigo de Jones (1986) buscou ilustrar problemas no uso de OKAPI (e de OPACs em geral) e oferecer um esboço do que poderia ser melhorado a fim de otimizar a recuperação. Para tanto, a facilidade de uso foi vista como a característica mais importante para o uso eficiente de um catálogo *on-line*, por isso, os *designers* do OKAPI consideraram a análise de logs para revelar várias áreas passíveis de melhorias na operação do catálogo: I) área de inserção de informações pelo usuário: erros de ortografia, colocação de informações no lugar errado; II) acesso ao assunto; III) tempo que os usuários levam para fazer pesquisas de itens específicos e IV) funcionamento das árvores de pesquisa: a resposta do usuário a opções dadas automaticamente pelo catálogo (JONES, 1986).

O OKAPI pretendeu facilitar a recuperação, incorporando técnicas que funcionavam automaticamente, por entender que a eficácia do catálogo *on-line* está relacionada à adaptabilidade e à tolerância com as entradas de seus usuários e procedimentos com erros, demonstrando bom senso no tratamento dos dados. Em alguns casos, em vez de transferir a responsabilidade pela correção de erros ao usuário, o catálogo tentava oferecer ajuda com sugestão automática de correção. No entanto, Jones (1986) enfatizou que não se deveria superestimar tal ajuda, pois o OKAPI não era um sistema inteligente, mas sim uma estratégia de *design* para desenvolver uma interação que poderia lidar automaticamente com algumas “limitações” das entradas dos usuários, de modo a facilitar a recuperação.

Em 1991, nos EUA, Zink monitorou o sucesso de pesquisas dos usuários pelos logs de transações no catálogo *on-line* WolfPAC, componente do sistema integrado de bibliotecas com registros de seis universidades. O estudo considerou como malsucedidas as pesquisas que não recuperaram nenhum registro bibliográfico. Foram analisadas 24 horas de registros, consideradas as pesquisas por autor, título e assunto. Zink verifica que o maior problema enfrentado pelos usuários do WolfPAC foi a falta de familiaridade com o catálogo e também com o conceito de vocabulário controlado. A maioria dos usuários utilizava a linguagem natural sem demonstrar consciência sobre a existência de uma lista de termos de assunto que poderiam classificar melhor e trazer mais respostas do sistema. Com os dados analisados, Zink (1991) enfatizou tanto a importância de aprimorar a interface do catálogo quanto de fornecer materiais informativos aos usuários, de modo a facilitar a recuperação de informações desejadas.

Blecic *et al.* (1998) dedicaram-se ao estudo comparativo de análise de log de transações derivados de dois conjuntos extraídos de Catálogos de Acesso Público *On-line* (OPAC) em um período de seis meses. A análise do primeiro conjunto de dados revelou que os

usuários tiveram dificuldade com as técnicas básicas de pesquisa. O segundo conjunto de dados, analisado após as alterações de tela terem sido feitas, mostrou diferenças estatisticamente significativas nos resultados da pesquisa, evidenciando que o monitoramento de OPACs por meio da análise do log de transações, pode levar a uma recuperação aprimorada quando as alterações são feitas em resposta a uma análise dos padrões de pesquisa do usuário. Logo, os autores concluíram que avaliar as telas introdutórias e propor a implementação de mudanças em resposta aos erros observados nos logs de transações, resultaram em alterações de comportamentos de pesquisa e maior índice de sucesso, provando que o método é uma ferramenta valiosa que poderia ser adaptada por outras instituições.

Jones *et al* (2000), ao identificarem a necessidade de entender as maneiras pelas quais os usuários interagem com os sistemas na prática, viram na TLA uma fonte de informações sobre o uso e o comportamento do usuário tanto para análises automatizadas (por meio do cálculo de estatísticas resumidas) quanto manuais (examinando sequências de consulta em busca de pistas semânticas sobre motivações e estratégia de pesquisa). Dessa forma, os autores conduziram uma análise do log de transações sobre a atividade dos usuários na coleção de relatórios técnicos de Ciência da Computação da Biblioteca Digital da Nova Zelândia e descrevem suas percepções sobre problemas de *design* da interface.

Os autores coletaram logs de usuários no período de abril de 1996 a julho de 1997, totalizando mais de 30.000 consultas registradas e analisadas. Como parâmetros investigativos, Jones *et al.* (2000) extraíram dados demográficos dos usuários disponíveis na biblioteca, discutiram o uso de operadores e opções de pesquisa em consultas e examinaram os padrões na construção e refinamento, além de descreverem erros comuns na pesquisa e examinarem a distribuição das palavras de consulta que aparecem nos logs.

A partir das análises realizadas, Jones *et al.* (2000) concluíram que a maioria dos usuários aceitam as configurações padrão oferecidas pelo sistema, independentemente de quais sejam essas configurações. A maioria das sessões de usuários são muito curtas e simples, indicando que eles buscam por resultados rápidos e com pouco investimento de esforço, o qual seria necessário para especificações detalhadas de suas necessidades de informação. Ademais, as investigações indicaram que a maioria dos usuários acabaram se concentrando apenas nos primeiros documentos devolvidos, portanto, sistemas mais eficazes devem apoiar os usuários oferecendo-lhes interface simples com recuperação de materiais em ordem de relevância, já que as pesquisas de exploração dos conjuntos retornados parecem ser mínimas.

Mahoui e Cunningham (2000) também reconheceram a potencialidade da TLA e propuseram uma análise comparativa dos logs de transações dos usuários em duas coleções de

computação, a coleção de Relatórios Técnicos de Ciência da Computação (CSTR) da Biblioteca Digital da Nova Zelândia e a coleção de Bibliografias de Ciência da Computação de Karlsruhe (CSBIB). Para tanto, retomaram as investigações feitas por Jones *et al.* (2000) e acrescentaram os registros de logs extraídos da CSBIB. Embora as ações pudessem ser agrupadas em sessões de usuário, usuários individuais não puderam ser identificados e usuários locais foram excluídos, pois, durante o período estudado, muitas consultas locais foram enviadas como testes de sistema.

Assim como Jones *et al.* (2000), Mahoui e Cunningham (2000) constataram que a maioria dos usuários registraram pesquisas curtas e rápidas. No CSTR, pouco mais de 82% das pesquisas utilizaram até três palavras de busca, já no CSBIB mais de 91% utilizaram até três palavras para pesquisa. Além disso, operadores booleanos não foram utilizados em 66% e 84% das pesquisas feitas pelos usuários no CSTR e CSBIB, respectivamente, indicando o predomínio de pesquisas simples. Os autores enfatizaram que, embora se esperasse que pesquisadores de computação pudessem explorar ativamente o software e suas funções, tais usuários aceitaram, na ampla maioria das vezes, as configurações padrão do sistema. As diferenças no comportamento de pesquisa entre os dois sistemas incluíram uma tendência dos usuários CSBIB de emitir mais consultas em uma sessão, e para consultas consecutivas em uma sessão para mostrar um menor grau de sobreposição de palavras. Porém, dadas as limitações da TLA, os autores sugeriram, ao final da pesquisa, estudos adicionais qualitativos baseados em entrevistas de pequenos grupos de usuários para determinar a base dessas diferenças.

Em 2001, Mahoui e Cunningham estudaram a TLA no comportamento de pesquisa de usuários da área de ciência da computação na biblioteca digital de literatura científica ResearchIndex. Os autores destacaram as diversas técnicas para analisar o comportamento dos usuários em bibliotecas e a TLA como método não intrusivo para coleta de dados a fim de identificar padrões na construção de consultas e conhecer o comportamento de pesquisa dos usuários. As variáveis analisadas foram a data e hora, o identificador IP e o termo de consulta com seus respectivos filtros e as sessões dos usuários delimitadas por 30 minutos.

Mahoui e Cunningham (2001) detalharam os métodos para pesquisar dados demográficos dos usuários, a duração considerada para delimitação das sessões, bem como a complexidade da consulta. Os autores relataram que as sessões se aproximaram da quantidade de usuários da biblioteca digital, porém não foi possível incorporar especificamente dados demográficos pela coleta dos logs de transações, uma vez que os usuários têm acesso público às coleções digitais sem a necessidade de realização de um controle de login e senha para identificá-los. No entanto, esse anonimato, embora traga algumas limitações para a análise, tem

também vantagens quanto a agilidade e a liberdade de pesquisa, posto que a consulta pode ser feita de modo imediato, sem necessidade de login, o que também não cria vínculos para configuração do perfil identificado do usuário, preservando sua privacidade.

A propósito da amostra e delimitação das sessões, Mahoui e Cunningham (2001) extraíram 1.541.148 consultas ou acessos de navegação divididos em 46.486 sessões. Os autores verificaram que a maioria das sessões começou com uma pesquisa que ignorou a tela principal de consulta, sobre isso ponderaram que o acesso possivelmente seja iniciado por meio de links de consulta resultantes de mecanismos de pesquisa como o Google. Os autores justificam também a delimitação temporal das sessões pelo intervalo de 30 minutos, por tratar-se de uma heurística plausível do ponto de vista do senso comum e ter sido adotada pela maioria da comunidade pesquisadora de TLA e *web mining*. Outro dado relevante trazido pelos autores foi que, nas sessões analisadas, apenas 53% incluíam palavras de busca dos usuários. Ademais, a respeito da complexidade, para cada coleção, pelo menos 80% das consultas dos usuários contêm três ou menos palavras, revelando a preferência por buscas curtas, a partir de poucas palavras (MOHOUI; CUNNINGHAM, 2001).

Avery e Tracy (2014) utilizaram a TLA para avaliar o comportamento de usuários estudantes na biblioteca da Universidade de Illinois, em Urbana-Champaign, a fim de verificar se estes seguiam as instruções dadas pelo professor durante a aula e, posteriormente, colocavam em prática as estratégias de pesquisa aprendidas. Os autores investigaram igualmente como os estudantes interagiram com as sugestões automatizadas da ferramenta de pesquisa e se foram persistentes nas revisões de suas próprias pesquisas. O estudo incluiu um total de 1.636 pesquisas exclusivas, realizadas por 579 alunos usando o *Easy Search*. Para cada evento, os campos de logs de transações incluíram informações da ação realizada pelo estudante, a pesquisa anterior feita por aquele aluno, o tempo da pesquisa, qualquer sugestão de melhoria feita pela interface *Easy Search* e quaisquer bancos de dados em que o aluno clicou na página de resultados. Também incluíram quaisquer cliques em links de sugestão de interface, como correções na ortografia das palavras, refazendo a pesquisa como uma pesquisa de autor em vez de uma pesquisa de palavra-chave e toda a tabulação foi feita em *Microsoft Excel*. Os autores justificam o uso da TLA por ela oferecer a oportunidade de observar discretamente o comportamento de busca dos estudantes em tempo real. No entanto, uma desvantagem potencial apontada é a disponibilidade dos logs, comumente difíceis de acessar.

Martínez-Comeche (2017), ao compreender que a análise de logs contribui para a análise da recuperação da informação em perspectivas diversas, destaca que, seja qual for o objetivo, três possíveis níveis de análise se mostram recorrentes nas pesquisas publicadas, a

saber: palavras de busca, duração e filtros utilizados nas sessões. As sessões são entendidas como um conjunto de buscas realizadas em um intervalo de tempo preestabelecido. Com o objetivo de melhorar a interface e a experiência dos usuários da Biblioteca Digital Hispânica (BDH), o autor estabelece um intervalo máximo de 24 horas para a delimitação das sessões, justificando limites impostos pelo método de compilação de informações do servidor da BDH.

Já em 2019, Mischo, Norman e Schlembach analisam uma amostra de 5.400 pesquisas de 988 mil logs coletados ao longo de um ano na Biblioteca Urbana-Champaign, Universidade de Illinois (EUA). O objetivo dos autores foi conhecer o percurso de pesquisa dos usuários por meio do rastreamento dos cliques feitos por eles. As variáveis analisadas foram o tipo de pesquisa (item conhecido ou pesquisa exploratória), a taxa de sucesso e os comportamentos específicos do usuário, tais como: palavras de busca, percentual de palavras oferecidas pelo sistema de pesquisa e percentual de uso dos usuários dessas palavras sugeridas.

Rafique, Ameen e Arshad (2019) verificaram, por meio da TLA, a frequência de visualizações de página, sessões, duração de sessão e tamanho dos dados a partir de logs extraídos durante um ano na Biblioteca Digital Comissão de Educação Superior (HEC), na Universidade de Engenharia e Tecnologia (UET), Lahore do Paquistão. Com esse estudo de caso, os autores adotaram uma abordagem quantitativa para identificar os padrões de uso de periódicos eletrônicos baseados em evidências, analisando o uso desses bancos de dados por cientistas ao longo do tempo. Todos os dados foram calculados com o software SAWMILL e inseridos em *Microsoft Excel* para apreciação, a fim de auxiliar a Comissão de Educação Superior (HEC) do Paquistão e os profissionais da informação para a melhor gestão do acesso às bases de dados.

Fischer *et al.* (2021) utilizaram a TLA para viabilizar a compreensão sobre como o *design* de um banco de dados pode ser melhorado e até que ponto a população está usando os recursos de pesquisa avançada da ferramenta *on-line*. Para tanto, analisaram transações do EBSCO *Discovery Service* registradas no *Google Analytics* e, conforme os autores, o aplicativo “*Google Analytics: Event Tracker*” da EBSCO permite a geração de um código de rastreamento HTML personalizado que pode ser colado na parte inferior da marca de EBSCO Admin. Este código permite que o *Google Analytics* registre palavras de pesquisa, uso de configurações, uso do histórico de pesquisa, cliques em links personalizados, acesso a registros detalhados de itens, uso da página de pesquisa avançada e outros recursos. Todavia, geram uma apresentação difícil de interpretar e, a fim de facilitar essa interpretação, os autores criaram um *script* de extração em linguagem R e tabelas dinâmicas do MS Excel para tabular o número de pesquisas avançadas e outras funções de interesse.

Em prosseguimento, Schultheiß *et al.* (2020) buscaram compreender as interações dos usuários com os sistemas de informação utilizados por duas bibliotecas alemãs, a fim de verificar suas estratégias de pesquisa e propor aprimoramentos para os sistemas de organização. Para tanto, dividiram sua investigação em quatro estudos que demonstram como a análise de transações de logs pode ser utilizada como meio de entender as interações reais entre usuários e sistemas de informação, por “[...] permitir uma visão em grande escala das necessidades de informação dos usuários, suas palavras de consulta e suas interações com os sistemas de pesquisa” (SCHULTHEIß *et al.*, 2020, p. 1, tradução nossa).

Nos estudos I e II de Schultheiß *et al.* (2020), foram analisadas 4.309 consultas com o objetivo de verificar a eficácia da recuperação para itens conhecidos. Para o estudo I, por inspeção manual, foi selecionada amostra aleatória de 1.981 consultas extraídas da biblioteca EconBiz em julho de 2014, cujos resultados indicaram a importância da qualidade dos metadados, posto que de todas as consultas com zero acertos, 20% estão associadas a metadados incompletos ou incorretos. Para o estudo II, houve a divisão de duas amostras extraídas da biblioteca Beluga (sessão definida pelo IP, cookie e um corte temporal) uma de três dias consecutivos e outra de três dias consecutivos uma semana após a primeira, com finalidade de avaliar o algoritmo (SCHULTHEIß *et al.*, 2020).

Schultheiß *et al.* (2020), nos estudos III e IV analisaram 4,2 milhões de sessões com o objetivo de verificar as estratégias de pesquisa utilizadas pelos usuários. No estudo III, são investigados os logs extraídos de novembro de 2013 a novembro de 2014 da biblioteca EconBiz divididos em dois conjuntos, um com todas as sessões para estatísticas descritivas, e outro com uma amostra aleatória, mas atendendo ao requisito de ter ao menos duas páginas visitadas, das quais pelo menos uma é um resultado da pesquisa com classificação manual das consultas de táticas de pesquisa. No estudo IV, a amostra de logs foi coletada da biblioteca Beluga, de abril a setembro de 2017, e foi dividida em dois conjuntos: o primeiro, para fins de estatísticas descritivas, com 717.092 sessões e o segundo, com amostra aleatória extraída do primeiro, com 2.700 das consultas e itens sessões com pelo menos uma consulta e de 2 a 100 ações registradas e classificação manual. Tanto o estudo III quanto o estudo IV buscaram investigar as estratégias de pesquisa dos usuários, porém o estudo III focalizou a pesquisa por itens conhecidos, já o estudo IV por itens desconhecidos. Os resultados dos estudos III e IV revelaram que a maioria das consultas contém entre duas e três palavras. O estudo III mostrou que 28% de todas as consultas foram feitas utilizando duas palavras, seguida por 19% com três palavras e 15% com apenas uma palavra. O estudo IV chegou a resultados semelhantes, com uma mediana de três palavras por consulta (SCHULTHEIß *et al.*, 2020).

De acordo com os trabalhos agrupados nesta categoria, é fundamental a busca por métodos que permitam acompanhar as estratégias de busca de usuários para que seja possível aprimorar suas experiências de busca, de modo a garantir o acesso e a usabilidade das informações e conhecimentos produzidos e armazenados nas instituições de ensino e pesquisa. Ademais, os estudos demonstraram a viabilidade de uso da TLA para redesenhar e criar sistemas melhores com relação à recuperação de informações desejadas, no entanto, apontam a dificuldade de acesso aos logs como obstáculo à ampliação das pesquisas na área e enfatizam igualmente a necessidade de estudos qualitativos adicionais dadas as limitações da TLA.

2.3 TLA aplicada a sites comerciais e buscadores da web

A utilização da TLA para investigação do comportamento dos usuários em sites comerciais foi tratada por Rice e Borgman (1983), Moukdad e Large (2001) e Zhou, Zhang e Wang (2016).

Em 1983, Rice e Borgman enfatizavam não apenas o uso de sistemas computadorizados para facilitar rotinas humanas, mas, principalmente, a possibilidade de compreender o funcionamento desses sistemas e monitorar como eles são usados para armazenar, trocar e recuperar informações (RICE; BORGMAN, 1983). Ao analisar um sistema para recuperação de informações pelo usuário, os autores afirmam que “[...] o monitoramento por computador de um sistema de informação ou comunicação é basicamente o registro automático do tipo, conteúdo, ou tempo de transações feitas por uma pessoa em um terminal com aquele sistema [...]” (RICE; BORGMAN, 1983, p. 247). A propósito, os autores citam o DIALOG, SDC ORBIT e Lexis como exemplos de sistemas comerciais de recuperação antes mesmo da popularização da internet.

O monitoramento do comportamento dos usuários poderia ser feito, já na época, pela coleta de dados como: I) números de identificação do terminal e do usuário; II) comandos de protocolo (específicos para idiomas de interação) para enviar mensagem, realizar pesquisa de autor, examinar diretório de usuário, imprimir, armazenar, transferir dados e mensagem; III) conteúdo de texto completo (de pesquisa, resultado, mensagem etc.) IV) tópico, palavra-chave como resumo do conteúdo, dentre outras partes das transações que poderiam ser capturadas de acordo com os recursos oferecidos pelos sistemas. Rice e Borgman (1983) destacam também os dois propósitos analíticos centrais dessas investigações, são eles: avaliar os usos e a utilidade de tais meios de comunicação pelos indivíduos e obter indicadores dos impactos de tal comunicação sobre os usuários e suas configurações sociais.

Os principais problemas, já na década de 1980, relacionavam-se à extensão dos dados e às estratégias gerenciais, bem como a questões éticas e de privacidade. Tendo em vista a exponencialidade de dados gerados, o consequente aumento do custo computacional para tratamento e a necessidade de salvaguardar a privacidade e garantir a permissão consciente para acessar e analisar os dados gerados sem incorrer em crime contra o direito à privacidade individual (RICE; BORGMAN, 1983). Ademais, a acurácia e a confiabilidade dos dados devem passar por testes constantes, já que “[...] os dados monitorados por computador obviamente não retratam toda a imagem da comunicação humana [...]” (RICE; BORGMAN, 1983, p. 251).

Moukdad e Large (2001) buscaram descrever as percepções dos usuários de buscadores da *web* a partir da TLA, tomando como amostra 2.000 consultas de pesquisa reais enviadas por usuários a um mecanismo de pesquisa da *web*, o WebCrawler, em duas sessões de captura separadas com 30 minutos cada. A amostra viabilizou uma análise quantitativa e qualitativa, pois, conforme os autores, foi possível observar estatisticamente as características das consultas e avaliar as próprias consultas na tentativa de entender como esses usuários veem a *web* a fim de propor melhorias na estrutura dos motores de busca, de modo que possam ser adaptados para se ajustar a modelos eficazes aos usuários.

As variáveis analisadas por Moukdad e Large (2001) consideraram:

- (1) Uso de modificadores de pesquisa WebCrawler individuais, aplicados corretamente ou incorretamente (três operadores booleanos e sinais +/-, capitalização da primeira letra de um termo, aspas, parênteses).
- (2) Uso incorreto de qualquer modificador de pesquisa WebCrawler.
- (3) Número de palavras individuais na consulta.
- (4) Categoria linguística à qual a consulta pode ser atribuída: palavra(s) -chave ± incluindo o número total de palavras-chave; frase completa, incluindo número de palavras na frase: ± corretamente / incorretamente formadas; ± humor imperativo; ± questão sem ponto de interrogação final; ± questão seguida de ponto de interrogação (MOUKDAD; LARGE, 2001, p. 352).

Por fim, os autores observaram que a estrutura linguística de muitas consultas se assemelhou a um modelo de comunicação humano-humano que provavelmente não produz resultados satisfatórios em um ambiente de comunicação humano-computador, como o oferecido atualmente pela *web*. Assim, há a necessidade de projetos de sistemas mais intuitivos a partir da compreensão mais completa do comportamento do usuário, que busca interfaces amigáveis e rápidas. Os autores apontaram que a amostra de consultas de uma seção transversal de WebCrawler trouxe registros bastante heterogêneos, que podem não ser representativos de todas as comunidades de usuários da *web*, todavia, a aplicação simples traz contribuições para

a compreensão geral do comportamento de usuários não restritos a comunidades específicas como a acadêmica, por exemplo (MOUKDAD; LARGE, 2001).

Zhou, Zhang e Wang (2016) se propuseram a examinar logs extraídos de sessões do taobao.com, o maior site de comércio eletrônico *Consumer to Consumer* (C2C) chinês, a fim de compreenderem as relações entre as tarefas em multitarefa na pesquisa de produtos. Os autores coletaram 2.910 consultas, de 47.387 usuários, em 18.102 pesquisas de produtos. Dentre as variáveis analisadas estiveram o endereço a partir do qual o clique é feito; o URL (*Uniform Resource Locator*) da página da *web* que o usuário visitou; as palavras de consulta inseridas pelo usuário (se houver), a data e o horário em que um usuário abriu um URL, considerando como intervalo de sessão 45 minutos. Após a extração dos logs, os autores empregaram o agrupamento hierárquico de palavras de consulta com base na similaridade de Jaccard (NIWATTANAKU *et al.*, 2013) de pares para identificar tarefas de pesquisa. Dessa forma, puderam concluir que os resultados obtidos forneceram uma compreensão exploratória das relações entre várias tarefas de pesquisa de produtos, fato que indicou a potencialidade comercial da recomendação de consultas e de produtos.

Os trabalhos com aplicações da TLA envolvendo sites comerciais e buscadores *web* evidenciaram a necessidade de respeito a questões éticas e de privacidade e também de aprimoramento na comunicação entre usuários humanos e sistemas, a fim de ampliar o acesso a dados e produtos de interesse. Destacou-se, sobretudo, o estudo de Zhou, Zhang e Wang (2016) que coletaram consultas e a sequência de cliques feitos pelos usuários nos intervalos considerados como sessões.

2.4 Modelos propostos de TLA e níveis de análise

Nessa categoria, são descritos catorze trabalhos os quais propõem modelos de TLA: Meister e Sullivan (1967); Kaske (1993); Sandore *et al.* (1993); Kurth (1993); Jansen (2008); Ozmutlu, Ozmutlu e Spink (2008); Bernstam, Herskovic e Hersh (2008); Shiri (2009); Herrerías e Gómez (2010); Huurnink *et al.* (2010); Park e Lee (2013); Li *et al.* (2015); Kacprzak *et al.* (2017); Zhang, Huang e Yang (2018).

Adicionalmente, apresenta-se também a investigação realizada por Laipelt (2015b), a qual não foi recuperada pela base de dados Scopus, no entanto, consiste em pesquisa pioneira no cenário brasileiro quanto ao uso de logs para conhecimento de vocabulário empregado por usuários em repositório digital.

As publicações citadas trazem a quantificação das principais variáveis de análise da TLA, as etapas definidas para coleta, preparação e a análise de registros, envolvendo os níveis

de pesquisa de termo, consulta e sessão, assim como algoritmos de identificação de padrões nos dados.

Em 1967, Meister e Sullivan, ao analisarem a recuperação de informações, identificaram o aumento em proporções geométricas na produção de dados que, por sua vez, exigiam soluções para sua organização. A efetividade da organização, contudo, já deveria ter acurácia e garantir acessibilidade e usabilidade, por meio de técnicas de representação as quais fossem, simultaneamente, simples, objetivas e fáceis de aprender.

Meister e Sullivan (1967) foram precursores quando propuseram a metodologia de Teste de Desempenho Controlado (CPT) e empregaram critérios, anteriormente utilizados em pesquisas *off-line*, em pesquisas no sistema *on-line*, de forma a comparar os resultados obtidos em ambas. Os autores definiram, então, como pesquisas bem-sucedidas aquelas que resultaram na visualização de um ou mais títulos, de modo a fornecer as duas variáveis de avaliação do estudo: frequência de uso e o grau dos usuários bem-sucedidos.

Kaske (1993) discutiu sobre o que era necessário analisar antes da escolha de um ou mais métodos de análise de logs e propôs um modelo que quantificasse as principais variáveis da interação humano-computador. Nessa discussão, o autor relatou como umas das principais limitações o tamanho e abrangência dos logs.

Um registro pode estar completo (registrando todas as ações realizadas pelo usuário, proprietário ou equipe, e todas as respostas do sistema para o usuário), ou um registro só poderia ser contado por diferentes tipos de comandos de pesquisa enviados ao sistema pelos proprietários e equipe sem informação sobre o que foi transmitido de volta pelo sistema (KASKE, 1993, p. 80, tradução nossa).

Assim, o modelo proposto por Kaske foi exemplificado por cinco variáveis: proprietário, usuário final, localização, acesso e sistema. O autor defendeu a criação de sistemas nacionais e padrões internacionais para dados de registro de transações a fim de fornecer uma base de conhecimento acessível para pesquisadores da área. Além disso, já naquele momento, o pesquisador apontava a necessidade de conferir maior atenção para a indústria da informação como um todo e de ampliar as análises de logs como fontes de informações confiáveis e úteis. Kaske enfatizou ainda uma tendência em complementar a TLA com outras metodologias para superar suas limitações no entendimento do comportamento de busca de informações pelos usuários (KASKE, 1993).

Ainda em 1993, Sandore *et al.* afirmaram que a natureza difícil do TLA, sem dúvida, dificultou o interesse em seu desenvolvimento como método para analisar a interação do usuário com sistemas *on-line*. Nesse sentido, os autores propuseram vários estudos para atestar

a disponibilidade e utilidade da TLA, assim como confirmá-la como método de pesquisa viável para aqueles profissionais interessados tanto nos comportamentos dos usuários quanto no desempenho dos sistemas utilizados.

Assim, Sandore *et al.* (1993, p. 105) apontaram a necessidade do desenvolvimento de “[...] padrões nacionais e internacionais para a coleta de dados de logs de transações, incluindo definições de palavras comuns associadas com análise de registro de transações”. Ademais, a disponibilização desses registros a bibliotecários é defendida como modo de promover uma melhor comunicação entre desenvolvedores para aprimoramento dos sistemas utilizados, visto que relatórios automáticos nem sempre fornecem as informações desejadas. Por fim, a importância do aprimoramento das aplicações da TLA é relacionada ao melhoramento dos sistemas, mas também da proficiência dos pesquisadores, vista a possibilidade de rastreamento de seus passos nos ambientes de busca que de outra forma seriam irrecuperáveis.

Para Kurth (1993), sobre as potencialidades e limitações da TLA, “[...] os dados do registro de transações descrevem com eficácia quais pesquisas clientes fazem e quando eles fazem, mas eles não refletem, exceto por inferência, quem faz a pesquisa, porque as fazem e quão satisfeitos ficam com seus resultados [...]” (KURTH, 1993, p. 98). Nesse sentido, Kurth destacou quatro elementos estruturais fundamentais para a TLA: o sistema; o usuário e seu processo de pesquisa; a análise dos dados do registro de transações; e as questões éticas e legais envolvidas no acesso a esses dados. Assim, empecilhos envolvendo a coleta de logs para posterior análise pode ser de ordem mecânica, temporal, espacial e intersistêmicas. O autor citou também a necessidade de se diferenciar os desempenhos do sistema e dos usuários e, a partir de revisão no catálogo *on-line* britânico realizada por Kinsella e Bryant (1987), corrobora que o limite mais crucial da TLA é sua incapacidade de isolar e caracterizar usuários de maneira individual, a fim de descrever os padrões de seu uso.

Em vista do exposto, Jansen (2008) citou as três etapas principais para uma produtiva análise de logs de pesquisa, a saber: coleta, preparação e análise de dados de acordo com os objetivos definidos para a investigação. O autor enfatizou que, para preparação de dados, é imperativo considerar os dados de registro em um banco de dados relacional (ou outro software de análise), a fim de atribuir a cada registro uma chave primária; limpar os dados, eliminando incorreções e identificar padrão de cálculo e métricas de interação que fundamentam a análise posterior.

Após a coleta, a etapa de preparação se subdivide em três níveis: limpeza de dados; compilação dos dados e normalização de episódios de busca. Na limpeza, são eliminados erros e dados corrompidos e Jansen (2008) recomendou que uma forma de identificar rapidamente

erros é classificar cada campo em sequência a fim de observar desvios, já que dados corrompidos ou fora do padrão para determinado campo tendem a se agruparem. Feita a limpeza, a compilação envolve a análise sintática dos logs; para tanto, geralmente, são considerados três campos: tempo, IP e URL de pesquisa. Há, portanto, uma teia de dados que contém consultas de usuários humanos ou não, a qual deve ser analisada conforme os objetivos da pesquisa. Por fim, a normalização de episódios de busca pode ser utilizada para diferenciar páginas de resultados recuperados que o usuário visitou de páginas abertas e consultadas efetivamente (JANSEN, 2008).

Especificamente sobre a etapa de análise, realizada após a preparação, Jansen (2008) também distingue três níveis: termo, consulta e sessão. Para esta análise define-se termo como “[...] uma sequência de caracteres separados por algum delimitador, como um espaço ou algum outro separador [...]” (JANSEN, 2008, p. 110, tradução nossa). Neste nível são consideradas a frequência de ocorrência das palavras, assim como a presença de palavras únicas e a coocorrência de palavras a propósito de pares comuns. Ainda neste nível, é possível calcular graus de associação de pares de palavras a partir de medidas estatísticas.

Quanto à análise a nível de consulta, define-se consulta como “[...] uma lista de *strings* de uma ou mais palavras submetidas a um mecanismo de pesquisa [...]” (JANSEN, 2008, p. 111, tradução nossa). Nessa compreensão, o autor distingue consulta inicial de consultas subsequentes (modificadas) e pondera que é necessário ter atenção às especificidades de registro de cada mecanismo de pesquisa *web*, pois o modo de registrar as consultas pode ser diferente, por exemplo, quando o pesquisador acessa uma nova página de resultados, a interação também pode ser registrada como uma consulta idêntica ou como lista de resultados que viabilizem a observação de padrões de visualização. Jansen (2008, p. 111-112, tradução nossa) distingue também consultas únicas e repetidas, as primeiras se referem a “[...] uma consulta que é diferente de todas as outras consultas no log de transações, independentemente do pesquisador [...]”, já uma consulta repetida é aquela “[...] que aparece mais de uma vez dentro do conjunto de dados por dois ou mais pesquisadores [...]”. Por fim, há ainda de se considerar a complexidade da consulta, a qual examina sua sintaxe, incluindo o uso de técnicas de pesquisa avançadas, como booleanas e outros operadores de consulta.

Na análise em nível de sessão, entende-se por sessão um intervalo de tempo considerado para a análise da interação,

“[...] uma interação de sessão é qualquer troca entre o pesquisador e o sistema (ou seja, enviar uma consulta, clicar em um hiperlink etc.). Um episódio de busca é definido como uma série de interações dentro de uma duração limitada

para atender a uma ou mais necessidades de informação [...]” (JANSEN, 2008, p. 112, tradução nossa).

Após distinguir os níveis de análise supracitados, Jansen (2008) enfatiza o potencial da TLA como métodos de coleta não intrusiva, barata e discreta e de centenas de milhares e mesmo milhões de interações, posto que o custo é relativamente baixo por envolver apenas o software e o armazenamento, já a coleta discreta permite o acesso ao comportamento inalterado dos pesquisadores. Quanto às limitações, o autor frisa que certos tipos de dados não estão no log de transações, indivíduos e identidades, por exemplo, assim também não é possível coletar dados demográficos. Outras limitações citadas são: a ausência de registros sobre os motivos para a pesquisa e outros aspectos qualitativos de uso; a incompletude dos dados devido ao armazenamento em cache de dados do servidor na máquina cliente ou servidores proxy.

Ozmutlu, Ozmutlu e Spink (2008) chamaram a atenção para os estudos de TLA a fim de identificar tópicos de consulta em mecanismos de pesquisa, desde a análise de palavras, a identificação de sessões e o agrupamento de consultas até a categorização de textos e o processamento de linguagem natural, considerando, por exemplo, a investigação da probabilidade condicional de uma palavra ocorrer dado que outra palavra ocorreu, já que em processamento de linguagem natural, as palavras são parâmetros observáveis morfológica, sintática e semanticamente.

Bernstam, Herskovic e Hersh (2008) versaram sobre a TLA no contexto biomédico, enfatizando que a análise de log de consulta em biomedicina tem uma história relativamente curta em comparação com a *web*, mas promissora para o aprimoramento terminológico, de modo a favorecer a recuperação de informações com estratégias que aproximem a forma de registro aos modelos mentais dos usuários.

Shiri (2009) enfatizou a interdisciplinaridade exigida pelo desafio de realizar pesquisas inclusivas e eficazes em uma gama diversa e heterogênea de bibliotecas digitais, bancos de dados baseados na *web* e mecanismos de busca. Assim, o autor propôs um estudo empírico que examinou, por meio da TLA e sua comparação com tesouros Inspect e Compendex da área de nanociência e tecnologia, a interdisciplinaridade nas consultas dos usuários feitas na biblioteca digital NANOnetBASE entre julho de 2004 e outubro de 2006. Os registros foram analisados para identificar padrões de comportamento de pesquisa dos usuários e verificar a presença da interdisciplinaridade, explorando o valor potencial dos tesouros no suporte à formulação de consultas pelos usuários. Para tanto, Shiri (2009) adotou três parâmetros comparativos, a saber, correspondência exata, correspondência parcial e correspondência de palavras, tanto ampla

quanto restrita, utilizando o método de análise de dados de transação de Jansen (2006) para conduzir análises de nível de termo, nível de consulta e nível de sessão.

Herrerías e Gómez (2010) descreveram uma aplicação da TLA no âmbito da computação forense para auxiliar na detecção de ataques, otimizando o trabalho do investigador forense. No artigo, os autores apresentam o *Automated Forensic Diagnosis System* (AFDS), um software implementado em Python, composto por um “Conhecimento Base de Ataque” e uma série de processos de análise de log funcionando de modo articulado com o objetivo de reconstruir as ações de ataque, após a ocorrência de um incidente de segurança. Conforme os autores:

[...] o eixo central desta proposta é o Módulo de Correlação de Eventos, que ajuda a detectar ataques de várias etapas, bem como reduzir a taxa de falsos positivos. O objetivo final do sistema é auxiliar o investigador forense no processo de busca de evidências diminuindo o tempo e a complexidade do processo. (HERRERÍAS; GÓMEZ, 2010, p. 659, tradução nossa).

Huurnink *et al.* (2010) mostraram a aplicação da TLA no contexto de consultas a materiais audiovisuais. Especificamente, a análise proposta pelos autores incluiu uma investigação dos pedidos comerciais feitos pelos profissionais da mídia e uma caracterização das sessões, consultas e o conteúdo das palavras registradas nos logs, a fim de caracterizar o comportamento de busca dos usuários e o modo como interagem com o sistema de recuperação. A análise é centrada no nível de palavra e os autores utilizam metadados de catálogo e um tesouro audiovisual estruturado para investigar o conteúdo das palavras de consulta de modo comparativo.

Park e Lee (2013) realizaram análise quantitativa de registros de logs oriundos de um ano de coleta no *National Science and Technology Information Center* (NDSL), totalizando 7.240.000 consultas e 20.700.000 registros. O NDSL é um sistema de RI baseado na *web* liderado pelo Instituto Coreano de Informação em Ciência e Tecnologia, com foco em conteúdo científico e técnico. Ele fornece diversos tipos de informações, incluindo periódicos (impressos / eletrônicos), teses eletrônicas, relatórios, patentes, normas e dados factuais. Nesse cenário, o objetivo geral do trabalho foi obter conhecimento do comportamento dos usuários, a fim de melhorar o sistema de RI. Para tanto, os autores buscaram a compreensão em três níveis: sessão, consulta e termo. O nível de análise da sessão verificou o número de consultas por pesquisador durante o processo de busca de informações; o nível de análise da consulta focou o seu comprimento e a análise a nível de termo trouxe dados estatísticos para identificar as tendências atuais na área de ciência e tecnologia. Os autores fizeram toda a programação em Java e optaram pela utilização de cookies para identificar usuários únicos porque os endereços IP nem sempre

são exclusivos dos usuários e podem ser compartilhados por grandes grupos ou proxies e consideraram como delimitação de sessão o intervalo de 30 minutos.

Laipelt (2015b) evidenciou a análise de logs como alternativa importante para a investigação direta das expressões utilizadas pelos usuários para recuperar informações e aprimorar o sistema de recuperação de informação estudado. Para tanto, desenvolveu uma metodologia para seleção de termos e descritores de tesouros em buscas realizadas pelos usuários do Portal LexML. Assim, cada pesquisa realizada pelos usuários gerou um registro (log) que informou: o número de IP da máquina utilizada para pesquisa; a data; o horário; o tipo de fonte (legislação, jurisprudência, doutrina etc.) e as lexias utilizadas pelos usuários em suas buscas.

Pela análise concentrada nos dados coletados ao longo de 15 dias, a autora identificou 65.536 lexias de buscas de usuários de todas as áreas do Direito, das quais selecionou 2.617 relacionadas ao Direito do Trabalho e ao Direito Previdenciário a partir de critérios de frequência de variação, funcionalidade e ocorrência. Dessa maneira, identificou três categorias de análise de lexias: I) aspectos verbais; II) recursos não verbais e III) aspectos conceituais.

A partir dessa categorização, Laipelt (2015b, p. 105), nos aspectos verbais, analisou as variantes denominativas de forma comparativa, ou seja, as lexias contidas no *corpus* coletado por meio dos logs comparadas ao *corpus* especializado. Os recursos não verbais contemplaram a análise dos registros de pesquisa feitas pela inserção de números de leis, súmulas, decretos e outros documentos jurídicos como palavras de busca. Tais recursos, no entanto, não puderam ser comparados aos tesouros especializados considerados pela autora, por estes não incluírem esse tipo de parâmetro. Por fim, a categoria “aspectos conceituais” foi “constituída por lexias que apresentam como característica as partes da definição de um termo ou uma pergunta” e centrou-se na investigação voltada à “ampliação da rede de remissivas e a identificação de demandas e necessidades informacionais dos usuários” (LAIPELT, 2015b, p. 132) em comparação às lexias oferecidas pelos tesouros e pelo Dicionário de Direito Processual do Trabalho, todos selecionados parâmetros referenciais.

Li *et al.* (2015) propuseram uma atualização de algoritmo de agrupamento transacional baseado em MapReduce para processamento de logs de consulta de *Domain Name System* (DNS) de uma universidade pública da Geórgia (EUA), investigando a taxa de processamento e a qualidade do agrupamento. Os autores coletaram 46 dias de logs, porém selecionaram aleatoriamente 5 dias para a realização da análise, de modo a tentar superar limites relacionados ao volume de dados a serem processados por meio da atualização da proposta.

Kacprzak *et al.* (2017) extraíram e analisaram logs de consulta de quatro portais de dados abertos: três pertencem aos governos nacionais do Reino Unido, Austrália e Canadá, o quarto do Escritório do Reino Unido para estatísticas nacionais. Os registros incluíram mais de 2,2 milhões de consultas, as quais foram analisadas quanto a seu comprimento, distribuição, estrutura e classificação. O objetivo dos autores foi obter um melhor entendimento do comportamento de busca de usuários dos portais, especificamente quanto aos tipos de consultas que eles emitiram; ademais, buscaram também enquadrar as descobertas no contexto mais amplo da pesquisa de conjunto de dados.

A pesquisa feita por Zhang, Huang e Yang (2018) surgiu da observação da desconexão entre o *design* de nível superior dos serviços de informações geoespaciais da China e as demandas reais dos usuários. Nesse sentido, os autores se propuseram a usar a TLA para identificarem padrões de comportamento dos usuários e, assim, melhor atendê-los. O sistema criado foi composto por cinco módulos: o primeiro de coleta de logs; o segundo de comunicação assíncrona; o terceiro de análise e cálculo em tempo real; o quarto de armazenamento e cobrança de dados e, por fim, o quinto módulo de visualização. De acordo com os autores, o lançamento do sistema foi feito de modo integrado ao *Map World* em junho de 2017 com sucesso e indicou que ele pode resolver com eficiência os problemas de coleta de dados de log em tempo real, análise, armazenamento e consulta em tempo real, bem como o armazenamento massivo de dados.

Essa categoria reuniu publicações concentradas na divulgação de modelos de TLA aplicados a diversos contextos, tais como, análises biomédicas, forenses, recuperação de materiais audiovisuais, entre outros. Nas pesquisas, os autores evidenciaram a necessidade de padrões para coleta e análise dos dados para garantir a confiabilidade das proposições feitas a partir da TLA; todavia, a heterogeneidade das variáveis (sistemas, usuários e processos de pesquisa) somada às restrições éticas e de respeito à privacidade são apontadas como desafios a serem superados para a proposição de um método padrão, flexível e adaptável a essa grande diversidade de ambientes digitais de pesquisa.

2.5 Aplicação da TLA para vocabulários controlados

A utilização da TLA para o aprimoramento de vocabulários controlados foi desenvolvida por Vállez *et al.* (2015) e Wu e Fan (2018).

Vállez *et al.* (2015) trouxeram um exemplo de utilização TLA para aperfeiçoar o processo de indexação com vocabulários controlados em periódicos acadêmicos, repositórios e intranet. Os autores realizaram a identificação e o processamento de consultas de usuários para

selecionar palavras candidatas à atualização de vocabulário controlado de forma semiautomática. Para tanto, os pesquisadores analisaram a interação do vocabulário controlado com um *corpus* de texto, através da mineração de palavras-chave utilizadas pelos usuários e de descritores aplicados na indexação. O estudo incluiu uma seleção aleatória de 100 artigos, indexados manualmente no portal Temaria, em HTML, da revista *BiD Textos Universitaris de Biblioteconomia i Documentació*, e eletronicamente pela ferramenta DigiDoc MetaEdit, a qual utiliza o mesmo dicionário de sinônimos do Temaria. A utilização do mesmo dicionário de sinônimos, pelo Temaria e DigiDoc MetaEdit, permitiu a comparação dos resultados da indexação manual e automática.

Os artigos selecionados por Vállez *et al.* (2015) foram indexados com os descritores do *Tesouro de Biblioteconomía y Documentación*, pertencentes ao Instituto de Estudos Documentários em Ciência e Tecnologia (IEDCYT). Os autores identificaram que as consultas de usuários continham, em média, 4,8 palavras-chave, excluindo da identificação consultas que continham 15 ou mais palavras, por serem consideradas inserção de passagem exata de um documento.

Vállez *et al.* (2015) utilizaram o processamento das informações e a análise estatística e o processo de derivação sobre o *corpus* do texto foram realizados por meio da biblioteca de Processamento de Linguagem Natural (NLTK), a qual permitiu o processamento de muitos idiomas, tais como, português, dinamarquês, holandês, inglês, finlandês, francês, alemão e húngaro. Por meio do estudo, os autores concluíram que o grau de correspondência entre as palavras-chave utilizadas pelos usuários e de descritores aplicados na indexação mostrou-se baixo, em torno de 23%. A baixa correspondência evidencia a necessidade de adaptar e atualizar vocabulários controlados embasados em novos termos candidatos para otimização do processo de indexação.

Por sua vez, Wu e Fan (2018) utilizaram a TLA como recurso para examinar o vocabulário livre utilizado pelos usuários para consultas no *Répertoire International de Littérature Musicale* (RILM) durante o mês de dezembro de 2015 e comparar esses achados com os vocabulários controlados usados no banco de dados, de forma a identificar possíveis falhas de recuperação. Os autores avaliaram comparativamente e classificaram as combinações em “perfeitas”, “parciais” ou “nenhuma combinação”, com base em 7.924 consultas, das quais foram selecionadas aleatoriamente 367. Dentre as principais limitações do estudo, Wu e Fan (2018) sublinharam que a análise de logs de pesquisa obtém melhores resultados quando combinada com outros métodos de pesquisa, como a intervisualização, a observação e a experimentação. Além disso, na investigação não foi possível diferenciar as consultas iniciais

das consultas modificadas enviadas pelo mesmo usuário, fato que impediu a observação de seu caminho de busca e revisão. Outro limitante foi a falta de acesso a dados demográficos para associações e formação de grupos de usuários, ao que os autores propuseram realização posterior de entrevistas qualitativas. Mesmo com todas as limitações, a pesquisa contribuiu para promover reflexões acerca do uso de bancos de dados para pesquisas em literatura musical e sobre as abordagens para sua indexação e recuperação.

2.6 Síntese dos resultados da revisão da literatura

Esta seção apresentou os principais resultados da revisão teórica e metodológica sobre a evolução da TLA aplicada ao aperfeiçoamento de sistemas de recuperação da informação. A leitura dos 39 estudos produzidos desde a década de 1960 permitiu observar dificuldades recorrentes, mas também potencialidades do método.

Conforme descrito no referencial teórico, estratégias computacionais têm sido amplamente utilizadas para aprimorar a organização de materiais e favorecer sua recuperação a partir da investigação das palavras de buscas empregadas pelos usuários, sendo a TLA uma estratégia viável, conforme evidenciaram Peters (1993); Jansen (2006); Jansen, Spink, Taksa (2008); Jansen (2008); Penniman (2008); Taksa, Spink, Jansen (2008); Tolle (1983); Jones (1986); Zink (1991); Jones *et al.* (2000); Mahoui, Cunningham (2000); Ozmutlu; Ozmutlu; Spink (2008); Shiri (2009); Park; Lee, (2013), Kacprzak *et al.* (2017); Zhang, Huang, Yang (2018).

Entretanto, tal tarefa permanece um desafio tanto para os cientistas da informação quanto para os cientistas da computação e mesmo para as instituições, uma vez que, além da enorme quantidade de dados produzidos diariamente, ainda precisam considerar divergências interinstitucionais na padronização, grande heterogeneidade de usuários e dificuldades para acesso a grandes volumes de dados que revelem a linguagem natural empregada pelos usuários ao realizarem suas pesquisas (VÁLLEZ *et al.*, 2015; WU, FAN, 2018).

Vale sublinharmos que, a partir da década de 2000, com a expansão dos Repositórios Institucionais, o uso da TLA se destacou a nível internacional com o trabalho de Vállez *et al.* (2015), o qual evidenciou a principal vantagem para a utilização da TLA como sendo o conhecimento direto de palavras de buscas utilizadas pelos usuários, ação que permitiria aperfeiçoar o processo de indexação. Posteriormente, inclusive, Wu e Fan (2018) propuseram a atualização de vocabulários controlados, no âmbito da organização de acervo musical, pela análise das palavras em linguagem natural empregadas pelos usuários.

De maneira geral, a implementação TLA, nos diversos trabalhos publicados, seguiu as etapas descritas por Jansen (2008): coleta, preparação e análise. Na primeira, foram obtidos os campos IP, time e palavras de buscas dos usuários. Na segunda, vários autores realizaram a limpeza dos dados, bem como o seu tratamento por meio de linguagens de programação, com o apoio de bibliotecas de processamento de linguagem natural. Na etapa final, os pesquisadores analisaram as palavras contidas nos logs em relação à frequência; à coocorrência em vocabulários controlados (VÁLLEZ *et al.*, 2015; WU; FAN; 2018); e as palavras empregadas para refinamento de busca ao longo do intervalo considerado como unidade de sessão (JANSEN, 2008; OZMUTLU; OZMUTLU; SPINK, 2008; SHIRI, 2009; PARK; LEE, 2013).

A revisão de literatura permitiu, portanto, comprovarmos a relevância da proposta ora apresentada e esclarecermos os próprios objetivos da pesquisa, posto que o quadro teórico delineado pela investigação exploratória forneceu as informações necessárias para reconhecermos o potencial e as limitações dos logs como fontes informacionais dos usuários de sistemas de informação.

Dado o exposto, confirmamos a viabilidade e relevância da TLA para o cumprimento do objetivo desta Tese de realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais. Sequencialmente, são apresentados os fundamentos teóricos que subsidiaram esta pesquisa.

3 FUNDAMENTOS TEÓRICOS

A busca por meios de melhorar a recuperação de informações surge com sua própria produção e registro. Ao se produzir um texto, a função comunicativa é prerrogativa inalienável que, para se cumprir, necessita da recepção por parte de seu público-alvo. Nesse sentido, estratégias para viabilizar a recuperação de informações ganham impulso e novas características, na medida em que ambientes digitais de armazenamento são desenvolvidos, bem como delineiam-se novos canais de acesso. Assim, nesta seção é exposta a fundamentação teórica sobre a constituição dos RI, suas características estruturais e de funcionamento, política de indexação, vocabulários controlados, sistemas de recuperação de informação, estratégias de busca e recuperação e, por fim, recursos computacionais e a análise de logs, especificamente no que tange a logs de busca como meio de pistas que podem revelar o percurso de pesquisa dos usuários e auxiliar no aprimoramento da indexação, proposição de estratégias de busca e recuperação de informações.

3.1 Constituição dos Repositórios Institucionais

A constituição dos RI está relacionada à observação de necessidades institucionais de organização e divulgação dos serviços prestados, bem como de pesquisas realizadas. De acordo com Rieger (2007), vários pontos precisam ser observados ao se avaliar o contexto de implementação de um RI, desde características de conteúdos até as necessidades dos usuários e os recursos humanos, financeiros e tecnológicos disponíveis.

Os repositórios institucionais têm ganhado espaço como canal de comunicação por permitirem a disponibilização de conteúdos em ambientes digitais, com acesso aberto e com rápida recuperação da informação. A expansão de tecnologias com baixo custo relativo e a crescente necessidade de organização, disponibilização e recuperação de informações têm avançado consideravelmente nas últimas décadas, desde os anos de 1990, com o *Open Archives Initiative* (OAI) e *Open Access* (OA). Tais avanços estiveram associados às discussões sobre o acesso livre das produções acadêmico-científicas (SANCHEZ; VECHIATO; VIDOTTI, 2019). Associativamente, o crescimento e o compartilhamento de dados pela organização de repositórios foram tão substanciais que contextualizar o ponto em que estamos torna-se uma tarefa arriscada, haja vista a ininterrupta, e mesmo vertiginosa, evolução dos estudos relacionados.

Em 2003, Lynch destacou os avanços do pensamento sobre preservação digital e, desde então, muitas são as iniciativas para promover a ampliação do acesso a coleções de artigos

de periódicos, dissertações, teses e uma série de produções digitais, ampliando inclusive os gêneros disponibilizados, com *podcasts*, infográficos e outros conteúdos associados, de modo a fortalecer a rede compartilhada de leitura e pesquisa (VIDOTTI *et al.*, 2016; FERNANDES; VECHIATO, 2017; SOUZA; SOUZA, 2018).

Nesse contexto, a compreensão dos repositórios institucionais nas universidades como “um conjunto de serviços que uma universidade oferece aos membros de sua comunidade para o gerenciamento e a disseminação de materiais digitais criados pela instituição e pelos membros da sua comunidade” (LYNCH, 2003, p. 328, tradução nossa) é oportuna, sobretudo, por se referir a iniciativa como “um conjunto de serviços” oferecidos que envolve desde a administração de materiais (em volume e diversidade sempre crescentes), passando pela organização, a fim de permitir a preservação a longo prazo, até a garantia de acesso e recuperação, mesmo que em suportes e tecnologias diferentes.

Sobre a importância e composição dos RI, Lynch (2003, p. 328, tradução nossa) enfatiza que “um repositório institucional maduro e totalmente realizado” deve conter:

[...] os trabalhos intelectuais de professores e alunos - tanto de pesquisa quanto materiais de ensino - e também documentação das atividades da própria instituição na forma de registros de eventos e performances e da vida intelectual em andamento da instituição. Também abrigará dados experimentais e observacionais capturados por membros da instituição que apoiam suas atividades acadêmicas.

O autor enfatiza o papel dos RI no reconhecimento, na representação e no compartilhamento da vida intelectual e dos conhecimentos produzidos nas universidades. Ademais, salienta a responsabilidade dessas instituições em administrar as riquezas que produzem, tanto para divulgá-las quanto para preservá-las, sendo o RI “o meio pelo qual nossas universidades endereçam essa responsabilidade aos membros de sua comunidade e ao público. É um novo canal para estruturar a contribuição da universidade para o mundo mais amplo” (LYNCH, 2003, p. 329).

Com a mesma compreensão, Farias e Santos (2017, p. 4) ressaltam que “os Repositórios Institucionais dispõem de mecanismos que possibilitam o aumento da visibilidade e da eficácia da preservação das produções científicas dos pesquisadores e das instituições acadêmicas”. Nesse sentido, os RI colaboram ao “fornecer um componente crítico na reforma do sistema de comunicação acadêmica”, visto que têm o potencial de dispor de “indicadores tangíveis da qualidade de uma universidade e demonstrar a relevância científica, social e

econômica de suas atividades de pesquisa, aumentando, assim, a visibilidade, o posicionamento e o valor público da instituição” (CROW, 2002, p. 1, tradução nossa).

Em complemento, Costa (2014, p. 125), ao analisar a literatura científica sobre RI, propôs cinco critérios para referir ao acesso aberto: “1. ser institucionalmente definido; 2. tratar da produção acadêmica da instituição; 3. disponibilizar textos completos dos documentos; 4. ser uma iniciativa de acesso aberto; 5. ser interoperável”. Na mesma compreensão, Dodebei (2009) enfatiza o caráter coletivo e cumulativo, o acesso aberto e a interoperabilidade como características importantes para que os RI cumpram seu papel de coletar, armazenar, preservar e disseminar as produções intelectuais da universidade.

Ao retomarmos a proposição de Lynch dos RI como “um conjunto de serviços”, observamos que, como o próprio autor defende, trata-se de “uma ideia muito poderosa que pode servir como um mecanismo de mudança para nossas instituições de ensino superior e, mais amplamente, para as empresas acadêmicas que elas apoiam”, desde que desenvolvidos de maneira adequada. Visualizando o potencial dos RI, Lynch (2003, p. 336) defendeu o investimento pesado por parte das universidades, porém alertou também que essa implementação deve ser feita refletida e cuidadosamente, “com ampla consulta e colaboração de toda a comunidade do *campus* (com liderança intelectual do corpo docente e a biblioteca trabalhando em parceria) e com total entendimento de que, se tiverem sucesso eles mudarão permanentemente o cenário da comunicação acadêmica”.

Atualmente, constatamos que a compreensão de Lynch já pode ser visualizada de forma mais clara, posto que a importância dos repositórios institucionais, nas universidades, como canais para comunicação científica em diversos gêneros e formatos, para ampliar a disponibilização de conteúdos, “têm sido tema de diversas iniciativas no contexto mundial em ambiente digital de acesso aberto” (VIDOTTI *et al.*, 2016, p. 2).

No Brasil, de acordo com a edição de fevereiro de 2022 do *Ranking Web of Institutional Repositories*, os cinco maiores repositórios institucionais são: Repositório Digital da Universidade Federal do Rio Grande do Sul - LUME; Universidade de São Paulo Biblioteca Digital de Teses e Dissertações; Repositório Institucional da Universidade Federal de Santa Catarina; Alice - Acesso Livre à Informação Científica da Embrapa e Repositório Institucional Unesp. Os repositórios citados figuram também entre os 30 maiores repositórios institucionais mundiais, nas posições 2º, 9º, 13º, 20º e 24º, respectivamente (WEBOMETRICS, 2022).

Para a construção de RI, Leite (2009) destaca três etapas: planejamento, implementação e participação da comunidade. Segundo o autor, no planejamento, a equipe deve considerar os custos, a constituição da equipe e das competências necessárias, além do

levantamento dos principais atores, seus interesses e funções, uma vez que é imprescindível planejar os serviços que serão oferecidos e avaliar as necessidades da comunidade. A etapa de implementação do RI envolve a escolha do software, forma de registrar os metadados, as diretrizes e os procedimentos seguidos para criação de comunidades/coleções, além do gerenciamento dos fluxos de submissão, pós-submissão e depósito de documentos e políticas de funcionamento, além da condução de um projeto-piloto. A fim de assegurar a participação da comunidade, última etapa definida por Leite (2009), desenvolve-se o marketing e o povoamento do repositório, uma política de depósito obrigatório e a avaliação e indicadores de desempenho do RI.

A partir da compreensão dos diversos fatores que precisam ser analisados para a constituição e caracterização de um RI, embora o RI Unesp já esteja implementado e em pleno funcionamento, a sua avaliação contínua envolve a observação dos metadados utilizados para organização dos materiais. Isso porque o desempenho do RI está relacionado à sua usabilidade, facilidade e rapidez de acesso, além do seu impacto na comunidade científica e acadêmica. Sequencialmente, descrevemos a estrutura de funcionamento do RI Unesp, local escolhido para a pesquisa desenvolvida nesta Tese.

3.1.1 Estrutura e funcionamento do RI Unesp

Lançado em 2013, o Repositório Institucional Unesp (RI Unesp) está disponível pelo link: <https://repositorio.unesp.br/> e insere-se no contexto global de universidades comprometidas com a disseminação e a preservação de sua produção intelectual. O RI Unesp tem como missão promover a Universidade em âmbito nacional e internacional, por meio do acesso aberto, da disseminação, da preservação, da valorização e do reconhecimento de sua produção científica, acadêmica, artística, técnica e administrativa.

O RI Unesp utiliza o software *DSpace* e o padrão *Dublin Core* para a definição de um perfil de aplicação de metadados com a finalidade de atender necessidades de representação e divulgação referentes a diferentes tipos de produção científica dos pesquisadores (VIDOTTI *et al.*, 2015).

A implantação do RI Unesp foi possível graças à aplicação dos processos de coleta, conversão e importação automática dos registros referentes à produção científica indexada de pesquisadores vinculados à Universidade em diferentes locais. Os registros da produção científica dos pesquisadores da Unesp foram inicialmente coletados nas bases de dados *Web of Science*, *Scopus* e em periódicos da *Scientific Electronic Library Online* (SciELO) (ASSUMPÇÃO *et al.*, 2014). Na sequência, a coleta teve continuidade na *PubMed*, Currículos

da Plataforma Lattes e registros MARC 21 (*Machine Readable Cataloging*) importados da Biblioteca Digital de Teses e Dissertações (C@THEDRA) e também da Biblioteca Digital de Trabalhos de Conclusão de Curso (C@PELO) da Unesp (VIDOTTI *et al.*, 2016).

Atualmente, o Repositório Unesp realiza, além da coleta automática, o arquivamento sob demanda e o autoarquivamento. Na modalidade de arquivamento sob demanda, é armazenada expressiva quantidade de recursos e documentos digitais de diferentes tipos de produção além da científica e acadêmica, tais como, a produção administrativa, artística, comemorativa e técnica da universidade. Na modalidade de autoarquivamento, é disponibilizada aos pesquisadores e aos autores de teses, dissertações e trabalhos de conclusão de curso, interface de preenchimento de metadados de descrição física e temática e submissão do original.

Em observação do RI Unesp, verificamos que a representação temática por palavras-chave adotada é a linguagem natural dos autores, disponibilizada nos registros de produção científica coletados, tanto no arquivamento sob demanda quanto no autoarquivamento pelos autores de teses e dissertações. Durante o autoarquivamento, o usuário tem a opção de buscar palavras-chave no Tesouro Unesp, fonte terminológica positiva e confiável, para preenchimento dos metadados, porém não se trata de uma exigência, ou seja, o autor tem a faculdade e não a obrigação de utilizar o vocabulário controlado para representar sua produção.

Adicionalmente, nenhum controle de vocabulário é realizado com as palavras-chave coletadas dos metadados que estão disponíveis em uma lista alfabética para consulta dos usuários durante a busca. A ausência de controle de vocabulário resulta em uma lista alfabética com variações de uma mesma palavra (singular ou plural, digitação com maiúscula ou minúscula, uso de aspas e outros sinais como hífen e assim por diante) e, apesar da interface de busca conter possibilidades de aplicação de diversos “filtros” para refinamento, de modo geral, a ausência de estratégia para controle de vocabulário pode dificultar a recuperação e, consequentemente, prejudicar a efetividade do RI.

Para contribuir com o processo de organização de documentos, em 2016, foi elaborada a *Política de Gestão do Repositório Institucional Unesp - Regulamento Interno*, de maneira a orientar o arquivamento de documentos no RI Unesp. Assim, de acordo com a política de gestão estabelecida, para integrar o RI é necessário atender alguns critérios, a saber:

- I - possuir entre seus criadores ao menos uma pessoa vinculada à Universidade;
- II - estar em formato digital;
- III - ser resultado de atividade científica, acadêmica, artística, técnica ou administrativa realizada nas dependências da Universidade ou em seu nome;

IV - estar completo, pronto para publicação ou já ter sido publicado e com a cessão de direitos patrimoniais, quando for o caso;

V - estar categorizado como:

- a) artigo publicado em periódico científico, em sua versão publicada ou aceita para a publicação;
- b) trabalho completo ou resumo publicado em anais de congresso, conferência, encontro, simpósio ou qualquer outra reunião de natureza técnico-científica;
- c) livro ou capítulo de livro;
- d) tese de doutorado, tese de livre-docência, dissertação de mestrado ou trabalho de conclusão de curso de graduação ou de especialização;
- e) notação musical;
- f) imagem;
- g) material cartográfico;
- h) gravação de som ou gravação de vídeo;
- i) software;
- j) patente;
- k) outra categoria a ser estabelecida pelo Grupo Gestor do Repositório Institucional Unesp (UNESP, 2016, p. 2-3).

De modo a ampliar a diversidade de materiais disponibilizados e atender aos interesses e necessidades da comunidade, o Artigo 6º dispõe que:

Documentos de tipos não listados no inciso V, do artigo 5º, deste Regulamento, poderão ser arquivados no Repositório Institucional Unesp, mediante encaminhamento de ofício para o Grupo Gestor do Repositório Institucional Unesp, juntamente com a lista de documentos a serem arquivados, que deverá ter sido previamente enviada ao gestor da comunidade a que pertence, que irá emitir o parecer (UNESP, 2016, p. 3).

Por fim, quanto à extensão, em acesso realizado em 18 de janeiro de 2022, às 16h35, visualizamos 29 tipos de materiais, desde artigos, dissertações e teses até partituras e *podcasts*. O número dos arquivos muda constantemente pela inserção contínua de materiais, contudo, na referida data e horário, o acervo do RI era composto por 172.960 documentos, como exposto na Figura 2:

Figura 2. Tipos de documentos disponibilizados no RI Unesp

Artigo: 95415	Dissertação de mestrado: 30038	Tese de doutorado: 16127
Trabalho de conclusão de curso: 9890	Trabalho apresentado em evento: 8902	Resumo: 6579
Resenha: 1984	Editorial: 754	Carta: 643
Capítulo de livro: 549	Livro: 485	Patente: 394
Podcast: 275	Errata: 254	Nota: 209
Tese de livre-docência: 166	Boletim: 78	Revista: 74
Dado de pesquisa: 44	Jornal: 34	Relatório: 26
Data paper: 13	Plano de gestão de dados: 9	Biografia: 7
Objeto educacional: 6	Partitura: 2	Regulamento: 1
Tese de Livre-Docência: 1	Vídeo: 1	

Total de registros: 172960

Fonte: Repositório Institucional Unesp (18 jan. 2022).

Frente à grande variedade e ao volume de materiais disponibilizados pelo RI Unesp, como demonstra a Figura 2, a forma como tais documentos são organizados, no momento da inserção na base de dados, é determinante para que possa ser executada a sua recuperação quando for de interesse dos usuários. A propósito, Lancaster (2004) destaca que os termos atribuídos aos documentos pelos indexadores funcionam como pontos de acesso para que os usuários possam localizar e recuperar informações conforme seus interesses. Em contexto informatizado, a captação de relações entre termos e conceitos se torna ainda mais fundamental para a recuperação, posto que a quantidade e a heterogeneidade, tanto de materiais quanto de usuários, se avolumaram substancialmente.

Nesse contexto, conforme ISO (*International Organization for Standardization*) 25964-1 (2011, p. 6, tradução nossa), os metadados são definidos como identificadores de atributos de um documento “[...] normalmente usados para dar suporte a funções como localização, descoberta, documentação, avaliação e/ou seleção”. No que se refere ao padrão *Dublin Core*, os metadados exigidos: “além do identificador, data, fonte e idioma já descritos, estes incluem contribuinte, cobertura, criador, descrição, formato, editora, relação, direitos,

assunto, título e tipo. Tudo exceto idioma e identificador são opcionais” (ISO 25964-1, 2011, p. 107).

Com a compreensão de que os metadados têm a função de descrever e identificar um documento, a fim de facilitar o processo da recuperação da informação, exemplificamos, após busca de artigo com uso do descritor “Ciência da Informação” e acesso aos metadados completos, a Figura 3 extraída do RI Unesp:

Figura 3. Metadados do RI Unesp

Uma perspectiva interacional da articulação dos profissionais da informação com o campo da Ciência da Informação

dc.contributor.author	Almeida, Carlos Cândido de [UNESP]
dc.date.accessioned	2015-02-24T13:57:45Z
dc.date.available	2015-02-24T13:57:45Z
dc.date.issued	2011
dc.identifier	http://www.dgz.org.br/fev11/Art_06.htm
dc.identifier.citation	Datagramazero, v. 12, n. 1, 2011.
dc.identifier.issn	1517-3801
dc.identifier.uri	http://hdl.handle.net/11449/115087
dc.description.abstract	Discusses the concepts occupational knowledge and scientific knowledge, taking into account the existence of the field of Information Science formed by sub-professional and scientific fields. Displays incursions in profession sociology and systems theory of profession, in

Fonte: Repositório Institucional Unesp (18 jan. 2022).

Na imagem, cada linha traz o metadado em negrito seguido pela informação correspondente, a qual permite a identificação detalhada do documento. Destaca-se, atualmente, o padrão *Dublin Core*, utilizado em ambientes *web* para representar, identificar, localizar, disseminar e acessar objetos digitais (GONÇALEZ *et al.*, 2018).

O metadado *Dublin Core* utiliza recursos de Linguagem de Marcação Extensível (XML, do inglês *Extensible Markup Language*) e a Estrutura de Descrição de Recursos (RDF, do inglês *Resource Description Framework*), usados para rotulagem, busca e navegação nos sistemas, principalmente com as ferramentas de pesquisa. Sua estrutura é composta de 15 metadados: Colaborador, Cobertura, Criador, Data, Descrição, Formato, Identificador, Linguagem, Publicador, Relação, Direitos, Fonte, Assunto, Título e Tipo (PÖTTKER; FERNEDA; MOREIRO-GONZÁLEZ, 2018).

Feitas as considerações sobre a estrutura e o funcionamento do RI Unesp, para melhor cumprir seu papel, os acervos, sejam físicos ou digitais, precisam organizar sua política de indexação de tal modo que a visão ampliada do contexto implicado, no gerenciamento de sistemas de recuperação da informação, encontre meios de se concretizar em procedimentos

operacionais claros, os quais garantam a sua qualidade e uniformidade. Nesse sentido, é fundamental aos manuais objetividade e estratégias para convergência de vocabulários dos indexadores e dos usuários.

3.2 Indexação

A organização e recuperação da informação são fundamentais no processo de divulgação do conhecimento científico e, por sua vez, a divulgação do conhecimento produzido fundamenta e impulsiona novas descobertas. Nesse contexto, as bases de dados e os repositórios vinculados a instituições de pesquisa desempenham papel essencial. Contudo, para que o conhecimento produzido seja efetivamente aproveitado, a organização dos dados deve ser executada de maneira que a recuperação ocorra de forma rápida e precisa. Para tanto, o processo de indexação é fundamental, a fim de representar e recuperar informações de modo a facilitar e agilizar o acesso, haja vista o grande e crescente número de publicações que requerem análise e catalogação.

No final da década de 1970, Borko e Bernier (1978, p. 8) definiram indexação como a ação de “analisar o conteúdo informacional dos registros do conhecimento e expressar o conteúdo informacional na linguagem do sistema de indexação”. Posteriormente, García Gutiérrez (1984, p. 105) citou a relação entre indexar e recuperar documentos, afirmando que indexar é “descrever o conteúdo de documentos ou demandas documentárias para possibilitar a elaboração de estratégias de recuperação mediante conceitos ou assuntos”.

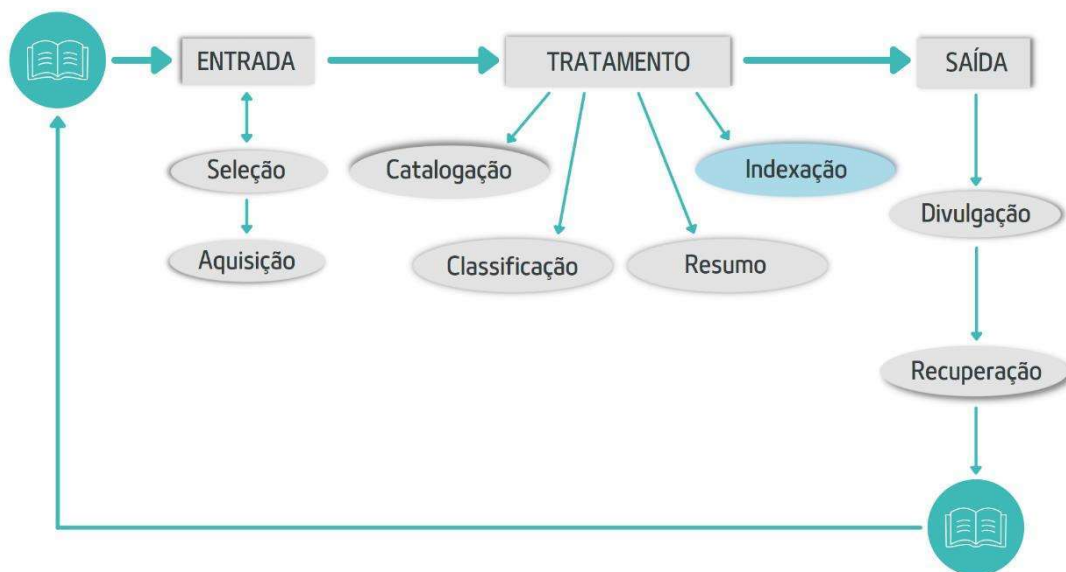
A compreensão da indexação como parte de um processo de mediação é ressaltada por Vieira (1988), para quem a indexação pode ser compreendida como um processo de condensação de informação, o qual viabiliza o encontro de materiais pelos usuários. Na mesma compreensão, Lancaster (2004, p. 6) salienta a conversão da análise conceitual em termos representativos durante a indexação de assuntos e a redação de resumos, compreendidas como atividades “[...] intimamente relacionadas, pois ambas implicam a preparação de uma representação do conteúdo temático dos documentos”.

Do ponto de vista normativo, o conceito de indexação tem sido compreendido como um processo de análise intelectual do assunto de um documento, com o propósito de identificar conceitos-chave capazes de possibilitar a recuperação da informação pelos usuários (ISO 25964-1, 2011, p. 5).

A indexação tem início, conforme Gil Leiva (2008, p. 55), com a contextualização do material como parte do processo técnico documental que, segundo o autor, compreende “[...] o conjunto de operações referentes à seleção, à aquisição, ao registro e ao tratamento dos

documentos com a finalidade de possibilitar seu armazenamento, recuperação e sua posterior difusão”. Nessa perspectiva, a indexação é entendida como parte dos processos de análise de estrutura e conteúdo, os quais compõem o tratamento documental. A Figura 4 exibe as etapas do processo técnico documental constituído por um conjunto de operações direcionadas à seleção, à aquisição, ao registro e ao tratamento dos documentos com o objetivo de viabilizar seu armazenamento, recuperação e divulgação:

Figura 4. Processo técnico documental de entrada e saída de informações



Fonte: Gil Leiva (2008, p. 59), tradução nossa.

A análise documental é compreendida, portanto, como um processo que visa o controle e a representação resumida dos dados formais e conteudistas de um documento. A partir da entrada de materiais selecionados ou adquiridos, o tratamento envolve a catalogação, isto é, a formulação de lista ordenada das entradas bibliográficas. Sequencialmente, há a classificação feita em linguagem documental pautada na representação estruturada de um ou vários domínios do conhecimento em classes cujas relações também são representadas. A indexação ocorre como operação documental voltada à representação por meio de linguagem documental ou natural dos dados resultantes da análise do conteúdo de um documento ou de uma demanda de informação e o resumo consiste no produto secundário correspondente à uma representação abreviada da macroestrutura do texto original (GIL LEIVA, 1997).

Realizada a etapa de tratamento, a qual conecta a entrada e a saída de informações, a divulgação dos documentos produzidos se efetiva pela transmissão ao usuário das informações

que necessita ou, ao menos, pela viabilização para que as obtenha quando consultar o banco de dados.

Em complemento, já no século XXI, o vínculo entre a indexação e a recuperação é reforçado por Cleveland (2001, p. 97, tradução nossa) quando o autor atribui à indexação o esforço de “identificar informação numa entidade de conhecimento (que seja texto ou não) e organizá-la para que esteja disponível num sistema de recuperação”, isto é, todo o esforço empreendido pelo indexador deve convergir para uma melhor recuperação.

Na “Apresentação” da obra *Política de indexação* (2012a), Fujita destaca a relação de causa e efeito existente entre os processos de indexação e recuperação da informação, posto que a especificidade dada à primeira repercute na especificidade desejada para a segunda. A autora enfatiza a necessidade de que os profissionais envolvidos no processo de indexação reflitam sobre a “elaboração de uma política de indexação que contemple a complexidade do contexto e da cultura organizacional dos sistemas de informação no que se refere à recuperação da informação” (FUJITA, 2012a, p. 14).

Destacada a importância da indexação para a recuperação efetiva de documentos, Gil Leiva e Fujita apontam, conforme a literatura, a existência de três perspectivas para a indexação de documentos: diferenciadas pelo foco nos documentos, nos usuários ou nos domínios (GIL LEIVA; FUJITA, 2012).

Ao citarem as pesquisas realizadas por Soergel (1985) e Lancaster (1991), Gil Leiva e Fujita afirmam que a indexação centrada nos documentos visa a sua descrição fiel, contudo não consideram o contexto ou necessidades informacionais dos potenciais usuários.

A respeito da indexação centrada no usuário, os estudos de Albrechtsen (1993) e Fidel (1994) são mencionados por se referirem a um tratamento da linguagem de indexação pensando no contexto dos potenciais usuários dos documentos, ou seja, são selecionados termos representativos, porém tais termos não são diretamente registrados na indexação, antes disso eles são convertidos em termos controlados que considerem o conhecimento e as necessidades de informação dos usuários.

Por fim, a indexação centrada no domínio é a mais completa por contemplar e expandir os elementos considerados na indexação centrada nos documentos e na indexação centrada nos usuários. Gil Leiva e Fujita (2012, p. 71) destacam que a indexação centrada no domínio foi proposta por Jens-Erik Mai (2005) e consistiu em “um método baseado no conhecimento profundo da organização (histórico, objetivos, metas, pessoas e relacionamentos, fluxos de informação etc.)” o que permite a subdivisão da indexação em quatro processos: “análise do domínio, das necessidades dos usuários, papéis desempenhados ou adotados pelos indexadores

e, finalmente, a análise do documento, tendo em conta os elementos acima referidos” (GIL LEIVA; FUJITA, 2012, p. 71).

Adicionalmente, sobre políticas de indexação, Fujita (2012b, p. 18) destaca aspectos relacionados à entrada e saída de documentos. A entrada envolve a seleção pautada nas necessidades da comunidade de usuários; a organização e representação dos documentos adquiridos; o armazenamento dos documentos no sistema e a construção de bases bibliográficas como suporte para a recuperação da informação. A saída de documentos contempla o atendimento às necessidades dos usuários “[...] pela elaboração de estratégias de busca de informação que devem ser compatíveis com a indexação e linguagem de indexação adotada pelo sistema [...]”, sendo “[...] a difusão seletiva da informação planejada entre profissionais e pesquisadores ou usuários ou, ainda, pelos próprios usuários” (FUJITA, 2012b, p. 18).

Pelo exposto, é inegável que a adoção de uma boa política de indexação é essencial e pressupõe uma compreensão gerencial complexa, pois envolve diversas ações relacionadas a todo o processo de tratamento de documentos, desde a seleção para integrar uma base de dados até a recuperação requerida pelo usuário desta base.

Com essa compreensão das políticas de indexação como “linhas mestras” para a tomada de decisões (ALMEIDA, 2005), a qualidade e uniformidade das atividades relacionadas dependem de manuais claros e detalhados.

Além das diferentes perspectivas e da exigência de políticas de indexação detalhadas, o avanço dos sistemas de informação permitiu que as ações requeridas pela indexação pudessem ser realizadas tanto pelo homem (indexação manual) quanto por softwares (indexação automática) (VIEIRA, 1988; GIL LEIVA, 2017). Com o incremento crescente do número de publicações, a indexação manual tornou-se inviável para a organização de grandes bases de dados, seja pela morosidade de análise, pelo custo elevado ou pela falta de uniformidade na escolha dos termos representantes dos documentos. Nesse contexto, os sistemas de indexação automática se constituíram como ferramentas importantes para a redução de custos operacionais e adoção de padrões programados para inserção de materiais em bases de dados (ROBREDO, 1991; CORRÊA; BAZÍLIO, 2017).

3.2.1 Indexação nos Repositórios Institucionais brasileiros

A busca por aprimorar a representação e recuperação das produções acadêmicas indexadas nos repositórios institucionais é uma constante, tendo em vista o desejo de ampliação de sua visibilidade. Com esse objetivo, a adoção de vocabulários controlados é uma ferramenta importante para a uniformização de termos e eliminação de ambiguidades.

Em mapeamento realizado por Fujita e Tolare (2019), foram identificados 107 repositórios brasileiros, dos quais 91 puderam ser acessados, apurando-se que 62 institucionais, 20 centrais e 9 temáticos. Quanto à utilização de controle de vocabulário, nos repositórios acessados, as pesquisadoras observaram três categorias: I) vocabulários controlados do tipo de modelos de relacionamento (tesauros, redes semânticas e ontologias), usado por 6 repositórios; II) uso de modelos semelhantes a metadados (arquivos de autoridades, diretórios, dicionários geográficos), por 74 repositórios e III) uso de vocabulário não controlado ou livre, por 11 repositórios (FUJITA; TOLARE, 2019).

Conforme Fujita (2012, p. 13), “[...] a recuperação da informação é o efeito e o objetivo da indexação [...]”, sendo assim, a qualidade de indexação por assunto é fundamental para a qualidade do repositório, pois dela depende o atendimento das necessidades dos usuários.

A propósito, Tartarotti e Dal’Evedove (2021, p. 586) destacam que:

Avaliar o grau de consistência da representação da informação por assunto, formalizada a partir da atribuição de termos aos conceitos derivados do processo analítico de um documento constitui-se uma prática essencial, notadamente quando ancorada nos ambientes digitais de característica complexa e de acesso multidimensional. Os resultados advindos da Avaliação da Indexação colaboram com a gestão do tratamento temático de conteúdos informacionais e melhoria contínua na qualidade do metadado assunto de repositórios institucionais.

Ponderando a complexidade dos ambientes digitais e o objetivo de viabilizar aos usuários um ambiente de uso intuitivo, uma política de indexação bem definida se constitui como um guia para a qualificação das ações profissionais, posto que uniformiza e amplia o potencial de resultados positivos associados à qualidade dos metadados. A estruturação de termos em hierarquias ou com agregações de navegação podem colaborar para a efetividade do acesso multidimensional oferecido pelo repositório. Ademais, a qualidade do metadado é essencial, uma vez que, a partir dele, as relações serão estabelecidas atendendo a especificidades na busca que gerarão recuperações igualmente específicas (FUJITA, 2012b).

Gil Leiva, Katahira e Fujita (2021) e Tartarotti e Dal’Evedove (2021), em pesquisas realizadas no RI Unesp, verificaram a total correspondência dos registros dos assuntos e os assuntos atribuídos ao artigo original. Essa correspondência indica que os registros são carregados juntamente com os demais metadados no momento da importação dos registros bibliográficos e que não há uma política de indexação que forneça diretrizes e normas para o processo de indexação neste sistema de recuperação da informação.

Ao fixar uma política de indexação e utilizar controle de vocabulário, o repositório pode ampliar sua visibilidade por considerar relações hierárquicas, de equivalência e associativas, uma vez que todo o processo é pensado e registrado para atender as necessidades dos diferentes usuários. Em complemento, um dos principais problemas da falta de padronização da indexação e da ausência de controle de vocabulário em um RI está associada a falhas na representação da informação, pois “[...] os usuários representam conteúdos com palavras que usam e/ou acreditam que entendem, sendo assim algumas palavras podem fazer sentido apenas a quem as elencou e nem sempre para a comunidade ou grupo de pessoas que utilizam o sistema” (SANTOS; NEVES, 2018, p. 1005).

Portanto, para ampliar a visibilidade do RI Unesp, a elaboração de uma política de indexação é fundamental, uma vez que fixará diretrizes a serem seguidas para o aprimoramento do processo analítico e consequente melhoria da representação da informação por assunto. Nesse esforço para melhoria contínua, a utilização de vocabulários controlados é uma importante ferramenta para padronização e busca de uniformidade na representação.

3.3 Vocabulário controlado

De acordo com a definição trazida pela NISO (2010, p. 5), vocabulário controlado corresponde a uma lista de termos disponibilizada a partir de vocabulário validado por registro de autoridade, de modo que cada termo tenha definição “inequívoca e não redundante”. Os termos são selecionados da linguagem natural para inclusão em um vocabulário controlado, cujo objetivo é descrever conteúdos por meio da atribuição de termos para representar metadados associados aos objetos de conteúdo que pretendem representar.

Em complemento, Barité Roqueta (2014, p. 99) define controle de vocabulário como: “[...] um conjunto de técnicas e procedimentos que são exercidos sobre expressões linguísticas chamadas 'termos', que representam conceitos em um dado contexto temático”. Por controlar sinônimos, diferenciar homógrafos, eliminar ambiguidades e redundâncias e, consequentemente, aumentar a consistência da descrição dos objetos, os vocabulários controlados colaboram para a eficácia dos sistemas de armazenamento e recuperação de informações; inclusive os RI.

Conforme a ISO 25964-1 (2011), os vocabulários controlados são projetados para aplicações nas quais é útil identificar cada conceito com um rótulo consistente. O gerenciamento de um vocabulário visa a desambiguação e padronização na representação conceitual. Tal controle é obtido por ações como a distinção entre homógrafos, de modo que cada um tenha apenas um significado, e pela escolha, dentre um conjunto de sinônimos ou

quase-sinônimos, daquele termo que deve ser preferido para uso na indexação. O objetivo dessas restrições é aumentar a probabilidade de indexadores e pesquisadores escolherem o mesmo termo para rotular um conceito específico.

Já em 2010, a NISO Z39.19 (2010) estabeleceu cinco propósitos para os vocabulários controlados: I) tradução; II) consistência; III) indicação de relações; IV) identificação e navegação; e, V) recuperação. A tradução se relaciona à conversão da linguagem natural de autores, indexadores e usuários em termos que possam ser usados de maneira inequívoca na indexação e que favoreçam a recuperação. A consistência se associa à uniformidade do formato e atribuição de termos, enquanto a indicação de relações estabelece vínculos semânticos entre os termos. Identificação e navegação fornecem hierarquias consistentes e claras em um sistema de navegação amigável, que auxilia os usuários a localizarem os conteúdos desejados e a recuperação auxilia na localização dos conteúdos:

Um dos principais objetivos do controle de vocabulário é garantir que cada conceito distinto seja descrito por uma única forma linguística. Se existirem várias formas, estes *devem* ser controlados ou regularizados para que as informações ou o conteúdo fornecido a um usuário não seja distribuído pelo sistema em vários pontos de acesso, mas reunidos em um só lugar. Eliminando a ambiguidade e compensando a sinonímia, o controle de vocabulário garante que cada termo tenha apenas um significado e que apenas um termo possa ser usado para representar um determinado conceito ou entidade (NISO, 2010, p. 12, tradução nossa).

No que concerne à estrutura, ainda de acordo com a NISO (2010), os vocabulários controlados podem ser organizados de acordo com a percepção de diferentes tipos de relacionamento: Listas; Anéis de sinônimos; Taxonomias e Tesouros, conforme exposto no Quadro 3.

Quadro 3. Tipos de vocabulários controlados

Estrutura dos vocabulários controlados	
Tipo	Definição
Listas	Uma lista (às vezes também chamada de lista de seleção) é um conjunto limitado de termos organizados como uma lista alfabética simples ou de alguma outra forma logicamente evidente. Usos: as listas são usadas para descrever aspectos de objetos de conteúdo ou entidades que têm um número limitado de possibilidades. Os exemplos incluem geografia (país, estado, cidade), idiomas (inglês, francês, sueco) ou formatos (texto, imagem, som).

Anéis de sinônimos	Os anéis de sinônimos desempenham um papel diferente dos outros tipos de vocabulários controlados, pois não são usados durante o processo de indexação, tendo aplicação apenas à recuperação. Usos: os anéis de sinônimos garantem que um conceito, que pode ser descrito por vários termos sinônimos ou equivalentes, será recuperado se algum dos termos for usado em uma pesquisa.
Taxonomias	Uma taxonomia é um vocabulário controlado que consiste em termos preferenciais, todos os quais estão conectados em uma hierarquia ou poli-hierarquia. Considera a listagem ordenada em grupos ou categorias com caracterização específica. Usos: relações de família, espécie, áreas e subáreas, entre outras.
Tesauros	Um tesauro é um vocabulário controlado organizado em uma ordem conhecida e estruturado de modo que as várias relações entre os termos sejam exibidas de forma clara e identificada por indicadores de relacionamento padronizados. Os indicadores de relacionamento devem ser empregados reciprocamente. Usos: descrição de documentos diversos com objetivo de facilitar a busca, localização e acesso subsequentes às informações.

Fonte: elaborado pelo autor a partir das normas NISO (2010); ISO 25964 – Parte 1 (2011); ISO 25964 – Parte 2(2013).

Além da adoção de um vocabulário controlado adequado para melhor representação dos objetos de conteúdo, como a linguagem é viva, um vocabulário não está estático no tempo. O surgimento de conceitos, terminologias e usos na linguagem natural exige que também os vocabulários controlados sejam atualizados, com a inclusão de termos novos e exclusão de termos obsoletos, para viabilizar a garantia do usuário. O processo de seleção de termos para inclusão em vocabulários controlados envolve a consulta de várias fontes de palavras e frases, além de critérios para garantia literária (a linguagem natural usada para descrever objetos de conteúdo), garantia do usuário (a linguagem dos usuários) e garantia organizacional (as necessidades e prioridades da organização) (NISO, 2010).

Como a elaboração e a atualização de vocabulários controlados exigem um processo analítico amplo e profundo, a atualização periódica é considerada importante, mas também complicada, uma vez que há o risco de “[...] perdas, distorção das palavras do autor, atribuição de termos errôneos, alto custo e pode haver dificuldades na interoperabilidade de informações por causa da incompatibilidade entre linguagens controladas” (FUJITA; TOLARE, 2019, p. 434).

Por manter em perspectiva a utilidade atrelada à capacidade do vocabulário controlado de atender seus usuários, tanto os que o buscam para descrever objetos conteudistas quanto os que procuram apoio para localização de informações, a realização de testes e avaliações

periódicas e contínuas é essencial para garantir que os termos controlados tenham a máxima utilidade representativa possível.

Em adição, para a máxima utilidade possível, a formulação de vocabulários controlados também considera categorias de usuários, de maneira que seja possível ponderar sobre as necessidades específicas de cada classe. A NISO (2010) refere três categorias: I) pessoal de manutenção de vocabulário controlado (usuários com experiência em indexação e especialistas no domínio do assunto de vocabulário controlado); II) indexadores e pesquisadores especializados (usuários com experiência em indexação, informações de recuperação e uso de vocabulários controlados, indexadores, provavelmente, têm domínio do assunto, porém pesquisadores podem não ter); III) usuários finais (possivelmente não terão experiência no vocabulário técnico, na percepção das relações complexas envolvidas na recuperação de informação e nos paradigmas de notação de vocabulário controlado).

Na sequência, apresentamos o Tesauro Unesp como vocabulário controlado considerado parâmetro para atualização por meio da análise de logs extraídos do RI Unesp.

3.3.1 Tesauro Unesp

Como já apresentado no Quadro 3, que sintetiza os tipos de vocabulários controlados quanto à sua estruturação e ao uso, a elaboração de um tesauro está condicionada à percepção das relações recíprocas entre os termos com usos voltados à descrição de documentos a fim de facilitar os processos de busca, localização e acesso posteriores. A principal norma sobre construção de tesouros é a ISO 25964 - *Information and documentation: Thesauri and interoperability with other vocabularies* (Informação e documentação - Tesouros e interoperabilidade com outros vocabulários) (ISO 25964, 2011).

Assim, o Tesauro Unesp é um vocabulário controlado construído como sistema de organização do conhecimento científico composto de palavras-chave (descritores) relacionadas semanticamente por áreas de conhecimento. O objetivo é que a comunidade de usuários utilize o Tesauro como ferramenta para ampliar os níveis de especificidade e exaustividade necessários para recuperar informações sem ambiguidades e redundâncias.

De acordo com informações contidas na página do próprio Tesauro, a construção e a manutenção do Tesauro Unesp são realizadas desde 2013 pelo Grupo de Linguagem da Unesp composto por catalogadores das bibliotecas e pesquisadores da área de Organização do Conhecimento, em conformidade com a Portaria CGB/Unesp nº 01 de 16 de maio de 2017.

Inicialmente, houve a internalização da base de registros de autoridades de assuntos da Lista de Cabeçalhos de Assuntos da Rede Bibliodata (LCARB) na base de dados Aleph do

catálogo Athena, ação que colaborou com a compatibilização dos registros de autoridades de assuntos com três outras linguagens: *Library Congress Subject Headings* (LCSH), Terminologia da Biblioteca Nacional (TBN) e *Medical Subject Headings* (MeSH) para áreas da Saúde. Posteriormente, o Grupo de Linguagem da Unesp foi criado e recebeu capacitação para substituir gradativamente a LCARB (SANTOS; MOREIRA; FUJITA, 2018).

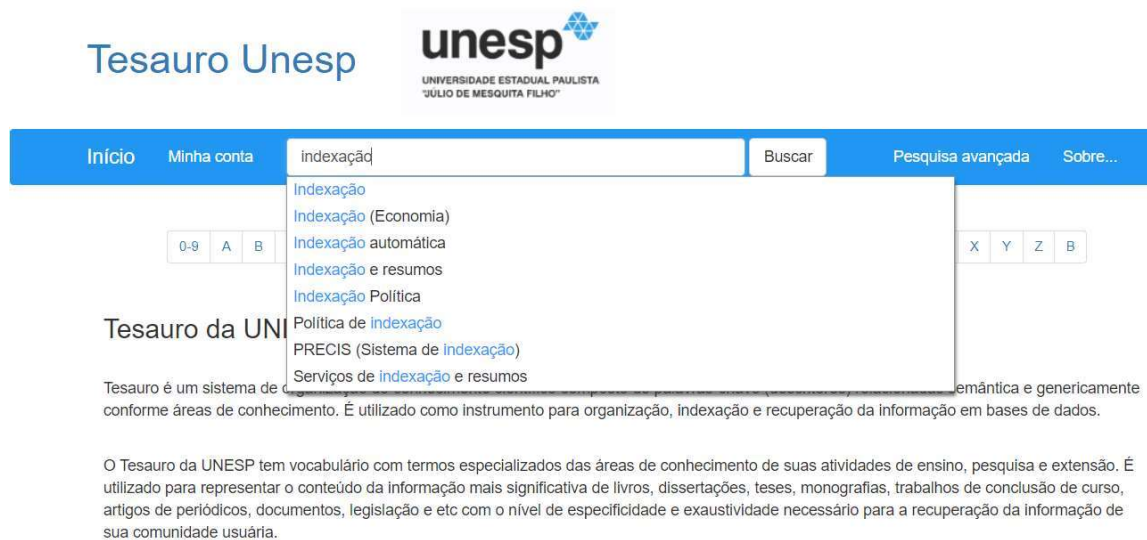
Em 2017, com a escolha do software TemaTres para gestão e manutenção da linguagem Unesp, foi possível a importação dos registros de autoridade compatibilizados (SANTOS; MOREIRA; FUJITA, 2018). Em consulta realizada em 24 de março de 2022, o Tesouro Unesp estava disponível junto à interface de busca integrada das bases de dados da Unesp e da busca do portal Parthenon, contando com 89.588 termos.

A título de comparação, ao pesquisar vocabulários controlados, Vállez *et al.* (2015) analisam mais detidamente o vocabulário controlado *Library and Information Science Thesaurus*, o qual continha 7.581 descritores no momento da pesquisa. Já Laipelt (2015), ao propor uma metodologia para seleção de termos equivalentes e descritores em tesouros, apontou o Vocabulário Controlado Básico (VCB) como a linguagem documental adotada na Rede Virtual de Bibliotecas (RVBI), sendo na época composto por 9.925 descritores. Em comparação com os vocabulários controlados utilizados como referências nas pesquisas de Vállez *et al.* (2015) e Laipelt (2015), a quantidade de termos do Tesouro Unesp é maior em 91,5% e 88,9%, respectivamente.

A importância de um vocabulário controlado bem estruturado é destacada por Fujita (2017) ao defender que a disponibilização do Tesouro Unesp para consulta de suas palavras-chave é parte fundamental da Política de Indexação das bibliotecas Universitárias da Unesp, colaborando para o aprimoramento da recuperação da informação por seus usuários durante a busca por assuntos. Além disso, o Tesouro aparece como hiperlink sugerido aos autores para cadastro de palavras-chave no autoarquivamento de produções acadêmicas no RI Unesp.

Ao acessar a página do Tesouro Unesp e digitar algum termo de busca, automaticamente são sugeridos termos associados, como ilustra a Figura 5:

Figura 5. Busca por termos no Tesauro Unesp



Fonte: Tesauro Unesp (24 mar. 2022).

No exemplo, observamos que os termos associados trazem maior especificidade ao termo buscado e, ao clicar em “Indexação Política” ou “Política de indexação”, obtém-se o resultado exposto na Figura 6:

Figura 6. Caracterização dos termos do Tesauro Unesp



Fonte: Tesauro Unesp (24 mar. 2022).

Na imagem, verificamos que há um menu tríplice com registro do termo, notas e metadatos. Ao clicar em metadatos, o usuário tem acesso às datas de criação e aceite do termo, número de termos relacionados, arquivos exportáveis em diversos formatos e possibilidades de

buscas cruzadas na Wikipédia, Google, Google Acadêmico, Google Imagens e Google Books, como exemplificado na Figura 7:

Figura 7. Formatos dos metadados do RI Unesp



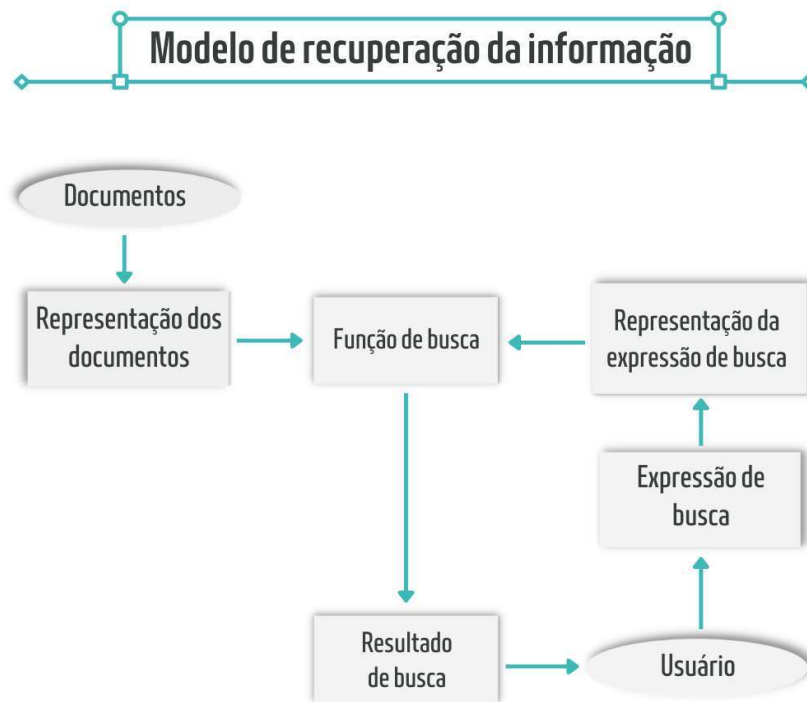
Fonte: Tesauro Unesp (24 mar. 2022).

Sublinhamos que o fornecimento dos metadados em diversos formatos corresponde a descrições das palavras utilizando diferentes linguagens computacionais (BS8723-5, DC, MADS, SKOS-Core, VDEX, XTM, Zthes, JSON e JSON-LD), o que contribui para a interoperabilidade.

Por fim, destaca-se que o controle de vocabulário tem função essencial para a organização do conhecimento, uma vez que contribui para o tratamento dos termos com a depuração, normalização da terminologia e estabelecimento de vínculos para recuperação eficaz de documentos dentro das bases de dados desejadas. Além disso, como o controle de vocabulário se estrutura para exibir de forma clara e padronizada os relacionamentos entre os termos, os sistemas de recuperação de informação podem ser enriquecidos por essas relações associativas.

3.4 Sistemas de recuperação de informação

Um sistema de recuperação de informação é um ambiente mediador entre um conjunto de documentos e seus potenciais requisitantes. Sua eficiência depende da qualidade da representação dos itens de informação (documentos) e das requisições de seus usuários (buscas) (NETO; FERNEDA, 2016), representado na Figura 8:

Figura 8. Modelo de recuperação da informação

Fonte: FERNEDA, 2012, p. 14, adaptada pelo autor.

Como ilustra a Figura 8, há um duplo processo de representação que precisa convergir para que o usuário localize os documentos de seu interesse. Em primeiro plano, em posse de um documento, o sistema de informação precisa representá-lo já pensando em expressões que o definam e sejam de reconhecimento intuitivo pelo usuário. Por outro lado, ao acessar o sistema, o usuário digita expressões de busca as quais representam a sua necessidade de pesquisa. Considerando essas duas representações (do sistema e do usuário), o sistema de recuperação será tão eficiente quanto “a expressão de busca seja representada de forma similar à forma como os documentos foram representados, para que seja possível uma comparação entre duas representações” (NETO; FERNEDA, 2016, p. 33).

Nesse sentido, um sistema de recuperação envolve um conjunto de elementos, operações, equipamentos e softwares, de forma que a estrutura organizada permita o armazenamento dos objetos de conteúdo para serem recuperados em resposta a pedidos dos usuários expressos por comandos que podem ser tratados pelo sistema. Lopes (2002, p. 60) destaca a complexidade desse tipo de sistema pela interdependência dos seus elementos constitutivos, e enfatiza que “a tecnologia eletrônica conduz os usuários ao acesso democrático à informação ampliando a busca de informação em bases de dados geograficamente distantes”,

no entanto, “o alcance da qualidade na informação recuperada requer o planejamento de estratégias de busca específicas para cada base de dados”.

Entre 1960 e 1970 surgiram os primeiros modelos de recuperação de informação, porém observamos que suas principais ideias ainda estão presentes na maioria dos sistemas de recuperação atuais, inclusive nos mecanismos de busca da Web. Peña, Baeza-Yates e Rodríguez Muñoz (2002) citam cinco modelos: busca livre (não exige indexação, acesso direto por combinações de cadeias de caracteres); modelo booleano (amplamente utilizado, conta com operadores para estabelecer inter-relações); modelo booleano estendido (utiliza operadores e atribui peso a eles); modelo vetorial (utiliza o sistema de atribuição de pesos para definir o ranqueamento dos materiais recuperados) e modelo de lógica difusa (baseado na inexistência de um limite definido para agrupamento de algo em um único grupo e com conjunto difuso).

Seja qual for o modelo utilizado pelo sistema de recuperação, três elementos se mantêm no centro de preocupação e podem determinar sua eficiência ou ineficiência, são eles: a linguagem de indexação, a linguagem de busca e o controle de vocabulário como estratégia viabilizadora de correspondência entre as duas linguagens.

Quando pensamos no verbo “recuperar” logo nos vem à mente o restabelecimento de vínculos e a ação de localizar. Associativamente, para (re)encontrar algo que foi afastado, são necessárias estratégias de busca, isto é, ações organizadas e estruturadas em prol da concretização do objetivo. Analogamente, estratégias de busca são ações organizadas para a recuperação de informações desejadas.

Desde a década de 1980, o aumento acelerado da produção de objetos de conteúdo impulsionou um movimento contundente no sentido de prover o usuário final com os instrumentos necessários para que o mesmo tivesse autonomia em suas próprias buscas (BACHA; ALMEIDA, 2013). Nesse contexto, a definição de ações sequenciais para alcançar resultados úteis se intensificou tanto no sentido de ampliar o acesso a informações e documentos quanto no sentido de restringir as pesquisas para uma compreensão mais específica e profunda do objeto. Portanto, as estratégias de busca passaram a ser foco de pesquisas de ampliação da recuperação da informação.

3.4.1 Estratégias de busca

Com o crescimento dos acervos digitais, os sistemas para armazenamento e recuperação de informação precisaram evoluir continuamente, fato que levou ao desenvolvimento de ferramentas capazes de auxiliar em pesquisas mais complexas, favorecendo a associação e integração de dados para ampliar a efetividade do acervo. A título de exemplo, temos a

possibilidade de aplicação de filtros em diversas bibliotecas digitais e repositórios, os quais permitem buscas por autores, título, palavras-chave ou expressões literais, ano de publicação, entre outras possibilidades (LOPES, 2002).

Segundo Lopes (2002, p. 61), estratégias de busca podem ser definidas como “uma técnica ou conjunto de regras para tornar possível o encontro entre uma pergunta formulada e a informação armazenada em uma base de dados”. Nessa compreensão, a autora acrescenta que muitas decisões são necessárias para facilitar esse encontro, tais como, a seleção do banco de dados mais adequado e a linguagem utilizada.

A respeito da importância das estratégias de busca, Lancaster (2004, p. 2-3) afirma que, para obter sucesso, os membros de uma comunidade, ao utilizar uma base de dados:

[...] devem converter uma necessidade de informação em algum tipo de “estratégia de busca”, a qual pode ser tão simples quanto a escolha de um único termo para consultar um índice impresso ou um catálogo em fichas, ou exigir a combinação de muitos termos numa estratégia mais elaborada e complexa, empregada para consultar uma base de dados mantida localmente ou conectada em linha a alguma rede de computadores.

Em complemento, Rowley (2004) e Lancaster (2004) definem estratégias de busca como o conjunto de decisões e ações tomadas durante uma busca com o objetivo de recuperar o máximo de registros relevantes e o mínimo de registros irrelevantes. Cumpre destacar que quanto maior a especificidade da busca, mais pontual tende a ser o retorno, evitando, assim, a recuperação de um número excessivo de registros.

A respeito da restrição ou ampliação da recuperação de objetos de conteúdo, Miranda e Alcará (2016, p. 3) enfatizam as estratégias de busca como sendo “um conjunto de expressões ligadas por operadores booleanos que permitem ampliar ou restringir os resultados, pode-se também utilizar sinais que auxiliem na operação ou melhore os resultados”.

No contexto dos RI, estratégias de busca bem definidas são fundamentais para o cumprimento do propósito de divulgação das produções acadêmicas e institucionais, pois é por meio delas que o usuário encontrará o objeto desejado. No entanto, para isso, é necessário ampliar as investigações sobre o público-alvo, isto é, os usuários finais dos RI.

Siebra, Oliveira e Marcelino (2014, p. 77), por meio de pesquisa em 16 repositórios institucionais brasileiros, destacam a existência de “problemas de acessibilidade, usabilidade e de interação com o usuário, especialmente no que concerne à funcionalidade de busca, uma das mais importantes nesse tipo de sistema”. Os autores apontam também dificuldades dos usuários devido a caminhos de navegação confusos até chegar a um documento. Para melhoria, pondera

a necessidade de considerar sistematicamente a organização, a navegação, a rotulagem e a busca de informações.

Também conduzindo investigação sobre a usabilidade dos repositórios institucionais brasileiros, Café e Kafure (2016, p. 39) focalizam o caso do Repositório Institucional da Universidade de Brasília e concluem que, embora seja um serviço de informação capaz de satisfazer as necessidades informacionais de usuários, “é necessária a inserção de funcionalidades para facilitar a interação com a interface do repositório”.

Em complementos, Marques e Vechiato (2017), no contexto do RI da Universidade Federal do Rio Grande do Norte (UFRN), observam insuficiência no dimensionamento dos usuários para a proposição de melhorias no que se refere à acessibilidade e usabilidade. Lima (2021) corrobora apresentando os resultados de pesquisas empíricas realizadas no Repositório da Universidade da Integração Internacional da Lusofonia Afro-Brasileira (Unilab) com alunos universitários, os quais demonstram que as instituições desconhecem a forma como os usuários utilizam os sistemas de informação e essa incompreensão se configura como empecilho para a melhoria da recuperação.

Diante do exposto, tendo verificado a ausência de pesquisas sobre as estratégias de busca dos usuários, no RI Unesp, confirmamos a necessidade de conhecer e listar as palavras que compõem o vocabulário desses usuários, para o que escolhemos a análise de logs de busca, nomeadamente aqueles que tenham algum indicativo de sucesso, a fim de que seja possível propor melhorias para a organização do RI.

3.4.2 Recuperação de informações no RI Unesp

Em repositórios baseados no DSpace, como é o caso do RI Unesp, que fazem uso da indexação automática, um mesmo metadado pode recuperar documentos completamente diferentes. Reis (2008) aponta o metadado palavra-chave como exemplo, destacando as funções de organizar, classificar, hierarquizar os documentos no repositório e facilitar a recuperação por assunto. Entretanto, “problemas simples como o de grafia, que provocam a recuperação de documentos diferentes, revelam a necessidade de padronização e de um estudo mais profundo sobre vocabulários controlados” (REIS, 2008, p. 12).

O uso de vocabulário controlado está diretamente relacionado à padronização terminológica para que se evite variações que inviabilizem ou prejudiquem a recuperação de informações. Como orientam tanto a indexação quanto a recuperação, os vocabulários controlados melhoram a qualidade e disponibilidade como suporte para buscas em várias bases de dados, ampliando assim as convergências e, consequentemente, o sucesso da recuperação.

Nesse sentido, os vocabulários controlados buscam alcançar a melhor representação dos conceitos pela percepção acurada das relações semânticas e hierárquicas entre os termos. Como afirma Reis (2008, p. 12), os vocabulários controlados “são instrumentos de controle terminológico utilizados em sistemas de informação para traduzir a linguagem dos documentos, dos indexadores e dos pesquisadores numa linguagem controlada, usada na indexação e recuperação de informações”.

Para compreendermos melhor a importância do controle de vocabulário no RI Unesp, observamos que os itens depositados necessitam de alguns metadados obrigatórios, tais como título, autoria, resumo e palavras-chave. Assim, a recuperação dos objetos digitais por assunto considera a navegação por palavras-chave, seja por meio da ferramenta de busca presente no RI, seja por uma lista automática gerada pelo próprio sistema contendo todas as palavras-chave indexadas no repositório. Todavia, a não correspondência literal entre as palavras tende a gerar perdas na recuperação que vão desde a exclusão de materiais representados por sinônimos até daqueles que foram representados por palavras com diferenciação de letras maiúsculas e minúsculas ou desvios gramaticais e mesmo de digitação. Portanto, a falta de controle de vocabulário da lista de palavras-chave atribuídas pelos autores faz com que materiais de interesse do usuário não sejam recuperados, como exibido na Figura 9:

Figura 9. Exemplo de potenciais perdas na recuperação por divergências de grafia

The screenshot shows the 'Repositório Institucional UNESP' search page. The left sidebar contains navigation links like 'Em todo o Repositório', 'Tipo de Produção', 'Data do documento', 'Autor', 'Título', 'Palavra-chave', 'MINHA CONTA', 'Entrar', 'SOBRE O REPOSITÓRIO', 'Estatísticas de acesso', and 'Acompanhamento de submissões'. The main area is titled 'Navegar por palavra-chave' and shows a search bar with 'Pesquisar' and a magnifying glass icon. Below the search bar, there's a list of search results for 'Cardiopatia' and 'Útero'. The results are organized into two columns: 'Palavras-chave' and 'Exemplo 1' / 'Exemplo 2'. The 'Palavras-chave' column lists terms like 'Cardiopatia [2]', 'cardiopatia [1]', 'Cardiopatia congenita [1]', 'cardiopatia congênita [1]', 'Cardiopatias [1]', and 'cardiopatias [1]'. The 'Exemplo 1' column shows 'Quantitativos e arquivos diferentes por variação de letras maiúsculas e minúsculas e uso do singular ou plural'. The 'Exemplo 2' column shows 'Útero [4]', 'Útero [19]', 'útero [3]', and 'Úteros [1]'. Arrows point from the 'Cardiopatia' and 'Cardiopatias' terms to the 'Exemplo 1' column, and from the 'Útero' terms to the 'Exemplo 2' column. The 'Exemplo 2' column also shows 'Quantitativos e arquivos diferentes por variação de acentuação, de letras maiúsculas e minúsculas e uso do singular ou plural'.

Palavras-chave	Exemplo 1	Palavras-chave	Exemplo 2
Cardiopatia [2]	Quantitativos e arquivos diferentes por variação de letras maiúsculas e minúsculas e uso do singular ou plural	Útero [4]	Quantitativos e arquivos diferentes por variação de acentuação, de letras maiúsculas e minúsculas e uso do singular ou plural.
cardiopatia [1]		Útero [19]	
Cardiopatia congenita [1]		útero [3]	
cardiopatia congênita [1]		Úteros [1]	
Cardiopatias [1]	Quantitativos e arquivos diferentes por variação de acentuação, de letras maiúsculas e minúsculas e uso do singular ou plural.		
cardiopatias [1]			

Fonte: elaborada pelo autor a partir do Repositório Institucional Unesp (24 mar. 2022).

Nos exemplos 1 e 2 (Figura 9), a montagem com duas palavras de pesquisa permite identificar que palavras idênticas, mas com letras em caixa alta ou baixa indexam materiais diferentes e podem ser perdidas durante a recuperação feita pela navegação por palavras-chave. Também não há tratamento de plural/singular ou de acentuação, fatores que podem gerar ruídos na busca de materiais desejados e impedir sua recuperação deles pelos usuários.

Além de medidas para eliminar as perdas de recuperação pela não correspondência literal das palavras, para aprimorar as estratégias de recuperação dentro do RI Unesp, mostra-se fundamental conhecer melhor a linguagem empregada pelos usuários, de forma a reconhecer a relevância desse vocabulário. Ao buscar inter-relações as quais permitam agregações semânticas e sintáticas que eliminem redundâncias e ambiguidades, o sistema de recuperação amplia sua efetividade, já que o processamento analítico é organizado para oferecer a recuperação com o máximo de informações úteis e o mínimo de documentos inúteis, de modo que o usuário encontre o que precisa, mas não perca tempo com resultados dissociados (LANCASTER, 2004).

No caso específico do RI Unesp, o modelo metodológico de avaliação do controle de vocabulário está sendo desenvolvido com base em Plano de Ações contendo seis projetos de pesquisa conduzidos pelo Grupo de Estudos Representação Temática e agrupados no título “O controle de vocabulário na representação e recuperação da informação em repositórios universitários” (FUJITA, 2022).

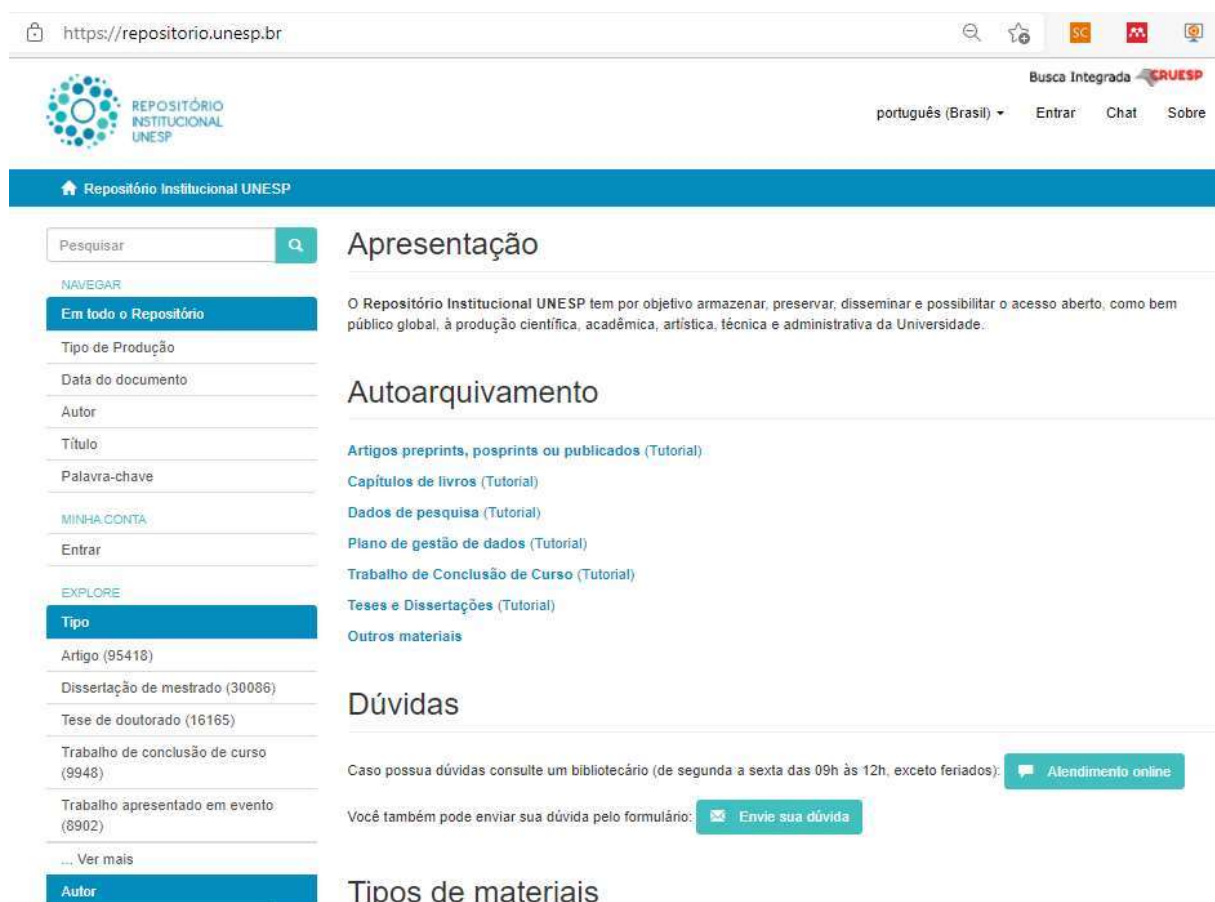
A sistematização propõe metodologias de avaliação de controle de vocabulário a serem aplicadas com objetivos definidos em processos de organização e representação da informação com precisão e especificidade. Para tanto, os pesquisadores do Grupo de Estudos Representação Temática buscam metodologias capazes de identificar os processos e sistemas de organização e representação da informação compatíveis. Entre os estudos integrantes da pesquisa, está o estudo teórico e metodológico sobre análise de logs em repositórios institucionais e a análise experimental dos logs de busca por assuntos dos usuários do RI Unesp como meio de viabilizar a atualização de vocabulários controlados, objetivos motivadores desta Tese.

Assim, para conhecer melhor a linguagem natural dos usuários, a investigação dos logs de busca empregados no RI Unesp se configurou como estratégia extensiva, discreta, com baixo custo relativo e passível de análise computacional capaz de processar o grande volume de dados coletados.

3.4.3 Estratégias de busca no RI Unesp

O RI Unesp é aberto à comunidade e não requer o cadastro de login e senha, o que facilita o acesso, mas também dificulta a identificação dos usuários para posterior categorização. No que concerne às estratégias de busca disponibilizadas pela plataforma, ao acessar a página inicial do RI Unesp o usuário já encontra diversas possibilidades como indica a Figura 10:

Figura 10. Página inicial do RI Unesp



Fonte: Repositório Institucional Unesp (24 mar. 2022).

Na Figura 10, canto superior direito, há a possibilidade de pesquisa integrada no Repositório da Produção Científica do Cruesp (Conselho de Reitores das Universidades Estaduais Paulistas). Um pouco abaixo, o usuário encontra a possibilidade de alterar o idioma para inglês ou espanhol, acessar por meio de login, esclarecer dúvidas por meio de chat e conhecer mais sobre o RI. Em coluna à esquerda, há barra de busca para digitação livre do usuário e abaixo são listados filtros para acesso e navegação, de modo que o usuário possa restringir os objetos de conteúdo que sejam de seu interesse de acordo com o tipo de documento,

a autoria, data de publicação, idioma, agência de fomento, palavra-chave, pós-graduação e *campus*.

Ainda na página inicial há informações e links sobre autoarquivamento, dúvidas com atendimento síncrono e assíncrono, tipos de materiais, tipo de produção e submissões recentes, além de acesso às estatísticas do próprio RI e acompanhamento das submissões.

Ao digitar um termo na barra de busca, o usuário tem a possibilidade de aplicar novos filtros, como exposto na Figura 11:

Figura 11. Página de busca avançada no RI Unesp

Fonte: Repositório Institucional Unesp (24 mar. 2022).

No exemplo, ao digitar o termo de busca “indexação”, há várias opções, inicialmente, onde se lê “Todo o repositório”, o usuário pode restringir por: “Produção acadêmica e científica”; “Produção administrativa”; “Produção artística”; “Produção comemorativa - Unesp 40 anos”; “Produção cultural” ou “Produção Técnica”. Sequencialmente, exibindo os filtros para refinamento, o usuário pode refinar a busca localizando documentos com títulos: iguais a expressão de busca, que a contenham ou não e as várias opções que podem ser selecionadas, tais como: título, autor, idioma, data de publicação e palavra-chave.

Observamos que após o preenchimento dos dados, aparece a quantidade de resultados, como também a possibilidade de ordenação dos documentos por: relevância, título em ordem alfabética, data da publicação crescente ou decrescente; e, por fim, mostrou-se a quantidade de documentos listados por página.

Ao clicar em um dos resultados, o usuário tem acesso a mais informações sobre o documento e pode acessar seus metadados, tais como: palavras-chave, citação, resumo, descrição, autor, idioma, *Uniform Resource Identifier* (URI) - que é um link direto do documento na página da internet.

Expostas as possibilidades fornecidas pela página do RI Unesp para aprimoramento das buscas realizadas por seus usuários, para avaliar sua efetividade, é preciso conhecer melhor o vocabulário utilizado por esses usuários e, embora haja na literatura estudos sobre as estratégias de buscas em Repositórios Institucionais, não há pesquisa específica no RI Unesp. Assim, análise de logs se configurou como um método discreto e viável para conhecer a linguagem empregada pelos usuários, bem como indicadores de sucesso na recuperação dos objetos de conteúdo desejados.

3.5 Logs de busca de usuários e Processamento de Linguagem Natural

Para a análise dos logs de busca dos usuários dentro de um sistema de recuperação de informações como o RI Unesp, o desenvolvimento de algoritmos capazes de realizar o processamento da linguagem natural dos usuários é de suma importância, haja vista o volume e variedade de registros coletados sem qualquer tratamento ou padronização. É necessário definirmos, portanto, o conceito de logs de busca e as técnicas já existentes para Processamento de Linguagem Natural (PNL).

3.5.1 Logs de buscas de usuários no RI Unesp

O valor representativo dos logs como “pegadas de informação digital” é enfatizado por Nicholas, Huntington e Waltkionson (2005, p. 250), uma vez que registram o caminho percorrido pelo usuário dentro do sistema utilizado. Em concordância, Jansen (2008) pontua que, enquanto navega, o usuário deixa rastros de sua interação com o sistema e esses vestígios podem ser utilizados pelos desenvolvedores para melhorar a experiência do usuário através da previsão de padrões de tráfego, intenções e necessidades.

Peter (1993) apresenta a TLA como o estudo de interações gravadas eletronicamente entre sistemas de recuperação de informações *on-line* e pessoas que procuram informações encontradas nestes sistemas. Nesse entendimento, para mapeamento do percurso dos usuários no RI Unesp, adotamos os fundamentos da TLA como técnica de coleta de dados sob a ótica da interação produtiva, isto é, aquela que satisfaz a necessidade do usuário, uma vez que os logs

se constituem como material primário que permite a consulta direta das expressões utilizadas pelos usuários para a recuperação da informação.

No que concerne aos logs de busca dos usuários no RI Unesp, há 34 campos em cada registro de acesso, no entanto, de acordo com o objetivo da pesquisa, podem ser feitos recortes. No caso específico, dado o objetivo de realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais, foi realizado um recorte no material coletado no RI Unesp, mantendo-se apenas os campos *IP*, *time*, *query* e *id* por fornecerem dados suficientes para conhecer a linguagem de pesquisa dos usuários. Apenas a título de exemplo, no Quadro 4, expomos parte do material coletado:

Quadro 4. Exemplo de armazenagem dos logs de buscas dos usuários no RI Unesp

IP (anonimizados)	Time Data e hora	Query Termo de busca do usuário	ID Identificador
000.00.000.001	2019-05-16T21:01:08.745Z	deficiencia visual	
000.00.000.002	2019-05-16T21:03:09.663Z	energia renovável	156039
000.00.000.0083	2019-05-16T21:03:59.461Z	author_keyword:Ortiz%2C\\\\Edmundo\\\\Cabral\\,subject_keyword:C\\\\-scan\\\\ test	
000.00.000.00207	2019-05-16T21:04:08.89Z	author_keyword:Ortiz%2C\\\\Edmundo\\\\Cabral\\,subject_keyword:C\\\\-scan\\\\ test	
000.00.000.00243	2019-05-16T21:04:16.407Z	author_keyword:Ortiz%2C\\\\Edmundo\\\\Cabral\\,subject_keyword:C\\\\-scan\\\\ test	
000.00.000.0066	2019-05-16T21:25:58.722Z	na labuta:	
000.00.000.0066	2019-05-16T21:26:06.505Z	na labuta:\\,type_keyword:Tese\\\\de\\\\ doutorado	
000.00.000.0066	2019-05-16T21:26:10.776Z	na labuta:	184547
000.00.000.0010	2019-05-17T03:10:34.593Z	celso júnior	
000.00.000.0010	2019-05-17T03:10:45.272Z	celso júnior	167487

Fonte: elaborado pelo autor.

Cada uma das linhas do Quadro 4 representa uma das 480.742 requisições dos usuários ao RI, no intervalo de tempo selecionado. Para a fundamentação metodológica, convém sublinharmos que as colunas dispõem os campos coletados com os respectivos valores,

destacando-se no campo *query* a possibilidade de verificação da estratégia/linguagem de busca dos usuários, utilizando (ou não) diversos filtros para auxílio na recuperação do documento desejado, o campo *id* contém registro apenas no caso de abertura de documento, ação com potencial indicativo de sucesso da busca.

Em posse dos logs, foi possível conhecer de maneira extensiva as palavras de busca dos usuários e considerar a linguagem utilizada por eles durante o processo de recuperação da informação, a fim de comparar com o vocabulário controlado que compõe o Tesouro Unesp e também com as palavras de linguagem natural empregados pelos autores. Pressupomos que essa tríplice comparação tem o potencial de favorecer a recuperação de informações, pois indicará convergências e divergências entre as linguagens por meio da abordagem quantitativa e qualitativa dos dados, como detalhado mais adiante na descrição dos procedimentos metodológicos. Contudo, para explorar o potencial do material coletado, mostrou-se necessário o conhecimento de técnicas de Processamento de Linguagem Natural.

3.5.2 *Processamento de Linguagem Natural*

A descoberta de conhecimento por meio de técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras pode contribuir muito para a recuperação eficiente de informações. Esse processo envolve o processamento de textos pela aplicação de algoritmos computacionais capazes de identificar padrões, informações úteis e mesmo implícitas (BARBOSA *et al.*, 2017). Nessa compreensão, as técnicas de PLN têm grande potencial para auxiliar na tomada de decisões tanto em âmbito empresarial sobre a aceitação de produtos quanto no cenário acadêmico para pesquisas relacionadas à eficiência de determinado serviço. No contexto do uso de repositórios, pode permitir mensurar a convergência entre vocabulários controlados e a linguagem natural dos usuários, por exemplo.

Conforme Bird, Klein e Loper (2009), para o processamento de linguagem natural há a necessidade de conjuntos de dados linguísticos amplos. Tais conjuntos de dados são essenciais por projetarem de forma balanceada materiais que viabilizam análises estatísticas e testes de hipóteses. Barbosa *et al.* (2017, p. 2) apontam que para problemas práticos “é adequado utilizar biblioteca que oferece recursos avançados de PLN”, nesse sentido, “uma das vantagens do Python é a existência de diversas dessas bibliotecas, como a NLTK”.

A linguagem Python, a qual é amplamente utilizada em computação e estatística para a análise de grandes volumes de dados (BARBOSA; PERICO, 2019), é baseada no conceito de software livre, além de ter pacotes estatísticos que podem ser utilizados para tratar volumes de

dados, conta também com curva rápida de aprendizagem, extensa documentação e bibliotecas da linguagem, que permitem acesso a pacotes como o *pandas*, o qual fornece ferramentas para manipulação e análise de dados de alta performance, com destaque por oferecer estruturas e operações para manipular tabelas numéricas e séries temporais (PANDAS, 2021).

Ademais, como apontam Barbosa *et al.* (2017), a programação em Python viabiliza a utilização da NLTK, um conjunto de bibliotecas e programas para processamento simbólico e estatístico da linguagem natural que é escrito justamente na linguagem de programação Python. Barbosa *et al.* (2017, p. 4-5) indicam:

Criada originalmente em 2001 como parte de um curso de linguística computacional do Departamento de Ciência da Computação e Informação da Universidade da Pensilvânia, a NLTK é uma plataforma usada para construir programas Python que trabalham com dados de linguagem humana para aplicação de PLN. O NLTK define uma infraestrutura que pode ser usada para construir programas de PLN em Python; fornece classes básicas para representar dados relevantes para o processamento da linguagem natural; interfaces padrão para executar tarefas como tokenização, classificação morfológica, análise sintática e classificação de texto; além de implementações padrões para cada tarefa que podem ser combinadas para resolver problemas complexos.

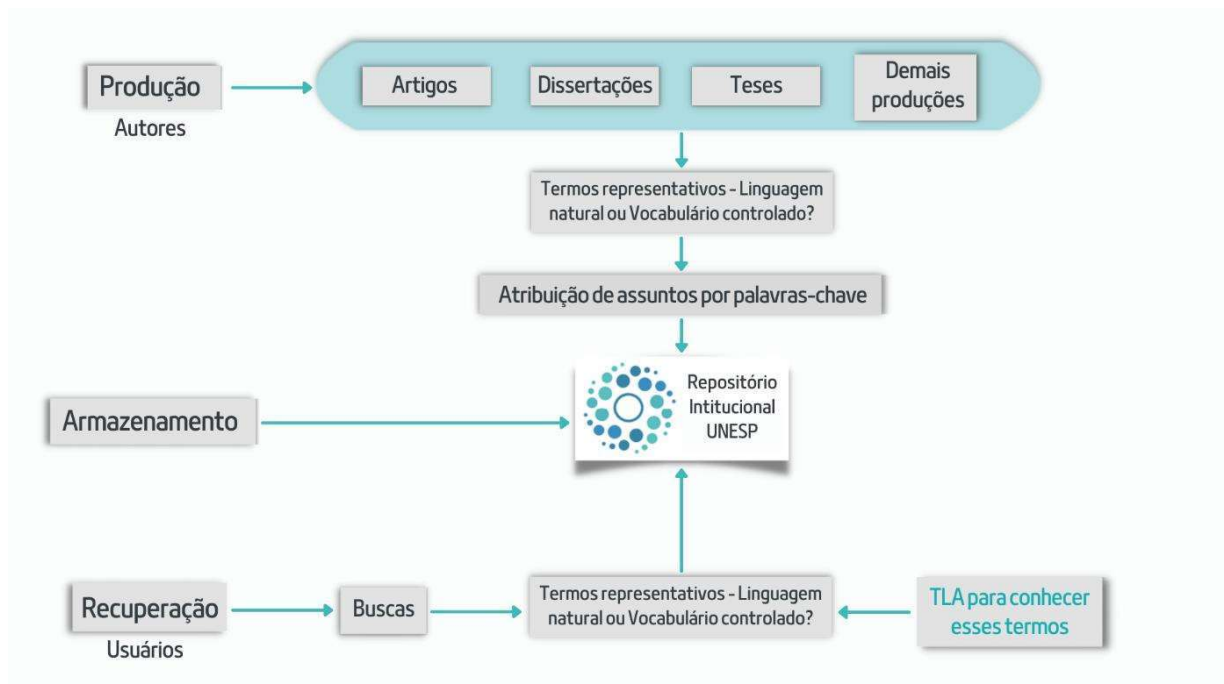
O acesso à NLTK é fundamental para o desenvolvimento desta pesquisa, pois, por ser uma biblioteca gratuita e de código aberto, é continuamente melhorada o que viabiliza a construção de aplicativos que decodificam grandes volumes de texto, além disso pode ser usada na elaboração de sistemas de extração e seleção de informações ou compreensão de linguagem natural por presença de padrões. A biblioteca viabiliza também o pré-processamento de textos para aprendizagem profunda, característica essencial para o processamento da linguagem utilizada pelos usuários para realização de buscas. Assim, a retomada e sistematização dos fundamentos teóricos expostos nesta seção é apresentada a seguir.

3.6 Síntese dos resultados dos fundamentos teóricos

A constituição dos repositórios institucionais, assim como o entendimento de sua estruturação e funcionamento é essencial para a proposição de melhorias tanto no processo de organização quanto de recuperação da informação. Dada a abrangência do tema, é necessário especificar a ação interventiva para viabilizá-la, entretanto, não se pode perder a visão do todo no que concerne à recuperação ser precedida pela organização dos documentos.

Para exibir a integração entre os conceitos tratados nesta seção 3 e favorecer o entendimento do fluxo documental (entrada, tratamento e saída) do RI Unesp, expomos a Figura 12:

Figura 12. Fluxo documental no RI Unesp: entrada, armazenamento e recuperação

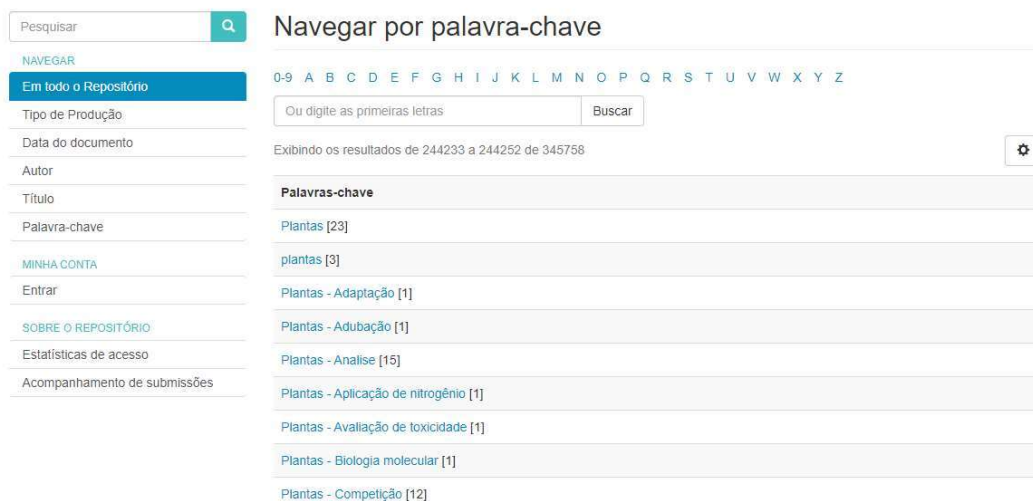


Fonte: elaborada pelo autor.

A Figura 12 exibe o processo percorrido pelos materiais no RI Unesp desde a entrada até o retorno aos usuários como resultado de busca. Para entrada, os documentos são inseridos a partir de metadados representativos que são adicionados ao sistema: colaborador; criador; data; identificador; linguagem; publicador; assunto; título; tipo e palavras-chave. Todos os metadados são preenchidos por autoarquivamento com finalidade de padronizar o registro de dados representativos do documento, para sua posterior recuperação pelos usuários do sistema. O tratamento envolve a atribuição de assuntos por palavras-chave que pode ser favorecida (do ponto de vista representativo) pelo uso de vocabulário controlado. Por fim, o processo de saída (recuperação) é condicionado pelas estratégias de buscas empregadas, as quais determinam a efetividade da recuperação dos documentos.

Exemplo de como documentos podem não ser recuperados devido a pequenas variações de registro pode ser visualizado na Figura 13, na qual observamos que o diferencial da inicial maiúscula na palavra “plantas” pode resultar em 23 ou 3 documentos identificados:

Figura 13. Duplicidade de palavras-chave dos autores



Fonte: Repositório Institucional Unesp (24 mar. 2022).

Como afirma Lancaster (2004), o acesso a uma base de dados resulta na divisão entre itens recuperados e itens não recuperados; nesta divisão, muitos fatores estão envolvidos, tais como, a cobertura da base de dados, sua política e prática de indexação, além da política e prática de resumos, a qualidade do vocabulário e das estratégias de busca previstas e viabilizadas.

Portanto, a Figura 12 exibe a visão geral de como a TLA pode intervir na atualização de vocabulários controlados eficientes e atualizados, posto que viabiliza o conhecimento da linguagem natural dos usuários ao capturar as palavras de busca utilizadas no momento das pesquisas, colaborando também para a compreensão das estratégias de buscas definidas para a recuperação dos materiais desejados.

Como posto pela ISO 25964 - 1 (2011, p. 6), a recuperação de informações se tornará mais eficiente “se o indexador e o pesquisador forem orientados a escolherem o mesmo termo para o mesmo conceito”, tal convergência é favorecida pelo uso de vocabulário controlado, uma vez que ambiguidades e redundâncias são eliminadas. Para buscar o melhor termo representativo para um conceito, torna-se latente a necessidade de os profissionais reconhecerem o valor da linguagem dos usuários, dado a língua ser viva e, portanto, em constante transformação.

A lista de palavras-chave quando extraída por ferramenta computacional traz grande variedade de registros e, por ser composta por palavras de linguagem natural, sem qualquer padronização, tem alguns problemas: letras e palavras sem sentido; variações quanto ao uso de

letras maiúsculas e minúsculas; singular e plural; acentuações; palavras só recuperadas se digitadas com caracteres especiais (aspas ou *hashtags*).

Nesse íterim, os logs de busca dos usuários são registros ricos, justamente por conterem a linguagem natural dos usuários e, por serem registrados automaticamente pelo próprio sistema assim que uma busca é realizada. Os logs são a forma mais discreta, rápida e abrangente possível de mapeamento do vocabulário dos usuários de um RI.

Convém destacarmos que a lista de palavras oriundas dos logs de busca dos usuários combinada com a lista de palavras-chave atribuídas pelos autores permite uma análise mais ampla da linguagem natural empregada pelos agentes que interagem no repositório.

Sequencialmente, a seção 4 apresenta a metodologia escolhida para o desenvolvimento desta Tese, a fim de esclarecer o caminho procedimental quanto à definição de abordagens adequadas a cada etapa da pesquisa, bem como os critérios do método para a coleta, preparação e análise de dados.

4 METODOLOGIA

Com o propósito de viabilizar o aprimoramento da organização e representação do conhecimento nos repositórios institucionais, a partir do estudo dos comportamentos dos usuários, a realização da presente Tese visa acompanhar as estratégias de busca dos usuários, para que seja possível aprimorar suas experiências de busca, de modo a garantir o acesso e a usabilidade das informações e conhecimentos produzidos e armazenados nos repositórios de instituições de ensino e pesquisa. Assim, nesta seção, é feita a caracterização da pesquisa.

4.1 Caracterização da pesquisa

Esta pesquisa se caracteriza como qualitativa e quantitativa e seu desenvolvimento se realizou em duas etapas: pesquisa exploratória e pesquisa descritiva. A pesquisa exploratória utilizou levantamento bibliográfico e análise documental e a pesquisa descritiva focalizou o trabalho com os dados coletados para realizar análises quantitativas e qualitativas as quais subsidiaram o desenvolvimento de uma metodologia empírica para a análise de logs. Tal processo descritivo e analítico resultou na ferramenta computacional RILogUser, a qual é apresentada na seção resultados.

Assim, nas duas subseções seguintes descrevemos a pesquisa exploratória com pesquisa bibliográfica e documental e a pesquisa descritiva.

4.1.1 *Pesquisa exploratória com pesquisa bibliográfica e documental*

Para atender ao primeiro objetivo específico da pesquisa – estudar os logs de buscas de usuários do Repositório Institucional Unesp para proposta de atualização do vocabulário controlado Tesouro Unesp – antes de partir para a análise do material coletado, foi preciso realizar a revisão de literatura, conforme resultados apresentados na seção 2. Destaca-se a importância da revisão de literatura como “[...] um meio de identificar, avaliar e interpretar pesquisas disponíveis e relevantes para uma determinada questão de pesquisa, uma temática ou fenômeno de interesse” (KITCHENHAM, 2004, p. 1, tradução nossa).

Assim, a fim de verificar a literatura publicada no período de 1967 a 2022, foram realizadas consultas às bases de dados *Scopus* e *Web of Science* ao longo de 2019, com revisões mensais periódicas até junho de 2022. Prosseguindo a revisão nas duas bases de dados, identificamos 227 publicações recuperadas na *Web of Science* e 113 na *Scopus*. Constatamos, pela leitura dos títulos, resumos e palavras-chave, que os documentos encontrados na *Scopus*

atendiam de modo mais específico o objetivo da pesquisa. Decidimos, portanto, continuar o procedimento de pré-seleção dos artigos com os 113 artigos da *Scopus*, conforme Quadro 5:

Quadro 5. Estratégias de busca para revisão de literatura

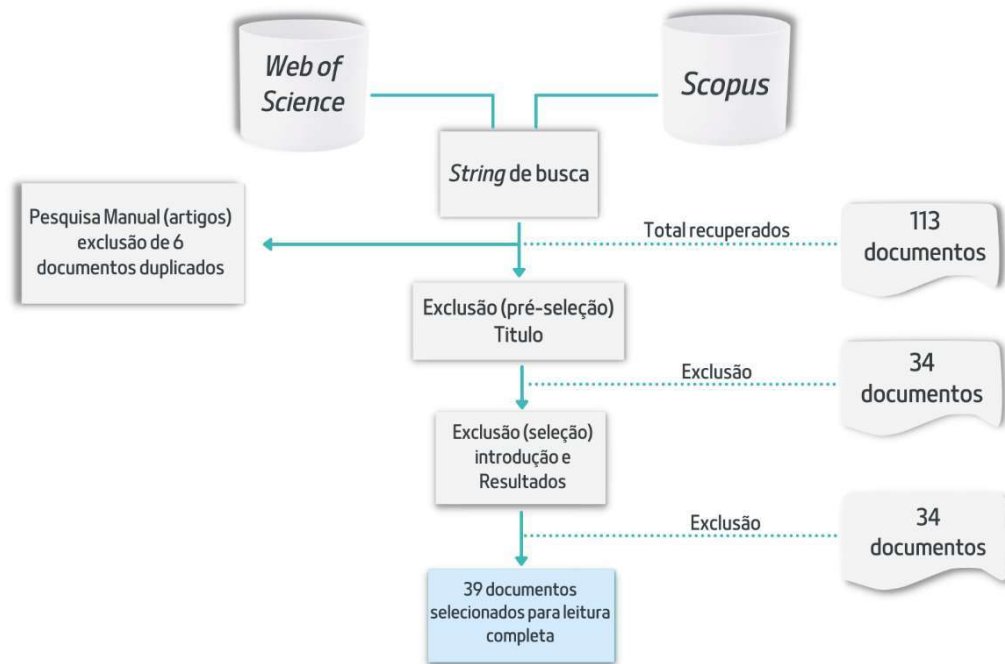
Estratégias de busca	Quantitativos recuperados
TITLE ("Transaction log analysis")	44
((TITLE ("log analysis") AND ABS ("log analysis") AND KEY ("log analysis"))) AND ((KEY (search) OR KEY ("information retrieval")))	21
(TITLE (logs) OR TITLE (queries) AND TITLE ("controll* vocabular*")) OR TITLE ("on-line information retrieval system") OR SRCTITLE ("Handbook of Research on Web Log") OR TITLE ("computer-monitored data") or TITLE ("digital libraries through log file analysis") OR TITLE ("research-oriented digital library")	48
Total:	113

Fonte: elaborado pelo autor.

Com a definição das estratégias de busca nos bancos de dados selecionados, inicialmente, foram recuperados 113 materiais, dos quais 6 estavam duplicados, resultando em 107 documentos únicos, após a leitura do título, 34 documentos foram excluídos por não indicarem claramente estarem relacionados ao tema. Na sequência, foi realizada a leitura da introdução e resultados com 34 documentos eliminados por não terem foco na aplicação de análise de logs associada ao estudo de vocabulário dos usuários de sistemas de informação.

Por fim, 39 foram selecionados para leitura completa e foram incluídos nesta revisão. Os critérios de inclusão e exclusão permitiram a seleção de trabalhos importantes para a compreensão do potencial dos logs para a melhoria de sistemas de informação a partir do conhecimento do vocabulário empregado por seus usuários. A Figura 14 representa as etapas de seleção de publicações acerca do tema:

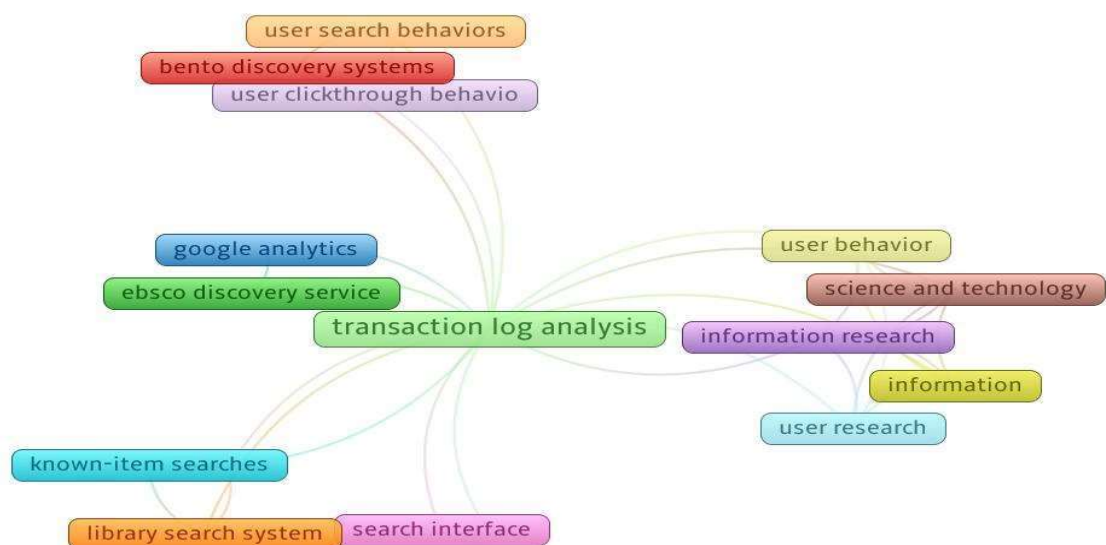
Figura 14. Seleção de estudos para revisão de literatura



Fonte: elaborada pelo autor.

A Figura 15 apresenta uma análise, na forma de mapa de rede, de coocorrência de palavras-chave, tal análise foi gerada a partir do tratamento dos dados obtidos nas buscas:

Figura 15. Análise de coocorrência de palavras-chave dos 39 documentos selecionados



Fonte: elaborada pelo autor.

As relações entre as palavras-chave representativas das publicações selecionadas reforçam a compreensão ampla das aplicações da TLA (*Transaction log Analysis*) para investigação dos comportamentos de busca dos usuários. Confirma-se, portanto, a TLA como uma grande oportunidade para os profissionais da Ciência da Computação e Ciência da Informação se envolverem com pesquisas relacionadas ao vocabulário empregado por usuários em repositórios institucionais, de modo que o conhecimento das palavras de busca utilizadas viabilize o aprimoramento da recuperação de informações nestes sistemas, a fim de acelerar os avanços nas áreas relacionadas.

A pesquisa documental cuja essência está, conforme Kripka, Scheller e Bonotto (2015, p. 57), em analisar “documentos que não sofreram tratamento analítico, ou seja, que não foram analisados ou sistematizados”, foi uma escolha metodológica oportuna, dada a importância de investigarmos diretamente a linguagem dos usuários, isto, é sem qualquer filtragem ou intervenção de outros agentes.

Isso posto, a pesquisa documental foi realizada mediante contato inicial da orientadora desta Tese, Prof.^a Dr.^a Mariângela Spotti Lopes Fujita, com a Coordenadoria Geral de Bibliotecas da Unesp (CGB), representada pela Prof.^a Dr.^a Flávia Maria Bastos, em 08 de maio de 2020. Posteriormente, foi feita solicitação formal de acesso aos logs pelo grupo de pesquisa “Representação Temática da Informação”, em 30 de junho de 2020. A solicitação foi analisada e aprovada pela Coordenadoria Geral de Bibliotecas da Unesp.

Dessa forma, com a solicitação aprovada, tivemos acesso aos logs dos usuários, os quais foram o material primordial para o *corpus* da pesquisa descritiva. O material foi obtido de forma mediada, sendo a Coordenadoria Geral de Biblioteca da Unesp responsável por fornecer o arquivo CSV o qual foi objeto de análise nesta Tese.

A pesquisa é realizada a partir da coleta, preparação e análise de três listagens de palavras: I) Logs de buscas dos usuários; II) Palavras-chave dos autores; III) Vocabulário controlado Tesouro Unesp.

Quanto aos logs de busca dos usuários, eles são referentes ao período de 01 de julho a 31 de dezembro de 2019, sem delimitação de sessões, uma vez que o objetivo não era identificar o número de usuários, mas sim o vocabulário empregado. Devido ao grande volume de dados, com 34 campos e 480.742 linhas de registros, inicialmente, buscamos selecionar os campos com potencial para investigarmos a linguagem empregada pelos usuários ao realizarem buscas no RI Unesp. Nesse propósito, foram selecionados quatro campos: IP para identificação da máquina utilizada na pesquisa; *time* para apuração da data e hora; *query* contendo palavras de

busca do usuário e *id* que indica as pesquisas que geraram abertura de arquivo. Com os campos reduzidos, o desafio de selecionar, tratar e interpretar os dados ganhou foco e especificidade associada ao objetivo de compreender a interação do usuário durante suas interações com o RI Unesp.

Quanto às palavras-chave dos autores, a coleta aconteceu no RI Unesp, no dia 24 de março de 2022, incluindo todos os registros feitos da implementação do RI até esta data. A coleta foi feita por meio de *web crawler*, programa com *script* simples, criado para coletar dados sequencialmente em diversas páginas Web, criando uma coleção de páginas (CHO; GARCIA-MOLINA, 1999), isto é, uma forma de coleta automatizada de todas as palavras-chave dos autores, capaz de percorrer as 32 páginas *web* do RI, cada uma com aproximadamente 10.000 palavras-chave, totalizando 325.524 palavras-chave de autores, de todas as áreas do conhecimento.

Por fim, a lista dos termos pertencentes ao Tesauro Unesp foi coletada de forma automática, por meio de um *script* em Python para acessar a API (*Application Programming Interface*). Uma API é uma aplicação programada para fornecer uma interface de programação geral que pode ser reutilizada de várias aplicações (GONZÁLEZ; NIETO, 2012), disponível pelo Tesauro (disponível em: <https://github.com/ikatahira/Doutorado>, módulo “thesauro_handler.py”). A coleta dos dados foi realizada em 24 de março de 2022. Nesta primeira listagem, constaram 40.351 termos autorizados e 49.237 não autorizados, totalizando 89.588 termos elencados no Tesauro.

Em posse das listagens de vocabulário oriundas da coleta computacional, o passo metodológico seguinte se concentrou na preparação do grande volume de dados para posterior análise. Sequencialmente, a próxima seção descreve o método proposto de atualização de vocabulário controlado por meio da TLA e as respectivas etapas definidas para execução.

4.1.2 Pesquisa descritiva

A pesquisa descritiva é aplicada à análise dos logs de busca dos usuários do RI Unesp a fim de viabilizar a compreensão de mecanismos de recuperação de informações pelos usuários, de modo a aumentar a compatibilização entre o registro e a recuperação da informação.

A utilização da pesquisa descritiva se justifica para cumprirmos o segundo e terceiro objetivos da Tese, que são: analisar, a partir de logs de pesquisa feitas pelos usuários no RI Unesp, a viabilidade metodológica de aplicação de recursos computacionais para realização de rotinas automatizadas para depuração / limpeza de logs de pesquisa para composição de lista

de palavras candidatas à inclusão no Tesouro Unesp, subsidiando metodologia específica de análise de logs; e, criar e investigar a aplicação experimental da análise de logs implementada por meio de ferramenta computacional desenvolvida durante a produção desta Tese, para a atualização de vocabulário controlado do Tesouro Unesp, com vistas à ampliação de seu uso no RI Unesp.

Conforme Marconi e Lakatos (2017, p. 304), “a finalidade da pesquisa científica não é apenas fazer um relatório ou descrição dos dados pesquisados empiricamente, mas também relatar o desenvolvimento interpretativo dos dados obtidos”. Com esse entendimento, a pesquisa descritiva foi escolhida, dado o objetivo de compreender de forma detalhada os significados e as características do objeto de pesquisa, a partir dos dados quantitativos oriundos dos logs.

Gil (1999, p. 35) afirma que “os procedimentos estatísticos fornecem considerável reforço às conclusões obtidas”, tornando-as bastante aceitas entre os pesquisadores. A análise quantitativa na pesquisa descritiva é, conforme Marconi e Lakatos (2017, p. 325), “uma garantia de confiabilidade, particularmente pelo rigor matemático estatístico e controle das variáveis”.

O enfoque quantitativo, para Sampieri, Collado e Lucio (2013, p. 30), vale-se da coleta de dados para testar hipóteses, com base na medição numérica e na análise estatística, para estabelecer padrões e comprovar teorias [...]. A finalidade das investigações quantitativas é “explicar e prever os fenômenos pesquisados, buscando regularidades e relações causais entre elementos” (MARCONI; LAKATOS, 2017, p. 327-328).

Adicionalmente, esta Tese traz também a avaliação qualitativa da representação da informação pelo estudo das variações de registros viabilizados pela extração de logs. Os logs resultantes dos acessos e processos de buscas empreendidas pelos usuários permitem a investigação de suas estratégias de busca no Repositório.

Destacamos, neste momento, o aspecto inovador da presente pesquisa na aplicabilidade da metodologia da TLA na avaliação da recuperação de informações pelo usuário no RI Unesp, a partir de indicadores contidos nos logs, sendo estes os responsáveis pelo nosso conhecimento das palavras de busca empregadas, bem como dos registros de abertura de arquivos recuperados.

Cabe esclarecer que, por se tratar de pesquisa de natureza aplicada, cujo interesse prático específico e imediato é analisar logs de busca de usuários como fonte informacional para a atualização de vocabulário controlado Tesouro Unesp, as análises quantitativa e qualitativa na análise dos logs se configuraram como as mais adequadas por ponderar variáveis

diferenciadoras (qualitativas) em um grande volume de registros (quantitativos). A pesquisa se limitou à proposição de modelo composto por listagens de palavras passíveis de revisão e atualização.

Nas seções 5 e 6, serão apresentados os resultados da análise descritiva referente ao desenvolvimento da metodologia e da ferramenta RILogUser, com base na metodologia de análise de logs, fundamentada pelas seções 2 e 3.

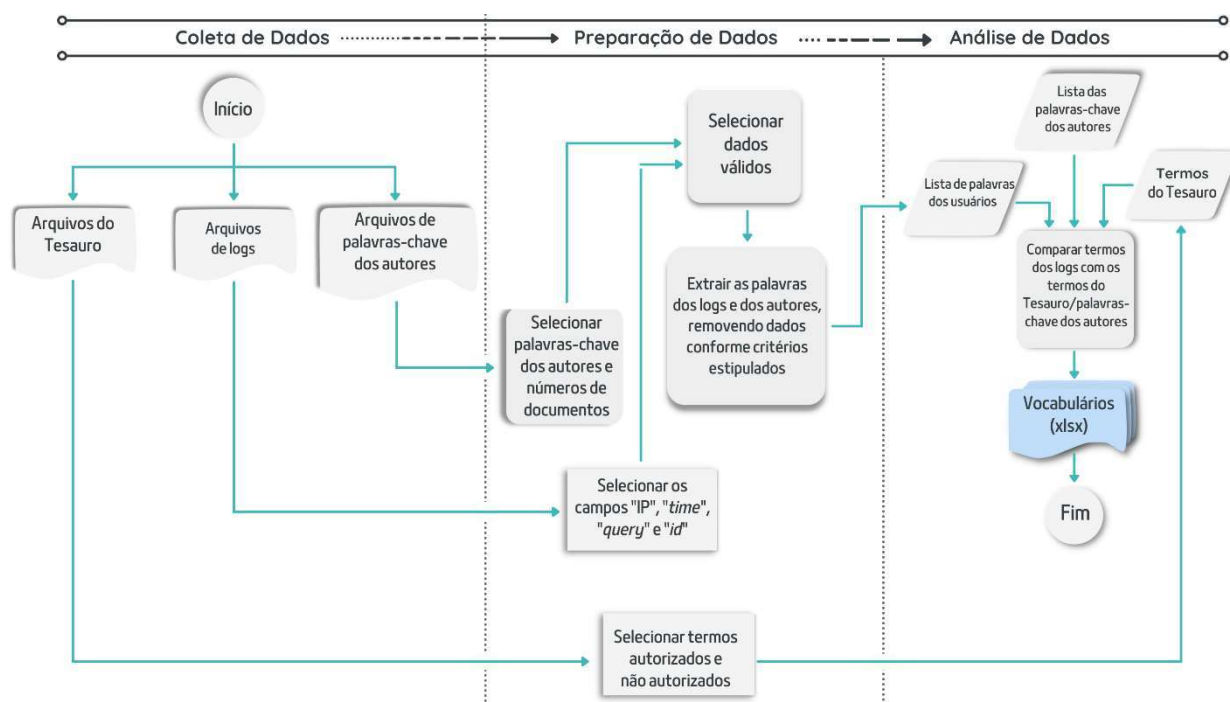
5 PROPOSTA PARA ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO POR MEIO DA ANÁLISE DE LOGS DE BUSCA

A partir do reconhecimento dos desafios relacionados ao processo de recuperação de informações em grandes volumes de dados nos repositórios institucionais, a presente pesquisa propõe um método de atualização de vocabulário controlado por meio da TLA. Enfatizamos que a TLA foi escolhida por permitir o contato direto com as palavras de busca empregadas pelos usuários a um custo relativamente baixo em relação a pesquisas exploratórias realizadas por meio da aplicação de questionários ou entrevistas.

O acesso aos logs de busca dos usuários permitiu um mapeamento amplo de seu vocabulário, no intervalo de tempo considerado, destacando-se que a atualização de um vocabulário controlado a partir da percepção de mudanças na linguagem natural dos usuários é uma atividade permanente, uma vez que o conjunto de palavras circunscritas à área é renovada constantemente (RAUTENBERG; SOUZA; KELNIAR, 2021).

No método proposto, são selecionados quatro campos de buscas dos usuários para investigação: IP para identificação da máquina utilizada na pesquisa; *time* para apuração da data, hora; *query* contendo palavras de busca do usuário e *id* que indica as pesquisas que geraram abertura de arquivo.

Com o conhecimento do vocabulário de busca dos usuários, há a possibilidade de adicionar outras listas de vocabulários, propondo uma análise comparativa entre o vocabulário dos usuários coletado nos logs, a lista de palavras-chave escolhidas pelos autores e até mesmo com vocabulários controlados como tesouros. No caso específico desta Tese, o vocabulário em linguagem natural dos usuários é comparado com as palavras-chave (também em linguagem natural) dos autores e com a lista de termos do Tesouro Unesp. A Figura 16 traz a representação das etapas programadas:

Figura 16. Etapas para análise dos logs de busca extraídos do RI Unesp

Fonte: elaborada pelo autor.

A coleta de dados foi composta pela entrada dos logs de buscas dos usuários no RI Unesp, em formato xlsx; entrada das palavras-chave dos autores via *crawler* e entrada dos termos do Tesauro Unesp via API, ambos também no formato xlsx. Sequencialmente, na preparação dos dados, houve diversas ações para selecioná-los pelo potencial de interesse, a seleção dos campos dos logs: *IP*, os quais passaram por processo de anonimização para garantir a privacidade dos usuários, *time*, *query* e *id*; a seleção de palavras-chave dos autores associada ao número de documentos os quais elas representam; e a seleção de termos do Tesauro Unesp como vocabulário controlado de referência.

De forma amostral, a Figura 17 exemplifica a planilha com dados extraídos pela coleta de logs já selecionando apenas as linhas de registro com *id*, por indicarem abertura do arquivo recuperado.

Figura 17. Coleta de logs extraídos contendo *id*

	ip	query	time	untryCo	referrer	page	city	id
1								
2	1	energia renovável	2019-05-16T22:00:00	BR	https://repositorio.unesp.br/di/	Formiga		156039
3	7	na labuta:	2019-05-16T22:00:00	BR	https://repositorio.unesp.br/di/	Sao Bernardo		184547
4	9	celso júnior	2019-05-17T03:00:00	BR	https://repositorio.unesp.br/di/	Guaratinguet		167487
5	40	malassezia	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Avare		11395
6	43	doce	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/ha	Sao Paulo		105366
7	45	malassezia	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Avare		116556
8	46	malassezia	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Avare		116556
9	49	wernick complexo socorro	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Sao Bernardo		66854
10	50	malassezia	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Avare		88439
11	51	malassezia	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Avare		88439
12	52	malassezia	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Avare		88439
13	53	malassezia	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Avare		86687
14	60	Paula Gripa	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Bauru		154759
15	65	Teoria de funções elípticas e aplicações em soluções de sistemas	2019-05-17T08:00:00	BR	https://repositorio.unesp.br/di/	Rio Claro		155723
16	71	motivação	2019-05-17T09:00:00	BR	https://repositorio.unesp.br/di/	Assis		134259
17	88	Babesia vivia gado leiteiro	2019-05-17T07:00:00	BR	https://repositorio.unesp.br/discover?scope=			100428
18	98	matite homeopatia	2019-05-17T08:00:00	BR	https://repositorio.unesp.br/di/	Tapira		97469
19	109	Gado de leite	2019-05-17T07:00:00	BR	https://repositorio.unesp.br/discover?scope=			141661
20	114	Ana Carolina Bosco Mendes	2019-05-17T08:00:00	BR	https://repositorio.unesp.br/discover?scope=			184457
21	158	Oral health of the aged: an unknown reality	2019-05-19T13:00:00	BR	https://repositorio.unesp.br/ha	Ourinhos		30761
22	170	juliano borba de moraes farinelli	2019-05-19T13:00:00	BR	https://repositorio.unesp.br/di/	Bertioga		90838

Fonte: logs extraídos do RI Unesp (2020).

Para a seleção das colunas de interesse, foi definido código com parâmetros editáveis, conforme Figura 18, na qual, nas linhas 23 e 24, são definidos os campos que serão extraídos para análise:

Figura 18. Código para seleção de dados dos logs

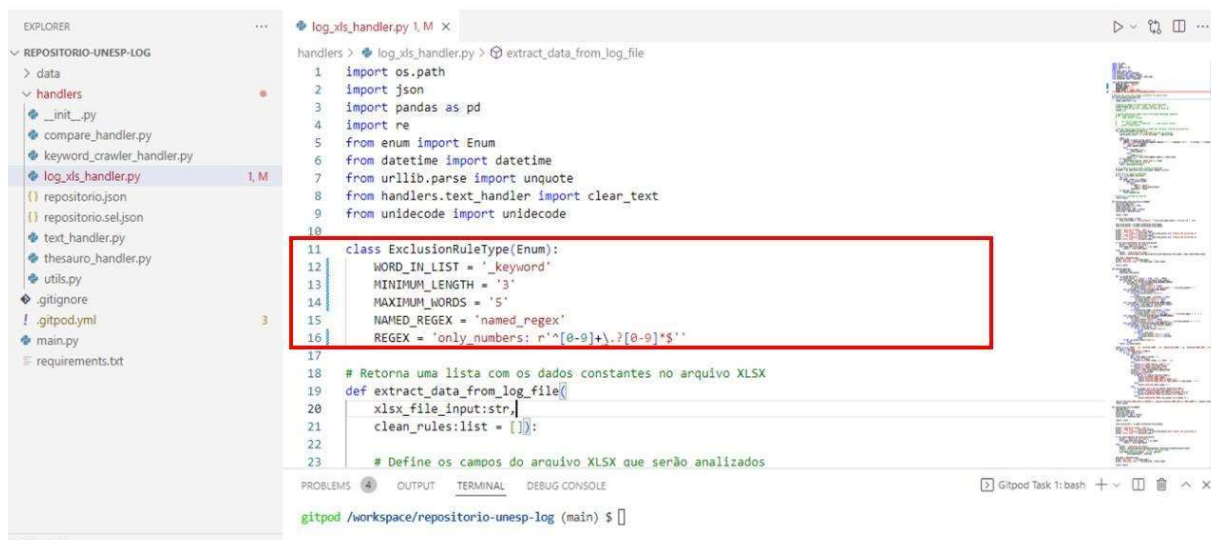
```

12 WORD_IN_LIST = 'word_in_list'
13 MINIMUM_LENGTH = 'minimum_length'
14 MAXIMUM_WORDS = 'maximum_words'
15 NAMED_REGEX = 'named_regex'
16 REGEX = 'regex'
17
18 # Retorna uma lista com os dados constantes no arquivo XLSX
19 def extract_data_from_log_file(
20     xlsx_file_input:str,
21     clean_rules:list = []):
22
23     # Define os campos do arquivo XLSX que serão analisados
24     select_fields:list = ['ip', 'time', 'query', 'id']
25     # Cria a variável que será retornada ao final da execução
26     result = []
  
```

Fonte: elaborada pelo autor.

Esclarecemos também que a preparação dos dados para contabilização dos campos selecionados nos logs de buscas incluiu a limpeza e seleção de dados válidos de buscas, isto é, retirada dos caracteres inválidos em cada log e consequente redução da quantidade de dados, também por meio de critérios editáveis, como demonstra e detalha as linhas de 11 a 16 da Figura 19:

Figura 19. Código para edição de critérios de exclusão



```

1 import os.path
2 import json
3 import pandas as pd
4 import re
5 from enum import Enum
6 from datetime import datetime
7 from urllib.parse import unquote
8 from handlers.text_handler import clean_text
9 from unidecode import unidecode
10
11 class ExclusionRuleType(Enum):
12     WORD_IN_LIST = 'keyword'
13     MINIMUM_LENGTH = '3'
14     MAXIMUM_WORDS = '5'
15     NAMED_REGEX = 'named_regex'
16     REGEX = 'only numbers: r'^[0-9]+\.[0-9]*$'
17
18 # Retorna uma lista com os dados constantes no arquivo XLSX
19 def extract_data_from_log_file(
20     xlsx_file_input: str,
21     clean_rules: list = []):
22
23     # Define os campos do arquivo XLSX que serão analisados

```

Fonte: elaborada pelo autor.

Executados os códigos, foram mantidas apenas as colunas de interesse. As linhas destacadas em vermelho foram excluídas automaticamente pelos critérios apontados nos códigos da Figura 19, os quais referem-se à exclusão de nomes próprios de pessoas; de municípios; palavras com menos de três caracteres; expressões com mais de cinco palavras ou constituídas apenas por caracteres numéricos. Tais critérios objetivaram a obtenção de uma listagem mais significativa, cuja aplicação é demonstrada na Figura 20:

Figura 20. Exclusão de logs conforme critérios estabelecidos

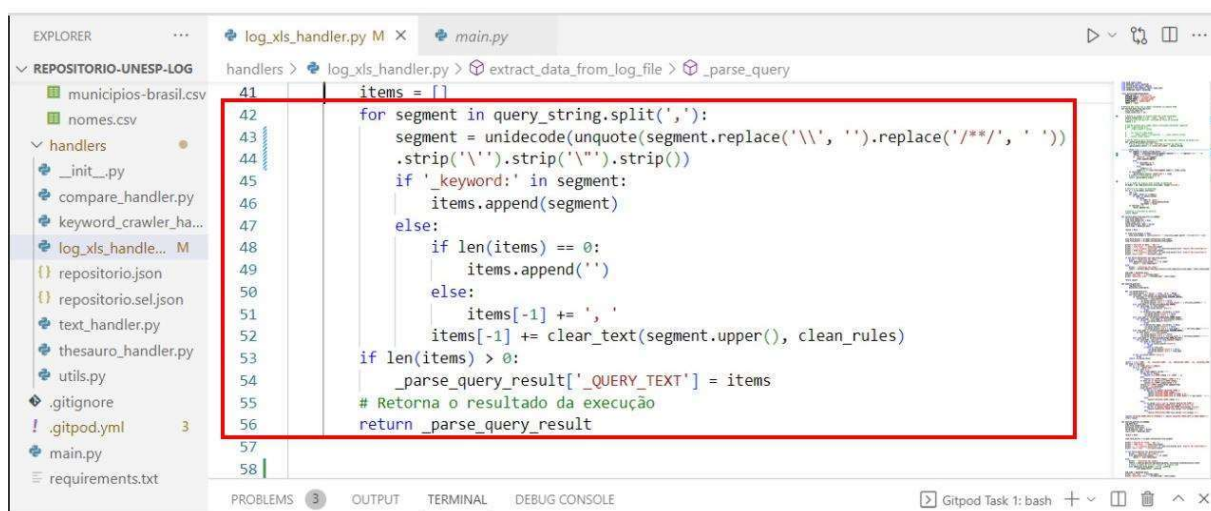
	ip	query	time	id
1				
2	1	energia renovável	2019-05-16T21:	156039
3	7	na labuta:	2019-05-16T21:	184547
4	9	celso júnior	2019-05-17T03:	167487
5	40	malassezia	2019-05-17T09:	11395
6	43	doce	2019-05-17T09:	105366
7	45	malassezia	2019-05-17T09:	116556
8	46	malassezia	2019-05-17T09:	116556
9	49	wernick complexo socorro	2019-05-17T09:	66854
10	50	malassezia	2019-05-17T09:	88439
11	51	malassezia	2019-05-17T09:	88439
12	52	malassezia	2019-05-17T09:	88439
13	53	malassezia	2019-05-17T09:	86687
14	60	Paula Grippa	2019-05-17T09:	154759
15	65	Teoria de funções elípticas e aplicações em soluções de sistemas periódicos em mecânica	2019-05-17T08:	155723
16	71	motivação	2019-05-17T09:	134259
17	88	Babesia vivia gado leiteiro	2019-05-17T07:	100428
18	98	mastite homeopatia	2019-05-17T08:	97469
19	109	Gado de leite	2019-05-17T07:	141661
20	114	Ana Carolina Bosco Mendes	2019-05-17T08:	184457
21	158	Oral health of the aged: an unknown reality	2019-05-19T13:	30761
22	170	juliano borba de moraes farinelli	2019-05-19T13:	90838

Fonte: logs extraídos do RI Unesp (2020), editados pelo autor.

As linhas 9, 49, 60, 114 e 170 foram excluídas, portanto, por meio da função definida na linha 12 da Figura 19 por conterem nomes de pessoas ou de municípios por meio da comparação com as listagens carregadas. Expressões com mais de cinco palavras foram excluídas por meio do código da linha 14 (Figura 19), como foi o caso da linha 65 da Figura 20.

Na sequência, ainda para tratamento dos logs, houve a seleção para remover dados duplicados, por meio do código apresentado na Figura 21:

Figura 21. Código para exclusão de palavras duplicadas nos logs



```

41 items = []
42 for segment in query_string.split(','):
43     segment = unicode(unquote(segment.replace('\\', ' ').replace('/**/', ' ')))
44     .strip('\\').strip('\"').strip()
45     if '_keyword:' in segment:
46         items.append(segment)
47     else:
48         if len(items) == 0:
49             items.append('')
50         else:
51             items[-1] += ', '
52             items[-1] += clear_text(segment.upper(), clean_rules)
53 if len(items) > 0:
54     _parse_query_result['_QUERY_TEXT'] = items
55 # Retorna o resultado da execução
56 return _parse_query_result
57
58

```

Fonte: elaborada pelo autor.

Com a execução, as linhas com dados duplicados (destacadas com fonte vermelha) foram excluídas, conforme exemplificado na Figura 22:

Figura 22. Exclusão de palavras duplicadas

		ip	query	time	id
1					
2	1	ANONIMIZAÇÃO	energia renovável	2019-05-16T21:03:09.663Z	156039
3	7		na labuta:	2019-05-16T21:26:10.776Z	184547
4	40		malassezia	2019-05-17T09:23:22.064Z	11395
5	43		doce	2019-05-17T09:24:01.605Z	105366
6	45		malassezia	2019-05-17T09:24:39.411Z	116556
7	46		malassezia	2019-05-17T09:24:39.494Z	116556
8	50		malassezia	2019-05-17T09:25:14.69Z	88439
9	51		malassezia	2019-05-17T09:25:14.729Z	88439
10	52		malassezia	2019-05-17T09:25:14.758Z	88439
11	53		malassezia	2019-05-17T09:26:22.29Z	86687
12	71		motivação	2019-05-17T09:03:11.316Z	134259
13	88		Babesia vivia gado leiteiro	2019-05-17T07:40:54.038Z	100428
14	98		mastite homeopatia	2019-05-17T08:02:44.86Z	97469
15	109		Gado de leite	2019-05-17T07:38:17.088Z	141661
16	158		Oral health of the aged: an unknown reality	2019-05-19T13:40:23.069Z	30761

Fonte: logs extraídos do RI Unesp (2020), editados pelo autor.

No que concerne à coleta das palavras-chave empregadas pelos autores para representação de seus documentos, a coleta foi realizada por meio de um módulo específico desenvolvido dentro da ferramenta RILogUser com a codificação de um *web crawler*. A Figura 23 traz parte do código de modo amostral:

Figura 23. Fragmento do código do *web crawler* para extração de palavras-chave

```

10 def get_repositorio():
11     offset = 0 # Ponteiro de início de paginação
12     npp = 10000 # Número de itens por página
13
14     keywords = {} # Dicionário que armazenará as palavras-chave encontradas
15
16     while(True):
17         # Variável que armazenará a string HTML da página
18         html = ''
19         # URL à ser consultada adicionando o "npp" e "offset" atuais
20         url = 'https://repositorio.unesp.br/browse?rpp='+str(npp)+'&sort_by=-1&type=subject&offset='+str(offset)+'&etal=-1&order=ASC'
21         # Realiza a requisição HTTP e armazena o retorno em "req"
22         req = requests.get(url)
23         # Captura o texto da página (HTML) para a variável "html"
24         html = req.text
25         # Realiza o parser do "html" retornando um objeto navegável em "soup"
26         soup = BeautifulSoup(html, 'html.parser')
27

```

Fonte: elaborada pelo autor.

Através do código desenvolvido, o *web crawler* percorreu todas as páginas, de A a Z, incluindo também as páginas com expressões iniciadas por números (0-9), do RI Unesp e coletou todas as palavras-chave empregadas pelos autores, como ilustra a Figura 24:

Figura 24. Coleta de palavras-chave no RI Unesp

Fonte: RI Unesp (2022), editado pelo autor.

Após extração, houve a limpeza e seleção de palavras-chave válidas, isto é, com mais de um caractere e menos que seis palavras (parâmetro configurável) e a eliminação de erros conforme exemplificado no Quadro 6:

Quadro 6. Exemplo de erros detectados na listagem de palavras-chave dos autores

Letras ou palavras sem sentido	B[1] ; P[1] ; S[1] ; V%[1]
Maiúsculas-Minúsculas	Árvores tropicais [1] / árvores tropicais [1]; Cardiopatia [2] / cardiopatia [1]; Cardiopatias [1] / cardiopatias [1]; C-BAND [4] / C-band [23] C-Fos [2] / C-fos [1] / c-Fos [2] E-Commerce [2] / E-commerce [2] / e-commerce [1] Rabbit [45] / rabbit [82] Y Chromosome [3] / Y chromosome [7]
Acentos	Karate-do [1] / Karate-dô [1]
Outros (aspas, hífen, datas,	Karl Marx [2] / Karl Marx (1818-1883) [1] N potencial mineralizável [1] / N potencialmente mineralizável [2] Y chromosome [7] / Y-chromosome [2] "Candidatus" Liberibacter spp. "Desenvolvimento Pacífico" "Ethos" della mobilità del servizio sociale brasiliano "Geração de 45" "Kissing spines" "Mãe Cristina" Municipal Nursery "Ovsynch" "Pivô" para a Ásia "São Paulo do School" Program "Upconversion" "de novo" assembly "Órfãos do Eldorado"

	#ErectileDysfunction #ProstateCancer #elenão #elesim –Active Learning –Aprendizagem Ativa –Educação Médica
--	--

Fonte: elaborado pelo autor a partir do Repositório Unesp.

Sequencialmente, o Quadro 7 exemplifica a perda de recuperação que pode ocorrer apenas devido à variação de letras maiúsculas ou minúsculas ou diferenciação de singular e plural, sendo, ao todo, cinco títulos a serem recuperados, mas que se encontram separados devido a mínimas diferenciações:

Quadro 7. Inconsistências de palavras-chave na recuperação de documentos

Palavra de busca digitada para recuperação	Título recuperado
Cardiopatía	Dispersão da onda P em cães com doença valvar mitral e cardiomiopatia arritmogênica do ventrículo direito
	Exercício resistido no cardiopata: revisão sistemática
cardiopatía	<i>Holter electrocardiography in dogs showing doxorubicin-induced dilated cardiomyopathy</i>
Cardiopatías	<i>Factors influencing psychotropic prescription by non-psychiatrist physicians in a nursing home for the elderly in Brazil</i>
cardiopatías	Remodelação miocárdica pós-lesão cardíaca induzida por videocirurgia, em tilápias-do-nilo

Fonte: elaborado pelo autor a partir do RI Unesp (24 mar. 2022).

A Figura 25 exemplifica outras ocorrências em que a ausência de padronização pode resultar quantitativos recuperados diferentes e perdas informacionais. Na figura, observamos que a mesma palavra de busca, apenas com variações de letras maiúsculas e minúsculas gera resultados diferentes indicados entre colchetes:

Figura 25: Exemplos da localização de erros na lista por tema



Fonte: RI Unesp (2022), editado pelo autor.

Por fim, ao que se refere à coleta de Termos do Tesauro Unesp, também foi programado um módulo integrante da ferramenta RILogUser, o qual, a partir de padrões definidos e disponibilizados pela API do próprio site do Tesauro Unesp, viabilizou a coleta de dados. Parte do código é exibido, de maneira amostral, na Figura 26:

Figura 26. Fragmento do código para extração de termos, via API, do Tesauro Unesp

```

EXPLORER  ...  log_xls_handler.py M  keyword_crawler_handler.py 2, M  thesauro_handler.py 2, M X
REPOSITORIO-UNESP-LOG
  products
    .gitignore
    repositario.json
    thesauro.json
  raw
    estatisticas-2sem2...
    municipios-brasil.csv
    nomes.csv
  handlers
    __init__.py
    compare_handler.py
    keyword_cra... 2, M
    log_xls_handle... M
    repositario.json
    repositario.sel.json
    text_handler.py
    thesauro_ha... 2, M
  OUTLINE
  TIRACI IAC

handlers > thesauro_handler.py > get_tesauro
18 def get_tesauro():
19     result = result = {'terms': {}, 'alternative_terms': {}, 'relationships': {}
20     {}, 'info': {'average': 0, 'count': 0, 'count_words': 0}}
21
22 def count_word(word: str):
23     result['info']['count'] += 1
24     result['info']['count_words'] += word.count(' ') + 1
25
26 # if len(result['terms']) == 0:
27 for letter in string.ascii_lowercase[:26]:
28     url = 'https://portal.biblioteca.unesp.br/tematres/vocab/services.php?task=letter&arg=' + letter
29     r = requests.get(url)
30     for child in ET.fromstring(r.content).iter('term'):
31         if child.find('string').text is not None:
32             s = prepare_string(child.find('string').text)
33             if child.find('term_id').text not in result['terms']:
34                 count_word(s)
35
PROBLEMS 7  OUTPUT  TERMINAL  DEBUG CONSOLE
gitpod /workspace/repositorio-unesp-log (main) $

```

Fonte: elaborada pelo autor.

Através do código desenvolvido, a API percorreu todas as páginas, de A a Z, incluindo também as páginas com expressões iniciadas por números (0-9), do Tesouro Unesp e coletou todos os termos, como ilustra a Figura 27:

Figura 27. Coleta de termos do Tesouro Unesp



Fonte: Tesouro Unesp (2022), editado pelo autor.

Exposta a metodologia empregada para a extração e seleção dos dados das três listagens, esclarecemos que a análise dos logs é realizada por meio da comparação às outras duas listagens: a de palavras-chave dos autores e a dos termos do Tesouro Unesp. Tal proposta se justifica pôr as palavras-chave dos autores ser outra fonte de dados representativos da linguagem natural e os termos do Tesouro comporem um vocabulário controlado já adotado pela Unesp, enquanto instituição. Tal comparação fornece subsídios para a proposta de atualizar o vocabulário controlado e otimizar o processo de busca e recuperação de documentos. A seguir, são detalhadas as etapas dos procedimentos de coleta, preparação e análise de dados.

5.1 Coleta de dados

No que tange aos dados coletados, temos: logs, palavras-chave dos autores e termos do Tesouro Unesp.

Quanto aos logs de busca dos usuários, são referentes ao período de 01 de julho a 31 de dezembro de 2019 e foram disponibilizados pelos responsáveis do RI Unesp, totalizando 480.742 requisições de usuários registradas em logs da base do Apache Solr, plataforma de

pesquisa corporativa de código aberto, escrita em Java, que alimenta os recursos de pesquisa e navegação de muitos dos maiores sites de internet do mundo por se tratar de uma plataforma altamente confiável, escalável e tolerante a falhas, fornecendo indexação distribuída, replicação e consulta balanceada com carga, *failover* e recuperação automatizados e configuração centralizada.

Na coleta primária de dados dos logs, obteve-se 34 campos em cada registro de acesso, no entanto, de acordo com o objetivo da pesquisa, foi realizado um recorte mantendo-se quatro campos (IP, *time*, *query* e *id*).

Realizada a coleta de dados, como primeira etapa de organização, para aumentar a consistência e a confiabilidade, seguiu-se com a etapa de preparação dos dados por meio da limpeza, de modo a eliminar as incorreções e restringir a análise das palavras buscadas que geraram abertura de documentos.

Quanto às palavras-chave dos autores, a coleta aconteceu no RI Unesp no dia 24 de março de 2022, por meio de *web crawler* capaz de percorrer as 32 páginas *web*, cada uma com aproximadamente 10.000 palavras-chave, totalizando 325.524 palavras-chave de autores coletadas, de todas as áreas do conhecimento, atribuídas no RI Unesp desde a sua implantação até a data de coleta das informações. O endereço para acesso às palavras-chave dos autores está disponível em: <https://repositorio.unesp.br/browse?type=subject>.

Por fim, não menos importante, a lista dos termos pertencentes ao Tesouro Unesp também foi coletada de forma automática, distinguindo-se os termos autorizados dos não autorizados. Para tanto, foi desenvolvido um *script* em Python para acessar a API disponível pelo Tesouro, isto é, as instruções programadas em Python permitiram o acesso a uma série de rotinas e padrões de programação para extrair os termos do Tesouro da plataforma na *web*, em formato xml (*Extensible Markup Language*). O formato xml foi escolhido por propiciar o compartilhamento de informações mediadas pela internet. A coleta dos dados foi realizada em 24 de março de 2022. Nesta primeira listagem, constaram 40.351 termos autorizados e 49.237 não autorizados, totalizando 89.588 termos elencados no Tesouro. O endereço da API está disponível em: <http://portal.biblioteca.unesp.br/tematres/vocab/services.php>.

Em posse das listagens de vocabulário oriundas da coleta computacional, o passo metodológico seguinte se concentrou na preparação do grande volume de dados para posterior análise.

5.2 Preparação de dados

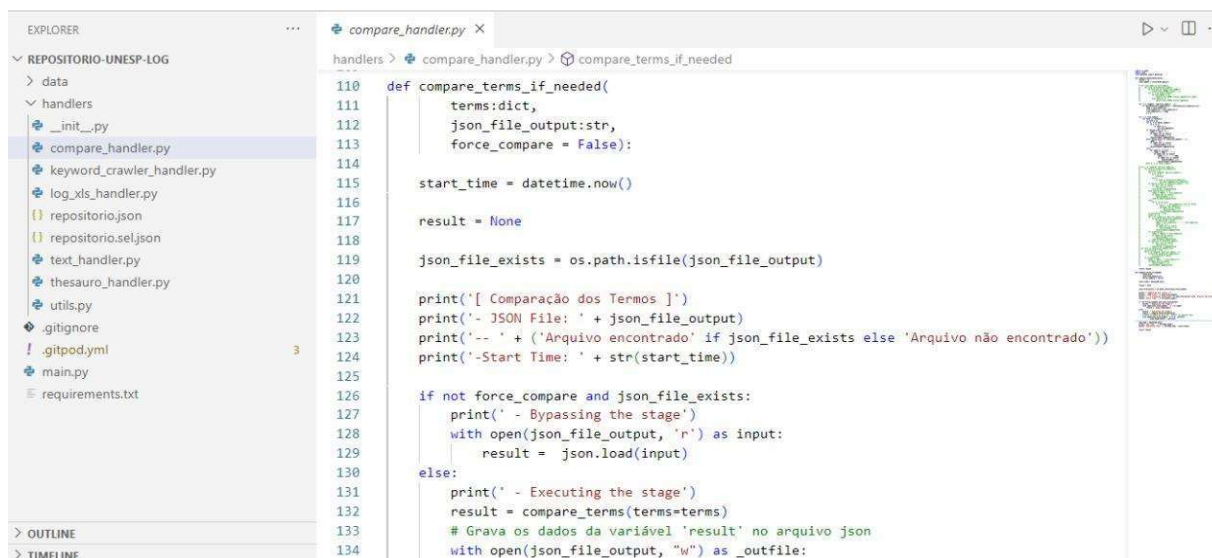
Em posse dos dados exportados da base do Apache Solr, para realizar o processamento inicial e a preservação do anonimato com relação à identidade dos usuários, respeitando o disposto pela Lei nº 13.709/2018, identificada como Lei Geral de Proteção de Dados Pessoais (LGPD), os arquivos referenciais foram submetidos a um processo de anonimização dos endereços IP. Por anonimização a Lei define, em seu Art. 5º: “dado relativo à titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento” isso para garantir o respeito à privacidade e a inviolabilidade da intimidade, da honra e da imagem, conforme disposto no Art. 2º da referida Lei (BRASIL, 2018).

Sequencialmente, já garantida a privacidade dos usuários, procedemos a limpeza dos dados em três etapas por meio de estratégias computacionais implementadas na ferramenta RILogUser. Na primeira, foram excluídos, por meio de expressões regulares, todos os registros: com campo *id* em branco; com campo *query* sem nenhum caractere ou com mais de cinco palavras.

Na segunda etapa, realizamos o processamento das palavras, por meio do agrupamento de palavras-chave que apresentavam apenas variações de caracteres maiúscula/minúscula (de duplicação). Com essas ações, foi possível a obtenção de uma lista depurada com palavras de buscas dos usuários.

5.3 Análise de dados

A análise de dados continua por meio do módulo de comparação, integrante da ferramenta RILogUser, o qual compara as três listagens coletadas e tratadas anteriormente. Ressaltamos: a primeira lista é composta pelas palavras de buscas dos usuários; a segunda, com as palavras-chave dos autores e a terceira, com os termos do Tesouro Unesp. O código elaborado para automatização da comparação dos termos é exibido na Figura 28:

Figura 28. Fragmento do código do módulo desenvolvido para comparação das listagens

Fonte: elaborada pelo autor.

As comparações realizadas geram produtos, isto é, listas secundárias formadas pela identificação de elementos comuns e exclusivos de cada listagem. A Figura 29 expõe a página *on-line* em que é possível fazer as solicitações de listagens consideradas produtos fornecidos pela ferramenta.

Figura 29. Página da RILogUser para extração de produtos: módulo *on-line*

REPOSITÓRIO INSTITUCIONAL UNESP **RILogUser**

ANÁLISE DE LOGS DE BUSCA COMO FONTE INFORMACIONAL PARA ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO EM REPOSITÓRIOS INSTITUCIONAIS

Preparação de dados

Dados completos analisados

Voltar

Os produtos das comparações entre as listas viabilizam a identificação de conjuntos compostos por elementos comuns, nomeados no diagrama: V4, V5, V6 e V7.

Coleta de logs de buscas: busca-usuarios.xlsx

Coleta de palavras-chave de autores: chave-autores.xlsx

Coleta de termos do Tesauro: tesauro.xlsx

Vocabulários criados:

- ☒ Palavras exclusivas dos logs de busca (V1)
- ☐ Termos exclusivos do Tesauro (V2)
- ☐ Palavras exclusivas das palavras-chave (V3)
- ☐ Palavras comuns: logs de busca e palavras-chave (V4)
- ☐ Termos comuns: logs de busca e termos Tesauro (V5)
- ☐ Termos comuns: palavras-chave e termos Tesauro (V6)
- ☐ Intersecção de todos (V7)

Limite máximo: 300 linhas

Produtos

Análise de dados

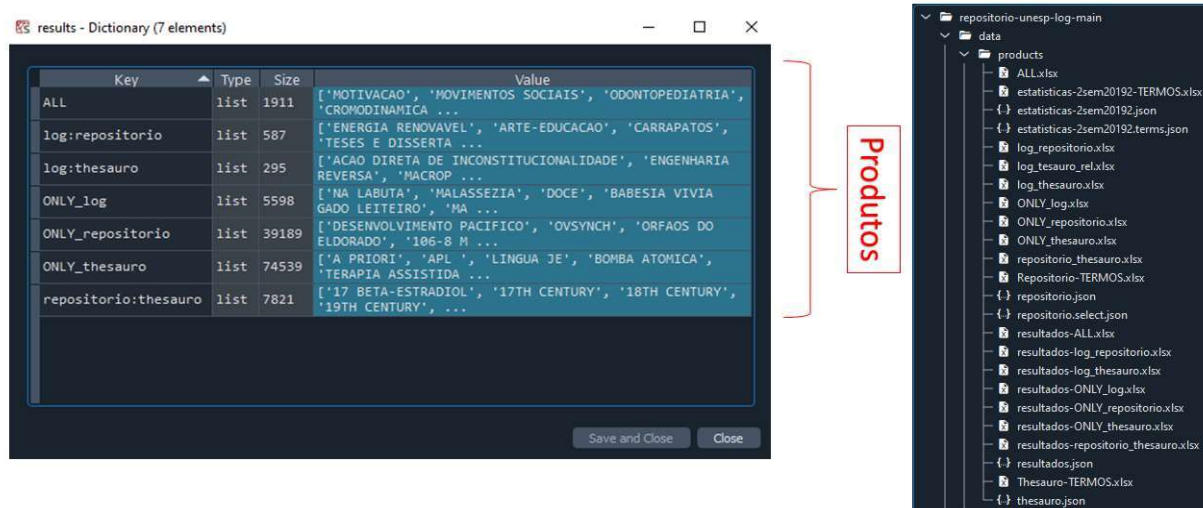
Diagrama de Venn:

- V1: Palavras de buscas dos usuários
- V2: Termos do Tesauro Unesp
- V3: Palavras-chave dos autores
- V4: Palavras comuns: logs de busca e palavras-chave
- V5: Termos comuns: logs de busca e termos Tesauro
- V6: Termos comuns: palavras-chave e termos Tesauro
- V7: Intersecção de todos

Fonte: elaborada pelo autor.

A fim de reduzir a latência, ou seja, o tempo entre a requisição e a efetiva liberação dos recursos solicitados, a análise dos dados completos é realizada pela ferramenta RILogUser módulo *off-line*, conforme exibido pela Figura 30:

Figura 30. Página da RILogUser para extração de produtos: módulo *off-line*



Fonte: elaborada pelo autor.

Para ilustrar a obtenção das listagens de vocabulários, as palavras comuns são nomeadas no diagrama da Figura 31 de V4, V5, V6 e V7; onde V7 corresponde à compatibilização dos três conjuntos; V5 corresponde aos termos comuns apenas ao Tesauro e ao vocabulário dos usuários; V6 aos termos comuns apenas ao Tesauro e às palavras-chave dos autores e V4 as palavras comuns apenas às palavras-chave dos autores e as palavras de busca dos usuários.

As palavras exclusivas são nomeadas de V1, V2 e V3; V1 as palavras exclusivas aos logs; V2 aos termos exclusivos ao Tesauro e V3 as palavras-chave exclusivas aos autores. Como exposto na Figura 31:

Figura 31. Diagrama das listas geradas com representação dos vocabulários



Fonte: elaborada pelo autor.

Como representado no diagrama, o vocabulário produto da pesquisa experimental, candidato à atualização do Tesauro Unesp, é composto pelos conjuntos V1 e V4. V1 por conter palavras do vocabulário dos usuários e V4 por representar um subconjunto com palavras comuns a dois vocabulários de linguagem natural. As demais intersecções contêm termos já integrantes do Tesauro da Unesp, portanto não trazem termos a serem incluídos. A delimitação de V4 se justifica, pois trata-se de palavras que estiveram presentes nas palavras-chave dos autores e efetivamente contribuíram para a recuperação de materiais pelos usuários, posto que geraram resultados em que documentos foram abertos.

A listagem com as palavras exclusivas dos logs de busca dos usuários (V1) destaca-se por representar a linguagem empregada pelos usuários quando intentam buscas no RI Unesp. Nessa compreensão, ainda que não constem nas duas outras listagens, devem ser analisadas precisamente por trazerem a linguagem dos usuários, de forma que a representação dos materiais se aproxime mais de seu público.

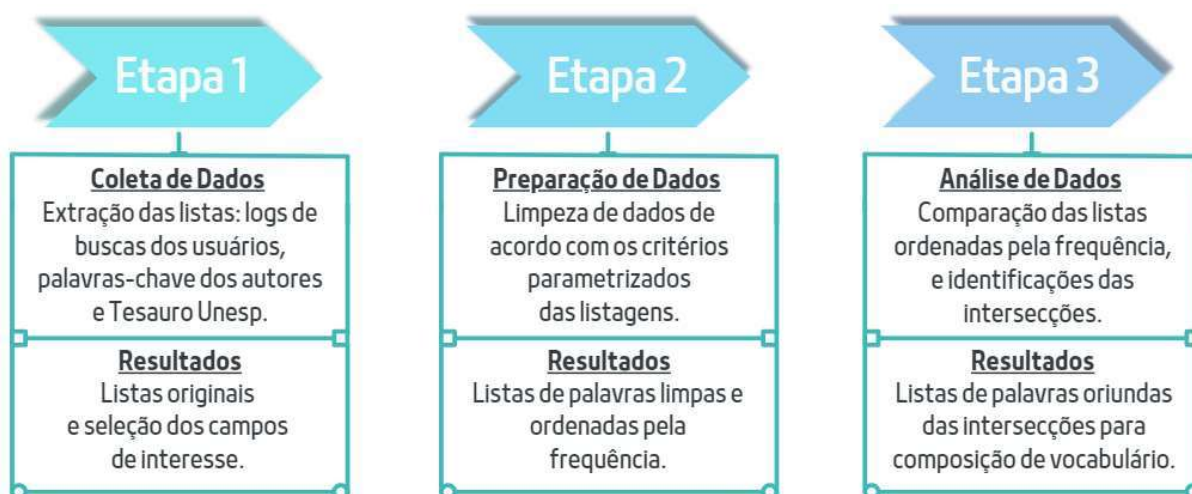
Sequencialmente, apresentamos na seção 6, de modo mais detalhado, a ferramenta computacional RILogUser como resultado da criação do método para atualização de vocabulário controlado por meio da TLA. Assim, são descritos os resultados obtidos pela aplicação experimental com a observação de palavras exclusivas e das intersecções de termos entre as listagens, finalizando com a indicação de lista como proposta de vocabulário controlado obtido pela análise de recorrência das palavras.

6 FERRAMENTA DESENVOLVIDA PARA ANÁLISE DE VOCABULÁRIOS E RESULTADOS

Em decorrência da metodologia para atualização de vocabulários controlados por meio da análise de logs de busca, apresentamos a ferramenta RILogUser. Ao retomar os objetivos específicos como ações essenciais à concretização do objetivo geral de realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais, nesta seção, são apresentados os resultados relativos ao segundo e terceiro objetivos específicos: analisar, a partir de logs de pesquisa feitas pelos usuários no RI Unesp, a viabilidade metodológica de aplicação de recursos computacionais para realização de rotinas automatizadas para depuração / limpeza de logs de pesquisa para composição de lista de palavras candidatas à inclusão no Tesouro Unesp, subsidiando metodologia específica de análise de logs; e, criar e investigar a aplicação experimental da análise de logs implementada por meio de ferramenta computacional desenvolvida durante a produção desta Tese, para a atualização de vocabulário controlado do Tesouro Unesp, com vistas à ampliação de seu uso no RI Unesp.

A relação entre as etapas metodológicas e os resultados da pesquisa é exibida pela Figura 32:

Figura 32. Relação entre etapas metodológicas de TLA e apresentação dos resultados



Fonte: elaborada pelo autor.

Na Figura 32, consta uma visão progressiva das etapas metodológicas associadas à obtenção de resultados da pesquisa. Na próxima subseção, é apresentado o módulo *off-line* da ferramenta RILogUser para atualização de vocabulário controlado a partir da análise de logs de buscas feitas por usuários no RI Unesp. Sequencialmente, para finalizar a apresentação dos resultados, é exposto também o módulo *on-line* da ferramenta RILogUser. No módulo *off-line* é realizado o processamento integral dos dados extraídos, já o módulo *on-line* configura-se como protótipo com interface mais amigável e acessível a não-programadores.

6.1 RILogUser: módulo *off-line*

O módulo *off-line* da RILogUser foi desenvolvido em linguagem Python, a qual é amplamente utilizada em computação e estatística para o processamento de grandes volumes de dados como os obtidos nas listagens de vocabulários descritas. Adicionalmente, foram agregadas várias bibliotecas para facilitar e ampliar o tratamento dos dados e obter listagens com informações produtivas. Todo o código fonte está disponível em: <https://github.com/ikatahira/Doutorado>.

Primeiramente, para a manipulação das listagens, foi adicionada à ferramenta a biblioteca pandas, que é desenvolvida para análise e manipulação de dados de forma eficiente. Sequencialmente, foi adicionada a biblioteca re (*regular expressions*) que fornece operações para correspondência de expressões regulares frequentemente chamadas de *regex*. Para homogeneizar as acentuações, foi adicionada a biblioteca unidecode. Por fim, a biblioteca *Beautiful Soup* foi aplicada para extrair as palavras-chave dos autores das páginas do RI Unesp (HTML e XML), conforme Figura 33.

Figura 33. Principais bibliotecas utilizadas para manipulação das listagens

```

1  import requests
2  import os.path
3  import json
4  import pandas as pd
5  import re
6  from bs4 import BeautifulSoup
7  from datetime import datetime
8  from handlers.log_xls_handler import ExclusionRuleType
9  from unidecode import unidecode

```

Fonte: elaborada pelo autor.

Nesse sentido, o método desenvolvido pode ser particionado em três grandes etapas: coleta; preparação e análise de dados. A coleta é composta por: I) entrada dos logs de buscas

dos usuários em formato *xlsx*; II) entrada das palavras-chave dos autores via *crawler* e III) entrada dos termos do Tesauro Unesp via API. Assim, algumas bibliotecas foram importadas para auxiliar na coleta de dados, como exibido no código exposto na Figura 34.

Figura 34. Bibliotecas utilizadas para a coleta de dados

```

1 import json
2 from pandas import DataFrame
3 from handlers.log_xls_handler import extract_data_from_log_file_if_needed, analyze_queries_if_needed
4 from handlers.utils import _get_nomes, _get_municipios
5 from handlers.text_handler import TextRuleType, get_text_combinations
6 from handlers.thesauro_handler import get_thesauro_if_needed
7 from handlers.keyword_crawler_handler import get_repositorio_if_needed, analyze_keywords_if_needed
8 from handlers.compare_handler import compare_terms_if_needed

```

Fonte: elaborada pelo autor.

Como exibido na Figura 34, cada linha do código possui uma função específica para o funcionamento da ferramenta RILogUser. A linha 1 é basicamente responsável por agilizar o carregamento de dados, posto que o “json” é um formato leve de troca de informações/dados entre sistemas para otimização. Na linha 2, é carregada a biblioteca *pandas* para auxiliar na manipulação e análise de dados. Na linha 3, são carregadas as funções responsáveis pela extração e análise dos logs de busca. Na linha 4, são carregadas as listagens dos nomes próprios para exclusão, conforme critérios estabelecidos. Na linha 5, são carregados os métodos que contêm as expressões regulares a serem excluídas para favorecer a manutenção apenas de dados relevantes. Na linha 6, é carregado o método de acesso à API do Tesauro Unesp. Na linha 7, é chamado o método de rastreamento de todas as palavras-chave dos autores por meio do *web crawler*. Por fim, na linha 8, é desenvolvido o método de comparação que gera as listagens de vocabulários.

Assim, em síntese os códigos das linhas 3, 6 e 7 (Figura 34) são responsáveis pela coleta de dados necessários para alcançarmos o objetivo de estudar, a partir de logs de pesquisa extraídos do RI Unesp, a viabilidade da aplicação de rotinas automatizadas para depuração/limpeza de logs de pesquisa. Portanto, inicialmente, realizamos a coleta do *corpus* a ser analisado, sendo este composto não apenas pelos logs de busca dos usuários do RI Unesp, mas também pela lista de palavras-chave dos autores de materiais incluídos no repositório e pela lista de termos do Tesauro Unesp. Os resultados da coleta de dados estão dispostos na Figura 35:

Figura 35. Resultados das listagens de palavras por fonte de coleta de dados



Fonte: elaborada pelo autor.

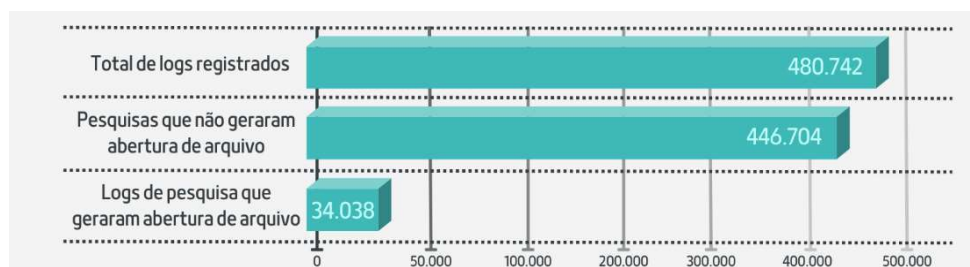
Sequencialmente, são apresentados os resultados obtidos pela preparação e análise de cada listagem de palavras coletada.

6.1.1 Palavras de buscas dos usuários extraídas dos logs

Nesta subseção são descritos os resultados associados à listagem de palavras empregadas pelos usuários para buscas de materiais no RI Unesp, de 01 de julho de 2019 a 31 de dezembro de 2019. Com o tratamento das palavras de buscas dos usuários, foi obtido um *corpus* terminológico proveniente diretamente da linguagem utilizada pelos usuários, de forma precisa e não intrusiva.

Dessa forma, foram contados 480.742 registros de logs nos seis meses coletados. Após essa primeira etapa de limpeza, tendo como critério de exclusão todas as buscas que não geraram abertura de arquivo, obteve-se uma listagem com 34.038 palavras, conforme Figura 36:

Figura 36. Quantitativos da 1ª etapa de seleção dos logs de busca dos usuários



Fonte: elaborada pelo autor.

Conforme Obrien *et al.* (2017), o rastreamento de *downloads* de arquivos pode ser considerado como uma métrica do sucesso do RI por auxiliar na verificação do tráfego do usuário. Como exibido na Figura 36, dos 480.742 logs extraídos inicialmente, verificamos que apenas 34.038 (7,08%) geraram abertura de arquivo pelos usuários. Essa seleção para análise

mais detalhada se justifica pela abertura de arquivo ser um indicativo de sucesso da busca, pois o usuário abre o material buscado quando acredita que este atende sua necessidade.

A partir da seleção de *queries* que geraram abertura de arquivo, a segunda etapa de delimitação do *corpus* composto pelos logs de busca dos usuários foi realizada pela exclusão de 5.731 *queries* em branco (16,83%), isto é, aqueles que tiveram origem por cliques do usuário no RI, sem digitação de qualquer palavra, assim obtivemos o total de 28.307 logs a serem analisados, conforme Tabela 1:

Tabela 1. Quantitativos da 2ª etapa de seleção dos logs de busca dos usuários

Registro de consultas dos usuários	
Período analisado	31-07-2019 a 31-12-2019
Logs de pesquisa que geraram abertura de arquivo	34.038
Logs com <i>queries</i> em branco	5.731
Logs com <i>queries</i> dos usuários	28.307

Fonte: elaborada pelo autor.

Em continuidade, foi observada a necessidade de refinar um pouco mais a limpeza dos dados, por meio da exclusão de *queries* que não traziam informações relevantes para o cumprimento do objetivo da pesquisa. O Quadro 8 exemplifica o processo de depuração:

Quadro 8. Exemplos de depuração de logs de busca dos usuários

Lista sem depuração	Processo de limpeza	Lista depurada
Abacate abacate	Abacate abacate	ABACATE ABATE COMPARATIVO
Abate comparativo abate comparativo	Abate comparativo abate comparativo	
Mônica Brasil Marília	Mônica Brasil Marília	
FRAGMENTOS FLORESTAIS E TESTES DE PROGÊNIES: OPÇÕES PARA A COLETA DE SEMENTES EM ESPÉCIES ARBÓREAS NATIVAS DO CERRADO NO BOLSÃO SUL- MATO-GROSSENSE	FRAGMENTOS FLORESTAIS E TESTES DE PROGÊNIES: OPÇÕES PARA A COLETA DE SEMENTES EM ESPÉCIES ARBÓREAS NATIVAS DO CERRADO NO BOLSÃO SUL- MATO-GROSSENSE	

Fonte: elaborado pelo autor.

Observamos, no Quadro 8, a unificação de expressões com variações no uso de letras minúsculas e maiúsculas, como em “Abacate” e “abacate”, “Abate comparativo” e “abate comparativo”. Também há a exclusão de substantivos próprios referentes a nomes de pessoas e localidades como exemplificado as palavras “Mônica”, “Brasil” e “Marília”. Por fim, a exclusão de “FRAGMENTOS FLORESTAIS E TESTES DE PROGÊNIES: OPÇÕES PARA A COLETA DE SEMENTES EM ESPÉCIES ARBÓREAS NATIVAS DO CERRADO NO BOLSÃO SUL-MATO-GROSSENSE” se deu por corresponder à expressão literal com reprodução do título de um material.

Após exclusões de acordo com os critérios exemplificados no Quadro 8, a Tabela 2 dimensiona quantitativamente os resultados desta 3ª etapa de limpeza:

Tabela 2. Quantitativos da 3ª etapa de limpeza com aplicação de rotinas automatizadas para depuração de logs de pesquisa

Registro de consultas dos usuários	
Logs com <i>queries</i> dos usuários	28.307
Logs com <i>queries</i> duplicadas	13.359
Nomes de pessoas	5.631
Nomes de municípios	572
Expressões literais (com mais de 5 palavras)	1.383
Total de <i>queries</i> válidos	7.362
Extensão média de palavras dos logs válidos	2,36

Fonte: elaborada pelo autor.

Como visualizamos na Tabela 2, a terceira etapa de limpeza reduziu o conjunto de dados conforme aplicação de critérios: I) 13.359 (47,19%) logs com registros duplicados diferenciados apenas por uso de maiúsculas e minúsculas; II) 5.631 (19,89%) logs removidos por constarem nos registros de nomes próprios pela identificação com nomes presentes na classificação de gênero em nomes brasileiros com base em dados do CENSO de 2010²; III) 572 (2,02%) remoções de substantivos próprios referentes a cidades brasileiras de acordo com dados

²Disponível em: <https://brasil.io/dataset/genero-nomes/nomes/>. Acesso em: 23 mar. 2021.

do Instituto Brasileiro de Geografia e Estatística (IBGE)³; e, IV) 1.383 (4,89%) expressões literais (remoção das frases com 6 ou mais palavras).

A Figura 37 apresenta, da linha 145 à 175, todo o processamento para limpeza e depuração dos logs de busca dos usuários. Destaca-se a variável 'QUERY_TEXT', responsável por armazenar as palavras de buscas dos usuários e a lista de dados nomeada *result*, a qual armazena todos os retornos quantitativos e conteudistas de cada um dos critérios descritos na Tabela 2:

Figura 37. Fragmento dos códigos responsáveis pela limpeza dos logs de busca

```

145 result = {'ALL_TERMS' : {}, 'SELECTED_TERMS' : {}, 'UNSELECTED_TERMS' : {}, 'SELECTED_TERMS_INFO' : {'AVERAGE' : 0, 'COUNT' : 0, 'COUNT_WORDS' : 0}}
146 for item in log_data:
147     for k, v in item['query'].items():
148         if k == 'QUERY_TEXT':
149             for term in v:
150                 key = term.upper().strip('\n')
151                 if key not in result:
152                     result['ALL_TERMS'][key] = {'_COUNT' : 1}
153                 else:
154                     result['ALL_TERMS'][key]['_COUNT'] += 1
155                 if term not in result['ALL_TERMS'][key]:
156                     result['ALL_TERMS'][key][term] = []
157                     result['ALL_TERMS'][key][term].append(item)
158                 is_valid = _is_valid(key)
159                 if(is_valid['result']):
160                     if key not in result['SELECTED_TERMS']:
161                         result['SELECTED_TERMS'][key] = 1
162                         result['SELECTED_TERMS_INFO']['COUNT'] += 1
163                         result['SELECTED_TERMS_INFO']['COUNT_WORDS'] += key.count(' ') + 1
164                     else:
165                         result['SELECTED_TERMS'][key] += 1
166                 else:
167                     if is_valid['rule'] not in result['UNSELECTED_TERMS']:
168                         result['UNSELECTED_TERMS'][is_valid['rule']] = {}
169                     if key not in result['UNSELECTED_TERMS'][is_valid['rule']]:
170                         result['UNSELECTED_TERMS'][is_valid['rule']][key] = 1
171                     else:
172                         result['UNSELECTED_TERMS'][is_valid['rule']][key] += 1
173
174 result['SELECTED_TERMS_INFO']['AVERAGE'] = result['SELECTED_TERMS_INFO']['COUNT_WORDS'] / result['SELECTED_TERMS_INFO']['COUNT']
175 return result

```

Fonte: elaborada pelo autor.

Após essa terceira etapa de limpeza das palavras de busca registradas nos logs, a qual foi executada pelos códigos expostos na Figura 37, obteve-se uma lista com 7.362 palavras, o que representa 21,6% da amostra de 34.038 logs que geraram abertura de arquivo e 1,5% do total de 480.742 logs obtidos inicialmente. Essa redução operacionalizada pela aplicação de critérios bem definidos foi essencial para a obtenção de uma listagem significativa.

Destaca-se ainda que a quantidade média de palavras de busca do usuário foi de 2,36 palavras, índice semelhante ao apurado por Mahoui e Cunningham (2001) no estudo do comportamento de pesquisa de usuários, na biblioteca digital de literatura científica ResearchIndex, que revelou a preferência desses usuários por utilizarem três ou menos palavras de busca, fato que evidencia a importância da seleção adequada e precisa de palavras para a representação.

³Disponível em: http://blog.mds.gov.br/redesuas/wp-content/uploads/2018/06/Lista_Munic%C3%ADpios_com_IBGE_Brasil_Versao_CSV.csv. Acesso em: 31 mar. 2021 (Valores Separados por vírgulas – CSV).

Observa-se, pois, que essa seleção de palavras permitiu uma grande depuração, a qual produziu um *corpus* terminológico proveniente diretamente da linguagem utilizada pelos usuários, alcançando-se o segundo objetivo específico traçado nesta Tese.

Conhecidas as palavras utilizadas pelos usuários do RI, foi fundamental investigarmos também as palavras-chave atribuídas pelos autores ao produzirem seus textos, uma vez que o sucesso da busca depende da convergência da linguagem dos dois agentes. Sob essa ótica, é oportuna a construção de uma listagem de palavras representativas, equilibrando a recorrência e a relevância, inclusive para futuras indexações, posto que tal listagem poderia aparecer como sugestão para os futuros autores (NARUKAWA; GIL LEIVA; FUJITA, 2009) e colaborar para a padronização da indexação de materiais.

Sequencialmente, são descritos os resultados associados à listagem de palavras-chave atribuídas pelos autores aos documentos indexados ao RI Unesp.

6.1.2 Palavras-chave empregadas pelos autores

Para criação da lista composta pelas palavras empregadas pelos autores, foi implementado, na ferramenta RILogUser, um *web crawler* que visitou 32 páginas do RI Unesp e coletou 325.524 palavras-chave nos documentos indexados até 24 de março de 2022. Por meio da coleta de dados, verificou-se que a lista de palavras-chave de autores do RI Unesp não possuía tratamento de linguagem natural, inclusive com relação a limites de expressões. Exemplo disso, foram os registros encontrados como representantes de palavras-chave, sendo o menor registro composto por um caractere e o maior por 411 palavras, indicando possível recuperação de resumo completo.

A falta de homogeneização, característica da linguagem natural, na indexação pode impactar e inviabilizar a recuperação destes materiais pelos usuários, fazendo-se necessário ponderarmos sobre esse fator para propor melhorias, a partir de estratégias computacionais as quais aumentem compatibilização entre a forma de indexar e a forma de buscar informações.

De acordo com Fujita e Tartarotti (2020, p. 335), a responsabilidade do autor ao atribuir palavras-chave ao material que produz é a de um indexador, pois, como o indexador, o autor “precisa pensar em outras funções além da representação do conteúdo do texto e adotar padrões que sejam compatíveis com diferentes empregos futuros”.

Em estudo publicado em 2020, Fujita e Tartarotti analisaram as palavras-chave da produção científica de pesquisadores em uma plataforma digital de currículos acadêmicos e verificaram a ausência de padronização dessas palavras-chave, fato que, como no estudo conduzido nesta Tese, resultou uma baixa convergência entre as palavras e evidenciou a

necessidade de oferecer aos autores algum meio de controle de vocabulário que favoreça a consistência da linguagem empregada.

Inicialmente, para reduzir as ocorrências de ruídos, já que os resultados experimentais nos levam à conclusão de que mínimas variações podem gerar ruídos na busca de materiais desejados e impedir sua recuperação pelos usuários, propusemos o tratamento das palavras-chave atribuídas aos documentos por seus autores, seguindo a retirada, por meio de expressão regular, de sinais gráficos, beneficiando 4.182 registros.

Sequencialmente, das 325.524 palavras-chave de autores, 47.322 estavam duplicadas. Entendemos como duplicadas aquelas que tiveram apenas variações de letras maiúsculas e minúsculas, isto é, nesta etapa não foram consideradas variações ortográficas nem flexões de gênero ou número. Dessa forma, a listagem inicial de 325.524, após eliminação das duplicações, foi reduzida para 278.202. A Figura 38 exemplifica o processo de tratamento das palavras-chave dos autores:

Figura 38. Exemplos de tratamento das duplicações de palavras-chave dos autores

The screenshot shows the REPOSITÓRIO INSTITUCIONAL UNESP interface. The search bar contains 'PSICOLOGIA HISTORICO-CULTURAL'. The results show a list of keywords with counts in brackets, such as 'Psicologia historico-cultural [1]', 'Psicologia histórico cultural [3]', and 'Psicologia Histórico-Cultural [22]'. Annotations include a red box labeled 'Documentos relacionados' pointing to the list, a green box labeled 'Módulo crawler: RLogUser' pointing to the search bar, and a blue box labeled 'Lista limpa: PSICOLOGIA HISTORICO-CULTURAL[79]' pointing to the unified result.

Fonte: RI Unesp (2022), editado pelo autor.

Destaca-se, na Figura 38, a unificação de palavras-chave dos autores com variações no uso de letras minúsculas e maiúsculas e de acentuação que geraram 9 variações para a expressão “PSICOLOGIA HISTORICO-CULTURAL” a qual na lista limpa foi unificada com a adoção do padrão caixa alta e sem acentuação, totalizando 79 documentos representados. Esse processo é realizado em todas as palavras-chave do RI Unesp a fim de colaborar para a concisão

da listagem. A Figura 39 exibe o código responsável pelo tratamento das duplicações das palavras:

Figura 39. Fragmento do código para eliminar duplicações

```

18  html = ""
19  # URL à ser consultada adicionando o "npp" e "offset" atuais
20  url = 'https://repositorio.unesp.br/browse?npp='+str(npp)+'&sort_by=-1&type=subject&offset='+str(offset)+'&etal=-1'
21  # Realiza a requisição HTTP e armazena o retorno em "req"
22  req = requests.get(url)
23  # Captura o texto da página (HTML) para a variável "html"
24  html = req.text
25  # Realiza o parser do "html" retornando um objeto navegável em "soup"
26  soup = BeautifulSoup(html, 'html.parser')
27  # Localiza a quantidade máxima de keywords, descrita em um item do HTML recebido, e armazena em "num_max"
28  num_max = soup.find('p', attrs={"class": "pagination-info"}).text.split('de')[-1]
29
30  # Localiza as palavras chaves na página HTML com base nas classes utilizadas para formatação e processa uma a uma
31  for kw in soup.find_all('td', attrs={"class": "ds-table-cell odd"}):
32      # Verifica se a palavra sendo processada não está no catálogo (variável keywords)
33      kw_clean = unicode(kw.a.text.strip(' ').strip(' ').upper())
34      if kw_clean not in keywords.keys():
35          # Caso verdadeiro adiciona a palavra com um contador "1"
36          keywords[kw_clean] = {'COUNT': 1, 'NUM_DOCS': int(re.search(r'\d+', kw.text).group(0)[1:-1])}
37      else:
38          # Caso contrário soma 1 ao contador relacionado à palavra
39          keywords[kw_clean]['COUNT'] += 1
40          keywords[kw_clean]['NUM_DOCS'] += int(re.search(r'\d+', kw.text).group(0)[1:-1])
41  # Incrementa a variável "offset" com o valor de "npp"
42  offset += npp

```

Fonte: elaborada pelo autor.

Entre as linhas de códigos expostos na Figura 39, destacamos que, para identificar palavras duplicadas, a programação inicialmente converte todas as letras dos registros em maiúsculas, linha 33 da Figura 39. Sequencialmente, da linha 34 à 40, o código agrupa palavras iguais, somando a quantidade de documentos que cada registro representa.

Feita essa etapa de limpeza básica das palavras-chave, foi utilizada a API disponível em: <https://detectlanguage.com/> para identificação de idiomas, sendo observada a maior ocorrência de três: inglês, com 172.057 termos (61,8%); português, com 48.732 (17,5%) e espanhol, com 18.520 termos (6,7%). Outros idiomas somados resultaram 38.893 termos restantes (13,9%). A identificação de idiomas é um recurso relevante para análise de base de dados multilíngues.

Ao pensarmos em indicadores de relevância dos termos, a frequência de utilização pelos autores foi quantificada, conforme Tabela 3:

Tabela 3. Coocorrência de palavras-chave dos autores

Lista de palavras-chave de autores em documentos armazenados no RI Unesp	
Período analisado	Até 24-03-2022
Total de palavras-chave não duplicadas	278.202
Total de palavras-chave que recuperaram apenas 1 Documento	183.986
Total de coocorrências* de palavras-chave	94.216
Extensão média das palavras-chave coocorrentes	1,98

*São consideradas coocorrentes as palavras-chave que representam dois ou mais documentos.

Fonte: elaborada pelo autor.

Os itens descritos na Tabela 3 foram obtidos pela execução do código exibido na Figura 40, no qual a média das palavras-chave dos autores foi armazenada na variável 'AVERAGE'(linha 126), após percorrer todo o RI Unesp resultou na extensão média de 1,98 palavras (linha 126 à 144). A quantidade de palavras empregadas uma única vez pelos autores é armazenada da variável 'UNSELECTED_TERMS', enquanto as coocorrências são armazenadas na variável 'SELECTED_TERMS'.

Figura 40. Fragmento do código para identificação de coocorrência de palavras-chave dos autores

```

125 result = {'ALL_TERMS': keyword_data, 'SELECTED_TERMS': {}, 'UNSELECTED_TERMS': {}, 'SELECTED_TERMS_INFO':
126         {'AVERAGE': 0, 'COUNT': 0, 'COUNT_WORDS': 0}}
127
128 for k, v in keyword_data.items():
129     if(min_count == 0 or v['COUNT'] >= min_count):
130         is_valid = _is_valid(k)
131         if is_valid:
132             result['SELECTED_TERMS'][k] = v
133             result['SELECTED_TERMS_INFO']['COUNT'] += 1
134             result['SELECTED_TERMS_INFO']['COUNT_WORDS'] += k.count(' ') + 1
135         else:
136             if is_valid['rule'] in result['UNSELECTED_TERMS']:
137                 result['UNSELECTED_TERMS'][is_valid['rule']] = {}
138                 result['UNSELECTED_TERMS'][is_valid['rule']][k] = v
139             else:
140                 key = 'MIN_COUNT:' + str(min_count)
141                 if key not in result['UNSELECTED_TERMS']:
142                     result['UNSELECTED_TERMS'][key] = {}
143                     result['UNSELECTED_TERMS'][key][k] = v
144 result['SELECTED_TERMS_INFO']['AVERAGE'] = result['SELECTED_TERMS_INFO']['COUNT_WORDS'] / result['SELECTED_TERMS_INFO']['COUNT']
145 return result

```

Fonte: elaborada pelo autor.

Do número total de 278.202 palavras-chave não duplicadas, disponíveis no RI Unesp, 183.986 (66,1%) foram utilizadas apenas uma vez, ou seja, foi considerada como representativa apenas de um documento e 94.216 (33,9%) foram utilizadas para representação de mais de um documento, o que identificamos como coocorrência.

Por entendermos a coocorrência como indicativo de maior representatividade, o *corpus* analisado é composto apenas por palavras-chave coocorrentes. Adicionalmente, de

maneira amostral, considerando a frequência de utilização das palavras-chave, a listagem composta pelas 947 palavras-chave que recuperaram mais de 100 documentos. O fragmento da Tabela 4 exibe as onze palavras do topo da listagem, as quais recuperaram mais de 1000 documentos.

Tabela 4. Palavras-chave dos autores que recuperaram mais de 1000 documentos

	A	B	C
1	Palavra-chave	Quantidade	Documentos
2	MALE	2	2536
3	BRAZIL	3	2405
4	FEMALE	2	2089
5	HUMAN	2	2043
6	EDUCACAO	3	1632
7	NONHUMAN	2	1479
8	CONTROLLED STUDY	2	1302
9	ADULT	2	1231
10	EDUCATION	5	1186
11	HUMANS	2	1126
12	ANIMALS	3	1020

Fonte: elaborada pelo autor.

Na Tabela 4, a primeira coluna exibe a palavra-chave em formato padronizado (todas as letras maiúsculas e sem acentuação) a quantidade de variações que foram unificadas e o número de documentos recuperados. Como complemento, a listagem com as 947 palavras-chave que recuperaram mais de 100 documentos também está disponível em: <http://riloguser.uh-app.com.br/Datasciencer-doutorado/datasciencer.do?action=disponibilizaAll>, página integrante do módulo *on-line* ferramenta RILogUser em seu módulo *on-line*.

De forma análoga, com o objetivo de verificar quais palavras-chave são privilegiadas e quais redes científicas podem ser identificadas relacionando-as aos autores de artigos selecionados, em 2013, quase dez anos antes da realização da investigação que fundamenta esta Tese, foi conduzida, por Miguéis *et al.* (2013), pesquisa com resultados semelhantes. Ao analisar as palavras-chave utilizadas pelos autores da Universidade de Coimbra, nos artigos publicados na área das Ciências Farmacêuticas, comparando-as com os termos da linguagem documental *Medical Subject Headings* (MeSH), foi observada baixa convergência entre os termos, fato que indicou a necessidade de refinamento das palavras-chave empregadas pelos autores para melhor representação documental durante o processo de indexação e, consequentemente, no processo de recuperação. Os autores indicaram ainda a necessidade de tratamento prévio dos dados coletados, uma vez que a incorporação de um conjunto de símbolos

que não representem as palavras-chave dos autores pode inviabilizar a recuperação de materiais desejados.

Em 2014, Lucas e Lara (2014) também investigaram as palavras-chave como elo entre o artigo e o autor. A amostra selecionada pelas autoras foi composta por 354 artigos, produzidos por 462 autores, sendo os materiais representados por 927 palavras-chave diferentes, das quais 648 (69,9%) foram utilizadas apenas uma vez, sendo a coocorrência característica apenas de 279 (30,1%) palavras-chave. Portanto, os resultados do estudo de Lucas e Lara (2014) possuem semelhanças com os resultados obtidos na análise desta etapa da Tese. Ainda que as pesquisas trabalhem com amostras de tamanhos bem distintos, os resultados analisados obtidos são semelhantes quando analisados proporcionalmente.

Diante das análises feitas no RI Unesp, comprovamos que embora o repositório ofereça ao autor a possibilidade de inserir palavras-chave do Tesouro Unesp, quando este vai indexar documentos, é possível também ignorar as sugestões contidas no Tesouro e inserir palavras-chave sem qualquer tratamento. Sem a utilização de vocabulário controlado, a consistência e o controle de qualidade dos registros ficam prejudicados pela falta de homogeneização da linguagem, repercutindo negativamente na recuperação de documentos e impedindo que ciclos informacionais se mantenham ativos.

Nesse sentido, possíveis soluções devem ser buscadas a fim de aumentar a precisão e a consistência dos metadados, nomeadamente, das palavras-chave atribuídas aos documentos pelos autores, de modo que o usuário possa acessar o mesmo vocabulário dos autores e aumentar a probabilidade de sucesso das buscas. Isso porque, se os autores continuarem usando vocabulário livre para carregar seus documentos, seguirão persistindo os mesmos problemas na hora de recuperar a informação por parte do usuário. Portanto, a disponibilidade de uma lista de termos a serem usados como palavras-chave é recurso fundamental para ampliar a compatibilização entre vocabulários, acrescentando-se a possibilidade de que, se o autor não encontrar um termo determinado que julga essencial para representar seu documento, possa acrescentá-lo, mas que essa palavra seja carregada como um metadado diferente. Dessa maneira, será fácil identificar quais são as palavras-chave normalizadas pelos gerentes do RI Unesp e quais as palavras-chave devem ser estudadas para sua normalização. Confirmando, assim, a importância da adoção de critérios para tratamento de palavras-chave, de modo a identificar e mesmo sugerir padrões que unifiquem materiais afins, para que documentos de interesse não sejam omitidos aos usuários.

6.1.3 Lista de termos do Tesauro Unesp

Em posse das listagens oriundas do emprego de linguagem natural - dos usuários e dos autores - buscou-se um vocabulário controlado como parâmetro comparativo, a fim de observar compatibilidades e propor a atualização desse vocabulário controlado para utilização no RI Unesp, já que tanto os usuários quanto os autores dispunham apenas de sua linguagem natural para realizar pesquisas e escolher palavras-chave para suas publicações, respectivamente. Destaca-se que, quando o autor realiza o autoarquivamento, há a possibilidade (porém não a obrigatoriedade) de selecionar termos listados no Tesauro Unesp. Nesse sentido, foi observado que a falta de homogeneização da linguagem acarreta perda informacional associada a não recuperação ou recuperação parcial de materiais.

Assim, o Tesauro Unesp foi escolhido como vocabulário controlado de referência por possuir termos especializados organizados por áreas de conhecimento e por ser utilizado na rede de bibliotecas da Unesp, inclusive, a atualização periódica dos termos indexados no Tesauro Unesp faz-se necessária, diante de sua importância na recuperação fidedigna de materiais (FUJITA; MOREIRA, 2021). Ademais, trata-se de referência com abrangência substancial por possuir termos especializados nas áreas de conhecimento do RI Unesp.

Para inclusão do Tesauro Unesp na ferramenta RILogUser, de modo a viabilizar o prosseguimento da análise comparativa, a etapa de coleta foi executada por meio de uma API disponível no site do Tesauro Unesp, com extração realizada em 24 de março de 2022, totalizando 49.237 formas alternativas não autorizadas e 40.351 autorizadas, conforme Tabela 5:

Tabela 5. Listagem dos termos do Tesauro Unesp

Tesauro Unesp	
Data da coleta dos dados	24-03-2022
Número de termos não autorizados	49.237
Número de termos autorizados	40.351
Número total de termos	89.588
Menor termo listado	1 palavra
Maior termo listado	17 palavras
Extensão média dos termos	1,80

Fonte: elaborada pelo autor.

Os quantitativos exibidos na Tabela 5 são obtidos pela execução do código exibido na Figura 41. Coletando os termos autorizados e não autorizados (linhas 54 e 55), bem como todo o processamento armazenado na linha 64. Destaca-se que há tratamento de conexão a cada 1 segundo, em toda a coleta de dados do Tesouro Unesp, fator responsável pela estabilidade do código para coletar a totalidade dos termos do Tesouro.

Figura 41. Fragmento do código para acesso à API do Tesouro Unesp e coleta dos termos

```

46 while True:
47     try:
48         for k, v in result['terms'].items():
49             if k not in result['relationships']:
50                 url = 'https://portal.biblioteca.unesp.br/tematres/vocab/services.php?task=fetchAlt&arg=' + k
51                 r = requests.get(url)
52                 result['relationships'][k] = []
53                 for child in ET.fromstring(r.content).iter('term'):
54                     result['alternative_terms'][child.find('term_id').text] = prepare_string(child.find('string').text)
55                     result['relationships'][k].append(child.find('term_id').text)
56             break
57     except:
58         print(" - Erro de Conexão")
59         time.sleep(1)
60         print(" - Tentando Novamente...")
61
62 result['info']['average'] = result['info']['count_words'] / result['info']['count']
63
64 return result

```

Fonte: elaborada pelo autor.

Com a lista dos logs de busca dos usuários, a lista de palavras-chaves utilizadas pelos autores para representação de suas publicações e a lista de termos do Tesouro Unesp, demos sequência à análise comparativa dos dados constantes com objetivo de observar tanto as palavras comuns às listagens como também, e principalmente, as palavras exclusivas. Assim, a análise de logs de busca de usuários do RI Unesp como fonte informacional para propor atualização do vocabulário controlado Tesouro Unesp foi implementada a partir da comparação dos termos presentes nas três listas já limpas e ordenadas alfabeticamente.

Como ponto inicial para apresentação da síntese dos resultados obtidos, consideramos os quantitativos apresentados na Tabela 6:

Tabela 6. Quantitativos dos vocabulários das listagens

Listas	Total de palavras/termos
Palavras de buscas dos usuários	7.362
Palavras-chave dos autores	94.216
Termos Tesouro	89.588

Fonte: elaborada pelo autor.

Cada lista identificada na Tabela 6 é, portanto, composta por um conjunto de elementos (palavras de buscas dos usuários, palavras-chave dos autores e termos do Tesouro Unesp) identificados computacionalmente pela execução dos códigos expostos na Figura 42:

Figura 42. Código para identificação dos vocabulários

```

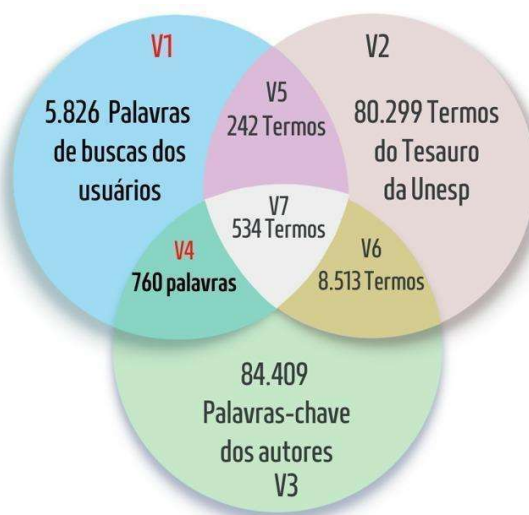
127 results = compare_terms_if_needed(
128     terms={
129         'repositorio' : list(repositorio_terms['SELECTED_TERMS'].keys()),
130         'log' : list(log_terms['SELECTED_TERMS'].keys()),
131         'thesauro' : _list_thesauro
132     },
133     json_file_output=_results_file_output,
134     force_compare=_results_force_compare)

```

Fonte: elaborada pelo autor.

Os produtos das comparações, as quais identificam termos com grafia idêntica, seguindo o método proposto na seção 5, item 5.3 “Análise de dados”, foram compostos pelos quantitativos apresentados na Figura 43, utilizando a ferramenta RILogUser com processamento *off-line* das listagens:

Figura 43. Diagrama de produtos das comparações e identificação de conjuntos comuns



Fonte: elaborada pelo autor.

A comparação entre as listagens é uma forma de verificar quão convergentes têm sido ou não os vocabulários. As palavras candidatas para a proposta de atualização do vocabulário controlado do Tesouro Unesp originam-se do V1 e V4, composto por palavras exclusivas

extraídas dos logs de busca dos usuários ou convergentes com as palavras-chave dos autores, respectivamente.

Pelo observado na coleta, preparação e análise comparativa das três listas de vocabulários, o fato do conjunto V7 ser composto por apenas 534 termos comprova a relevância da pesquisa no que concerne à busca por métodos que permitam ampliar a convergência de vocabulário, bem como acompanhar as estratégias de busca dos usuários para que seja possível aprimorar suas experiências de busca. Assim, ao aumentar a compatibilização de linguagens entre os diversos agentes envolvidos, é possível garantir o acesso e a usabilidade das informações e conhecimentos produzidos e armazenados nas instituições de ensino e pesquisa.

É importante destacarmos que há uma diferença significativa entre os tamanhos dos conjuntos (listagens), posto que as palavras de busca capturadas pelos logs têm a delimitação temporal de 6 meses de coleta, enquanto os demais conjuntos são compostos por todos os registros desde a fundação do RI e do Tesouro Unesp.

Contudo, mesmo considerando as variáveis, V4 como subconjunto composto pela intersecção dos logs de busca e palavras-chaves atribuídas pelos autores apresenta 760 palavras, quantidade relativamente pequena vista a quantidade de elementos dos conjuntos. O conjunto intersecção ser composto por poucas palavras evidencia a pouca convergência de vocabulário entre o produtor e o potencial receptor dos documentos. Esse dado, por sua vez, é indicativo de falhas no processo de representação que podem inviabilizar a recuperação do material e interromper o ciclo informacional.

Assim, todas as listagens completas estão disponíveis para acesso pela página da RILogUser: <http://riloguser.uh-app.com.br/Datasciencer-doutorado/>. Destacamos que os conjuntos V1 e V4, conforme identificado na Figura 43, são a base do vocabulário oriundo das pesquisas realizadas nesta Tese, os quais serão fornecidos à equipe gestora do Tesouro Unesp como listas de palavras candidatas à atualização.

Assim, a análise das listagens dos termos exibidos e das intersecções permite visualizar termos utilizados tanto pelos usuários do RI quanto pelos autores dos materiais disponibilizados. Essa verificação comprovou a relevância do experimento para a atualização de um vocabulário controlado que otimize a recuperação de informações pelos usuários no RI.

A partir dos resultados obtidos, algumas ações são indicadas para que haja uma melhor representação e recuperação de documentos por assuntos no RI Unesp, inclusive, com potencial de replicação em outros repositórios institucionais que têm sua função de divulgação de informações prejudicada pela falta de convergência entre os vocabulários empregados pelos diferentes agentes envolvidos na produção, armazenagem e recuperação de materiais.

Em síntese, para iniciar uma intervenção planejada e sistemática, é imperativo o fortalecimento do Grupo Gestor do RI Unesp. Com esse propósito, o Grupo “Representação Temática da Informação”, composto por professores da área da Ciências da Informação e Ciências Sociais Aplicadas, em parceria com a Coordenadoria Geral de Bibliotecas (CGB) têm muito a contribuir. Ademais, o Grupo Gestor (Grupo Gestor da Política do Repositório Institucional Unesp – GRI-Unesp) pode agregar docentes especialistas e bibliotecários catalogadores-indexadores representantes das seis áreas do conhecimento: Área de Artes e Humanidades, Área de Biomédicas, Área de Ciências Aplicadas, Área de Conhecimentos Gerais, Área de Exatas e Área de Tecnológicas para composição de diretrizes e planos de trabalho e ampliação das parcerias com o Grupo de Trabalho em Política de Indexação da Rede de Bibliotecas Unesp.

Em adição, a elaboração/formalização de uma política de indexação específica para atender as particularidades do RI Unesp é fundamental, consolidada pela divulgação de um Manual de Política de Indexação complementar aos já existentes tutoriais para autoarquivamento, disponíveis no site do RI Unesp. O propósito do Manual de Política de Indexação é orientar sistematicamente a indexação de assuntos no RI Unesp pelos profissionais que realizam o tratamento temático dos recursos informacionais e, especialmente, para os autores que armazenam suas produções no RI.

Em complemento, a elaboração de um vocabulário próprio e único no âmbito da Unesp, o qual seja proposto a partir de investigações feitas no próprio RI, como a que realizamos nesta Tese, é viável e relevante. Como nesta investigação realizamos o levantamento dos termos já atribuídos aos documentos no RI, obtivemos uma base confiável dos termos considerados representativos pelos autores (uma vez que foram escolhidos por eles para representarem suas produções) e que, por isso, podem subsidiar a compilação dos termos para formalização do Vocabulário Controlado da Unesp, o qual pode ser utilizado para atualização do Tesouro Unesp.

Como desdobramento importante para o sucesso da representação e posterior recuperação dos materiais armazenados, são essenciais estratégias que ampliem a utilização dos usuários desses vocabulários controlados, de modo que a linguagem empregada seja compartilhada pelo maior número de pessoas.

Isso posto, a padronização dos registros (palavras-chaves dos autores) no RI Unesp quanto à representação descritiva, tende a facilitar também a recuperação por parte dos usuários, visto que duplicações e inconsistências serão minimizadas já no momento da indexação.

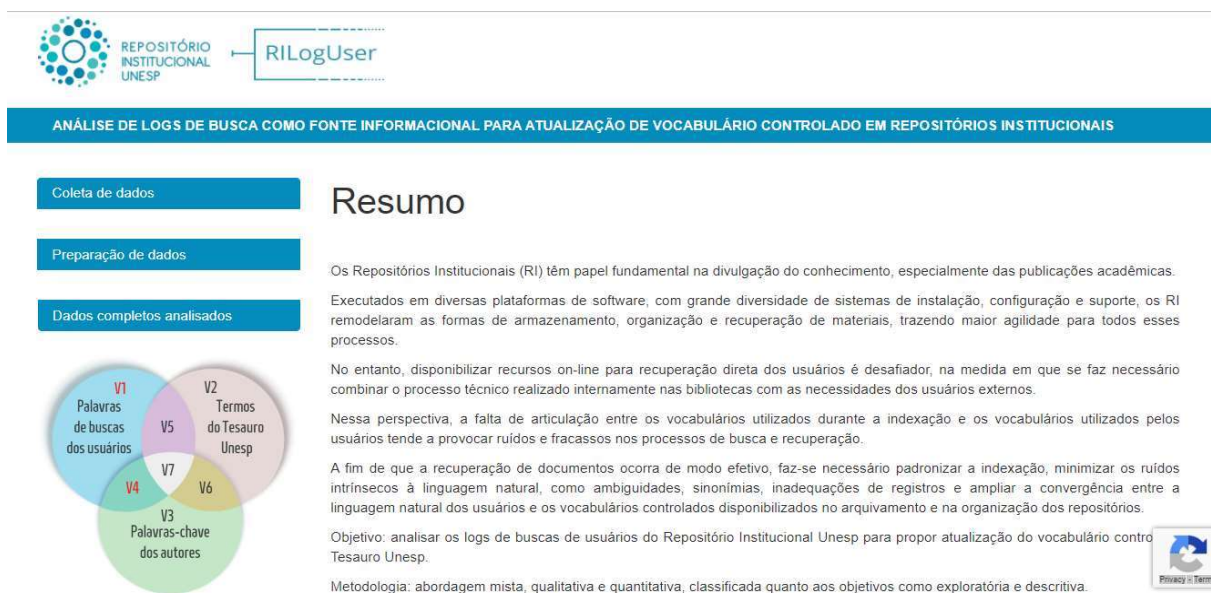
6.2 RILogUser: módulo *on-line*

O uso de programas instalados vem sendo integrado pela utilização de plataformas *on-line*, pois, além da economia de tempo, a praticidade e usabilidade têm se destacado, uma vez que o acesso em rede rompe limitações físicas ou restritivas com relação à máquina do usuário. Ademais, o uso *on-line* favorece o desenvolvimento colaborativo com a possibilidade de edições simultâneas e controladas, como na plataforma Github de hospedagem de código-fonte e arquivos com controle de versão.

Nessa perspectiva, como protótipo, apresentamos a ferramenta RILogUser Web, como forma de demonstrar, em uma interface mais amigável, o desenvolvimento de análise de logs e outras listas de vocabulário.

O desenvolvimento deste projeto foi planejado basicamente em 3 etapas, seguindo a proposta de Jansen (2008), sendo elas: coleta, preparação e análise de dados. Inicialmente, ao acessar o site: <http://riloguser.uh-app.com.br/Datasciencer-doutorado/>, o usuário da ferramenta tem acesso a um resumo da pesquisa que motivou a criação da RILogUser, de forma a contextualizá-lo. Também na tela inicial (Figura 44), no canto superior esquerdo, são expostas as funcionalidades básicas nas quais o usuário poderá carregar dados, prepará-los e obter análises comparativas.

Figura 44. Tela inicial da ferramenta RILogUser



Fonte: elaborada pelo autor.

Das três partes constantes do menu, as duas primeiras (coleta e preparação) realizam ações programadas em um *Servlet* Java (classe principal, com 2.370 linhas) e o último é apenas

um repositório para *download* de arquivos já processados. Por limitações técnicas do servidor UOL Host que hospeda o serviço Java (possui apenas 512 MB de memória principal), não foi possível colocar na versão *on-line* toda a capacidade de processamento da versão *off-line*, que gerou os arquivos disponíveis no terceiro item do menu (dados completos). Mesmo assim, na versão *on-line*, é possível realizar testes que comprovam a tese de doutorado. Todo o código-fonte está disponível em: <https://github.com/datasciencer2022/Doutorado/>.

As telas são páginas do tipo *Java Server Pages* (JSP) e serão detalhadas a seguir. Quando necessário, será dada a explicação do(s) método(s) chamados na classe principal.

Ao clicar na primeira opção do lado esquerdo da tela inicial (Coleta de dados), iniciamos o trabalho de investigação de vocabulário a partir de diversas listagens de palavras que compõem o *corpus* de pesquisa. Após fazer o *upload* de listagens, o usuário pode obter análises comparativas parciais. A Figura 45 expõe a página em que é possível testar a ferramenta com listagens amostrais de até 300 registros, identificando os diferentes conjuntos obtidos por intersecções:

Figura 45. Tela secundária para carregamento e análise de listas de vocabulários

REPOSITÓRIO INSTITUCIONAL UNESP RILogUser

ANÁLISE DE LOGS DE BUSCA COMO FONTE INFORMACIONAL PARA ATUALIZAÇÃO DE VOCABULÁRIO CONTROLADO EM REPOSITÓRIOS INSTITUCIONAIS

Preparação de dados

Dados completos analisados

Voltar

Os produtos das comparações entre as listas viabilizam a identificação de conjuntos compostos por elementos comuns, nomeados no diagrama: V4, V5, V6 e V7.

V1 Palavras de buscas dos usuários

V2 Termos do Tesouro Unesp

V3 Palavras-chave dos autores

V4, V5, V6, V7

Preparação de dados

Coleta de logs de buscas: Escolher Arquivo Nenhum arquivo escolhido

Coleta de palavras-chave de autores: Escolher Arquivo Nenhum arquivo escolhido

Coleta de termos do Tesouro: Escolher Arquivo Nenhum arquivo escolhido

Vocabulários criados:

- ☐ Palavras exclusivas dos logs de busca (V1)
- ☐ Termos exclusivos do Tesouro (V2)
- ☐ Palavras exclusivas das palavras-chave (V3)
- ☐ Palavras comuns: logs de busca e palavras-chave (V4)
- ☐ Termos comuns: logs de busca e termos Tesouro (V5)
- ☐ Termos comuns: palavras-chave e termos Tesouro (V6)
- ☒ Intersecção de todos (V7)

Limite máximo: 300 linhas

Análise de dados

Fonte: elaborada pelo autor.

A Figura 45 exhibe a tela na qual é possível fazer o carregamento de listas em *xlsx* com dados coletados durante a pesquisa. Destacamos que foram coletadas três listagens:

palavras de busca de usuários, via logs; palavras-chave dos autores, via *crawler*; e termos já constantes no Tesauro Unesp.

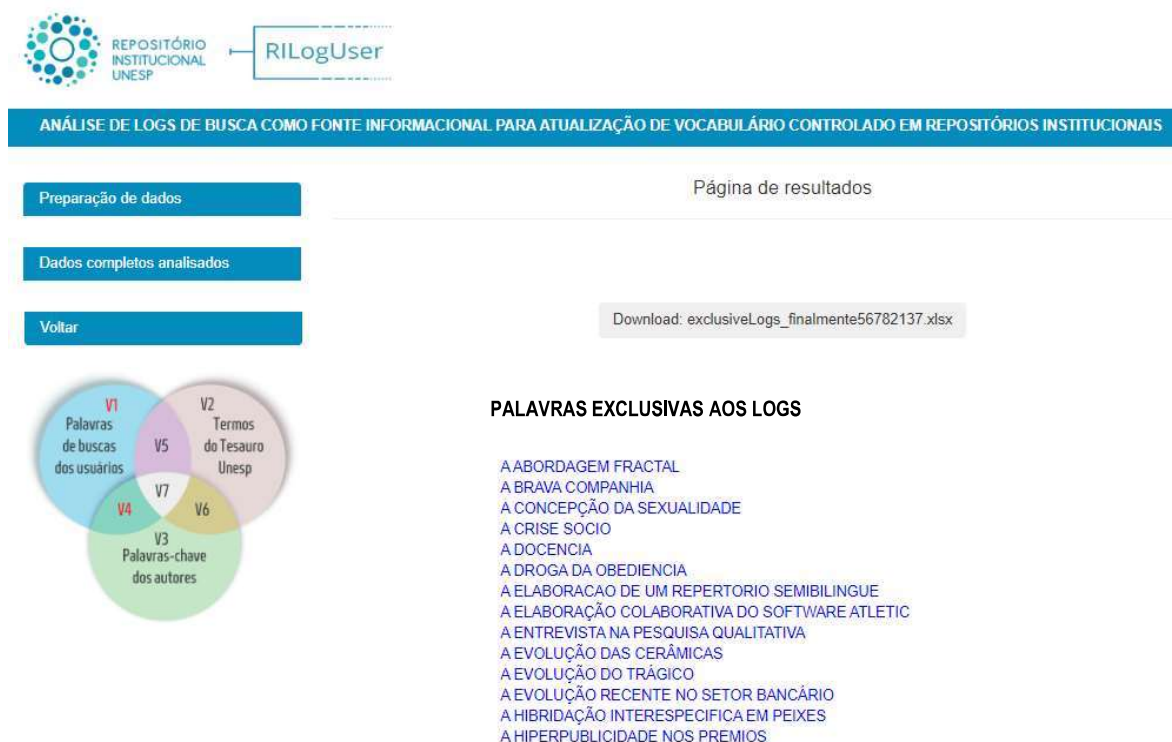
Ao fazer o *upload* dessas listagens e considerá-las como conjuntos de dados, por meio da RILogUser, podemos prosseguir a análise comparativa, selecionando a intersecção que for de interesse. Tais intersecções formam subconjuntos os quais nomeamos V4, V5, V6 e V7.

Observamos, que o usuário pode obter diversos resultados comparativos entre as listas carregadas, termos comuns aos logs, às palavras-chave e ao Tesauro (V7); aos logs e termos do Tesauro (V5); às palavras-chave e aos termos do Tesauro (V6) e aos logs e às palavras-chave (V4). Também relevantes são os subconjuntos de palavras ou termos exclusivos V1 (exclusivo aos logs), V2 (exclusivo ao Tesauro) e V3 (exclusivo às palavras-chave).

Assim, a apresentação dos resultados se torna mais flexível e a leitura dos arquivos pode ser feita de acordo com as necessidades do usuário ou da equipe que deseja analisar os vocabulários.

A título de exemplo, após selecionar o subconjunto de interesse e clicar em “Análise de dados”, o usuário tem acesso ao resultado solicitado tanto em tela quanto em versão xlsx para *download*, como visualizado na Figura 46:

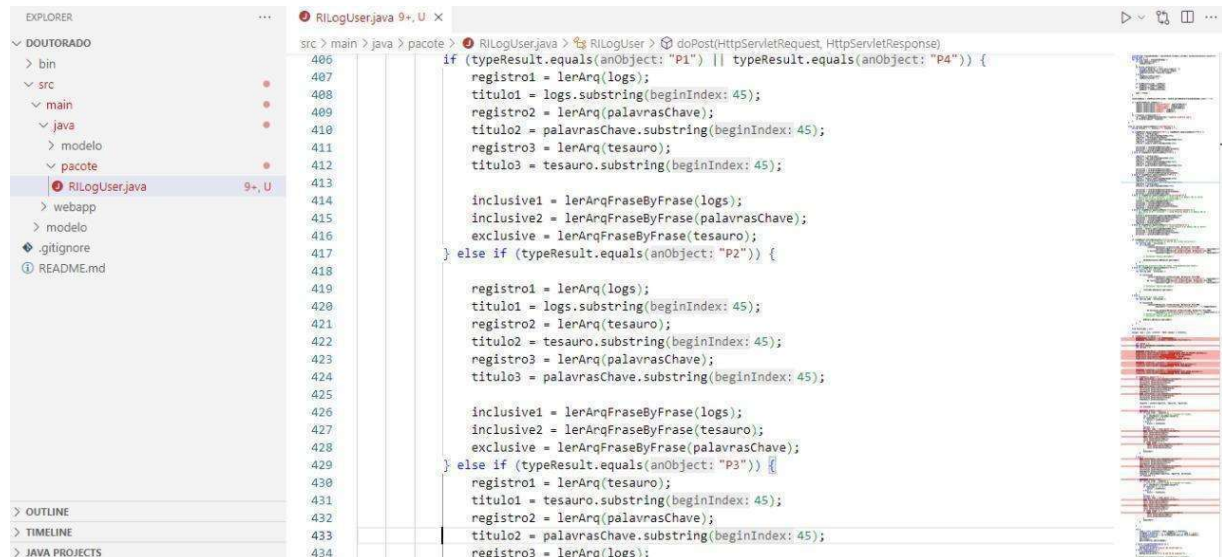
Figura 46. Página de resultados para palavras exclusivas aos logs



Fonte: elaborada pelo autor.

No código-fonte a seguir, é demonstrado o carregamento dos registros V1, V2 e V3, no fragmento, a ordem dos arquivos lidos pela função lerArq(...) muda, de acordo com a Figura 47 para gerar o resultado correto:

Figura 47. Código para carregamento de registros



Fonte: elaborada pelo autor.

Sequencialmente, a Figura 48 exibe o fragmento do código e as configurações de leitura dos conjuntos oriundos de intersecções ou de elementos (palavras ou termos) exclusivos. O termo “*inclusive*” destaca os itens integrantes das intersecções e o termo “*exclusive*” resulta os itens que estão presentes em apenas um conjunto:

Figura 48. Código para carregamento de registros de modo comparativo



Fonte: elaborada pelo autor.

Portanto, as frases e as palavras que não estão nas listas “*inclusive 1, 2 ou 3*”, são adicionadas as listas de *PalavrasExclusivasLogs* (V1), *TermosExclusivosTesouro* (V2) e *PalavrasExclusivasAutores* (V3).

Para que fossem desconsideradas diferenças de acentuação e letras maiúsculas e minúsculas, foi adicionado o código exibido na Figura 49 em várias partes do programa:

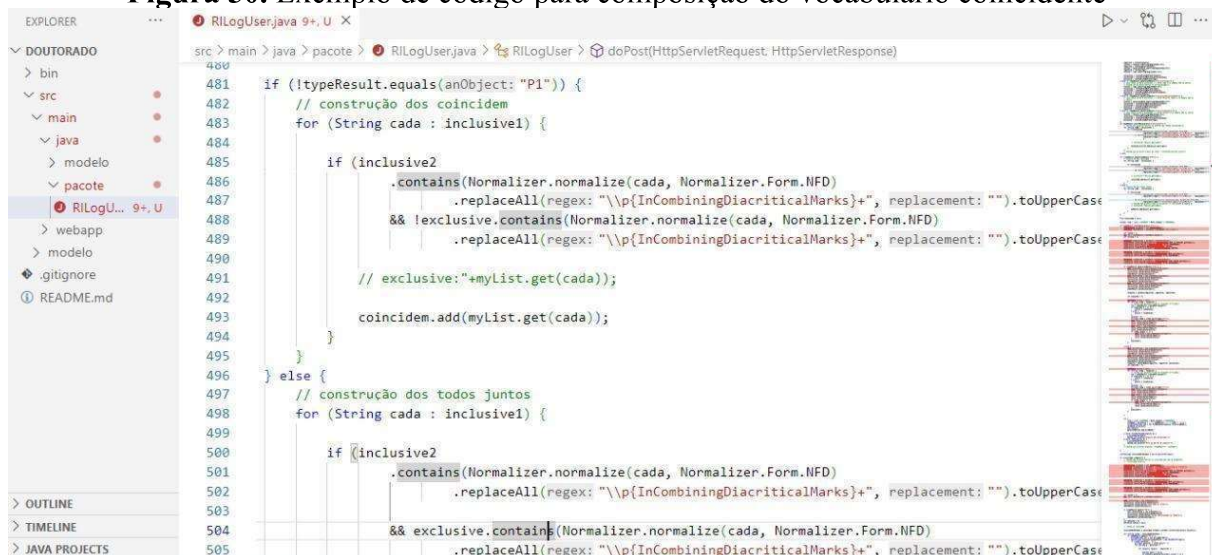
Figura 49. Código para tratamento de acentos e conversão das letras para maiúsculas



Fonte: elaborada pelo autor.

Já para o processamento dos vocabulários comuns, temos a seguinte instrução se (*if*) seguida de senão (*else*), como exemplo para delimitar V7, como representado na Figura 50:

Figura 50. Exemplo de código para composição do vocabulário coincidente



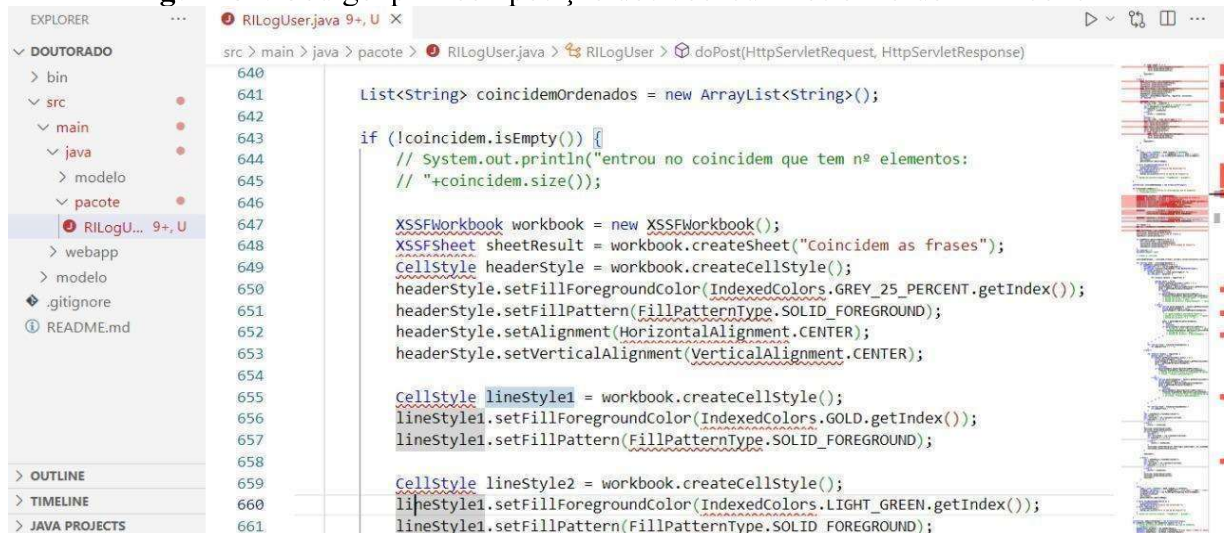
Fonte: elaborada pelo autor.

No caso de V7, são lidos todos os registros de “*inclusive1*” e caso estejam também em “*inclusive2*” e “*inclusive3*”, são adicionados à lista “*p1Result*”. E nos demais casos de pesquisas (V4, V5, V6), são lidos todos os registros de “*inclusive1*” e verificado se também estão em

“*inclusive2*” ou “*inclusive3*”, sendo então adicionados à lista “*coincidem*”. Este trecho de código é que dá sentido ao nome dessas listas, no restante do código elas foram reaproveitadas visando economia de memória.

Em seguida, é verificado se a lista “*coincidem*” foi preenchida e colocada em ordem alfabética, como colocado, na Figura 51:

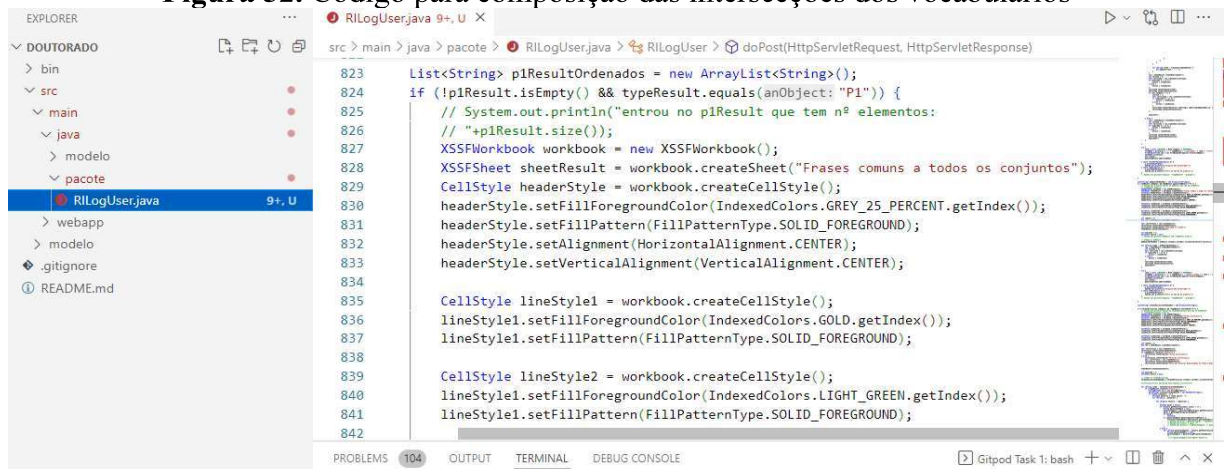
Figura 51. Código para composição dos vocabulários em ordem alfabética



Fonte: elaborada pelo autor.

E o mesmo ocorre para o caso de V7 (intersecção de V1, V2 e V3): palavras dos logs de pesquisas dos usuários, palavras-chave dos autores e termos do Tesauro, conforme exposto no trecho de código na Figura 52:

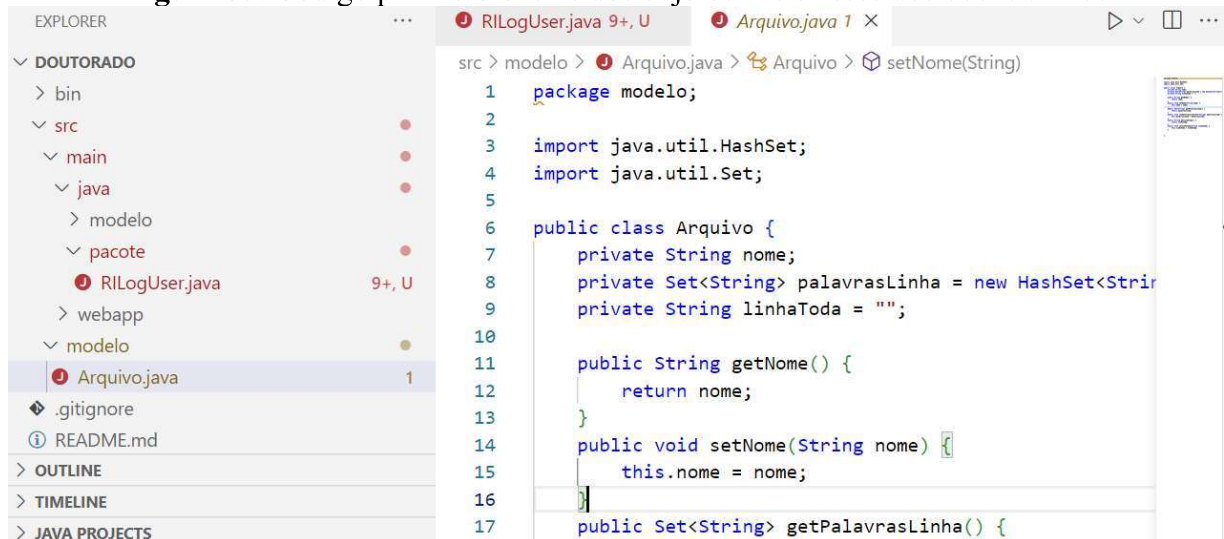
Figura 52. Código para composição das intersecções dos vocabulários



Fonte: elaborada pelo autor.

As variáveis de referência “registro1, 2 e 3” são objetos do tipo da classe definida como Arquivo.java, a qual armazena o nome do arquivo, a frase constante em cada linha, bem como uma lista de palavras constantes em cada linha, conforme exposto pelo código da Figura 53:

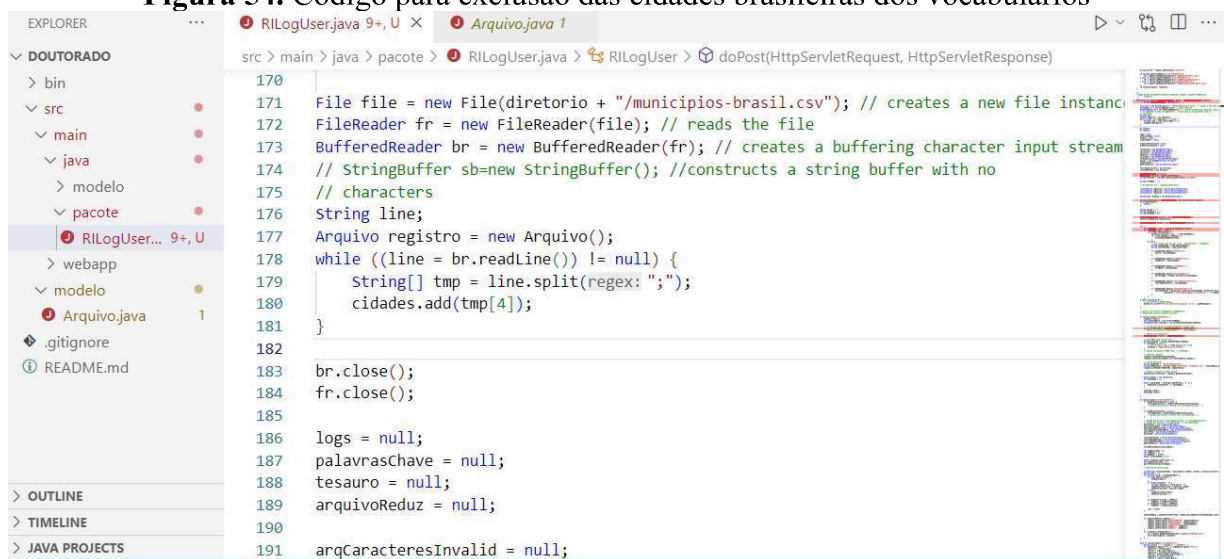
Figura 53. Código para referências dos objetos nas classes dos vocabulários



Fonte: elaborada pelo autor.

Na sequência da execução, é criada a lista “cidades”, com os nomes de todas as cidades do Brasil. Posteriormente, há uma verificação e, se qualquer cidade constar em uma das linhas dos demais arquivos, essa linha não será carregada, conforme exposto no trecho da Figura 54:

Figura 54. Código para exclusão das cidades brasileiras dos vocabulários



Fonte: elaborada pelo autor.

Finalmente, o nome do arquivo é adicionado ao atributo “*generatedFiles*” que será registrado na seção para ser lido pelo JSP de resultados, disponibilizando, assim, os arquivos para *download* da lista selecionada, conforme exposto pela Figura 55:

Figura 55. Tela de *download* do vocabulário selecionado



Fonte: elaborada pelo autor.

Outra funcionalidade disponibilizada no protótipo *on-line* é a preparação de dados, na qual os usuários podem estabelecer parâmetros para limpeza de suas listagens, conforme exibido na Figura 56:

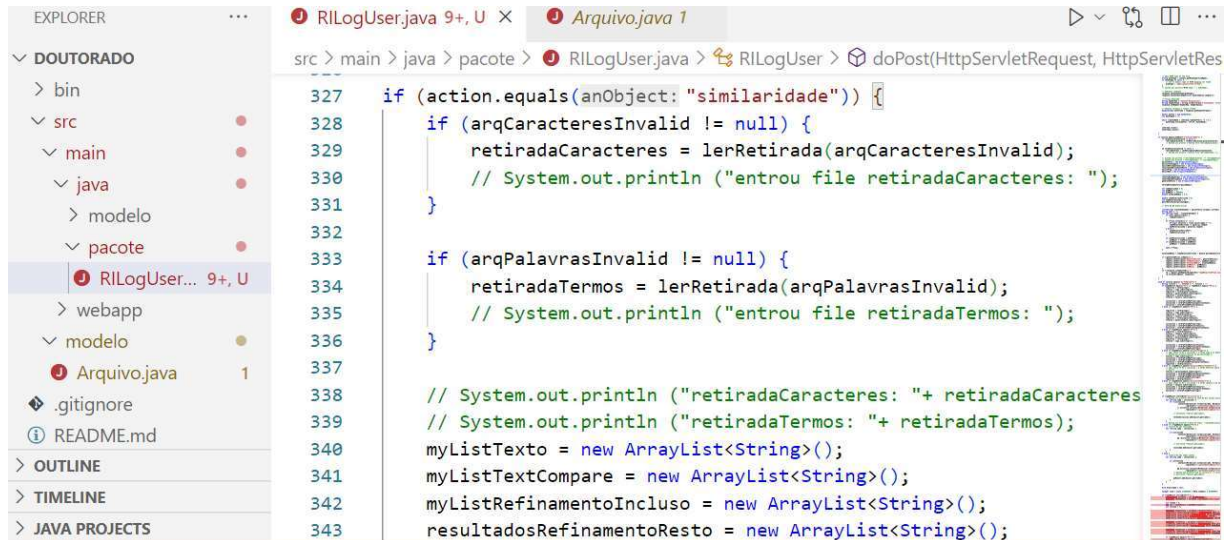
Figura 56. Tela para limpeza e depuração dos dados



Fonte: elaborada pelo autor.

Na Figura 56, primeiramente são carregados os arquivos ou digitadas as palavras ou expressões que contém os caracteres inválidos a serem retirados, juntamente com o arquivo propriamente dito que será trabalhado, conforme trecho da Figura 57:

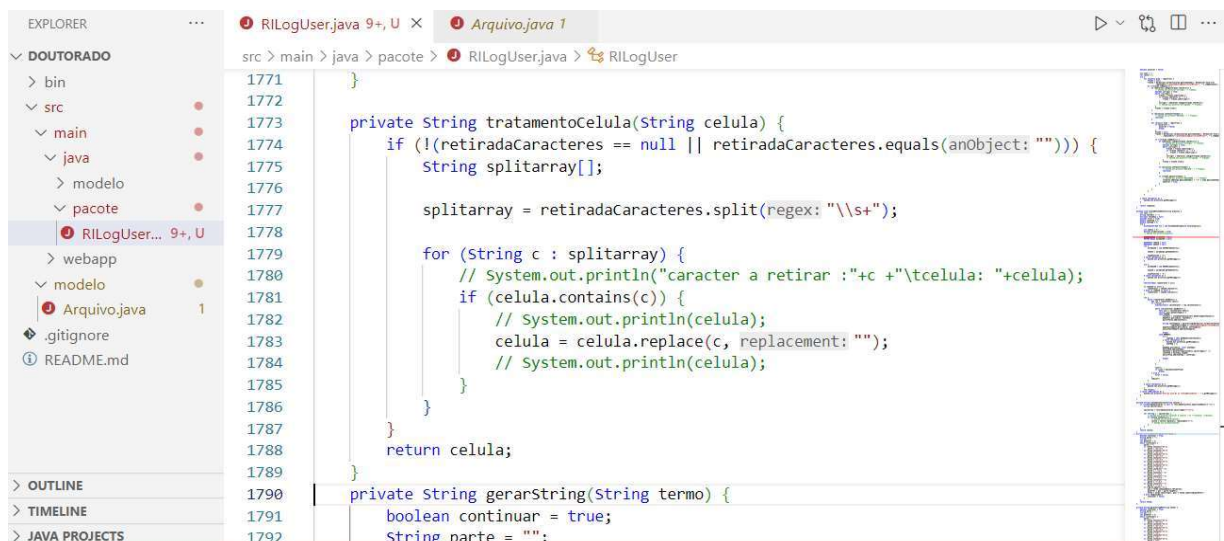
Figura 57. Códigos para limpeza e depuração dos dados



Fonte: elaborada pelo autor.

Isso é feito através dos métodos *lerRetirada(...)* e *lerArqRefinamento(...)*, respectivamente. Outra opção para entrar com a lista dos caracteres e palavras inválidas é pela digitação direta na página. Inclusive dentro do método *lerArqRefinamento* há a chamada para o método chamado *tratamentoCelula*, especificado na Figura 58, que faz a retirada dos caracteres inválidos:

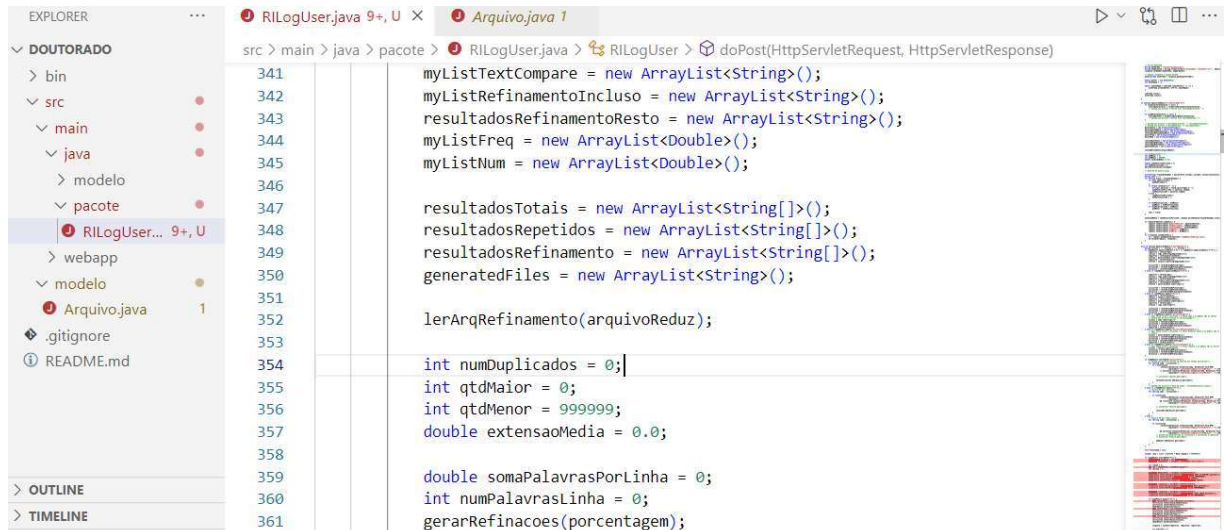
Figura 58. Códigos para retirada de caracteres inválidos



Fonte: elaborada pelo autor.

Na sequência, na Figura 59, são criadas as variáveis estatísticas e chamado o método principal desta seção para gerar métricas de refinamento como número de palavras duplicadas, extensão mínima, máxima e média das expressões.

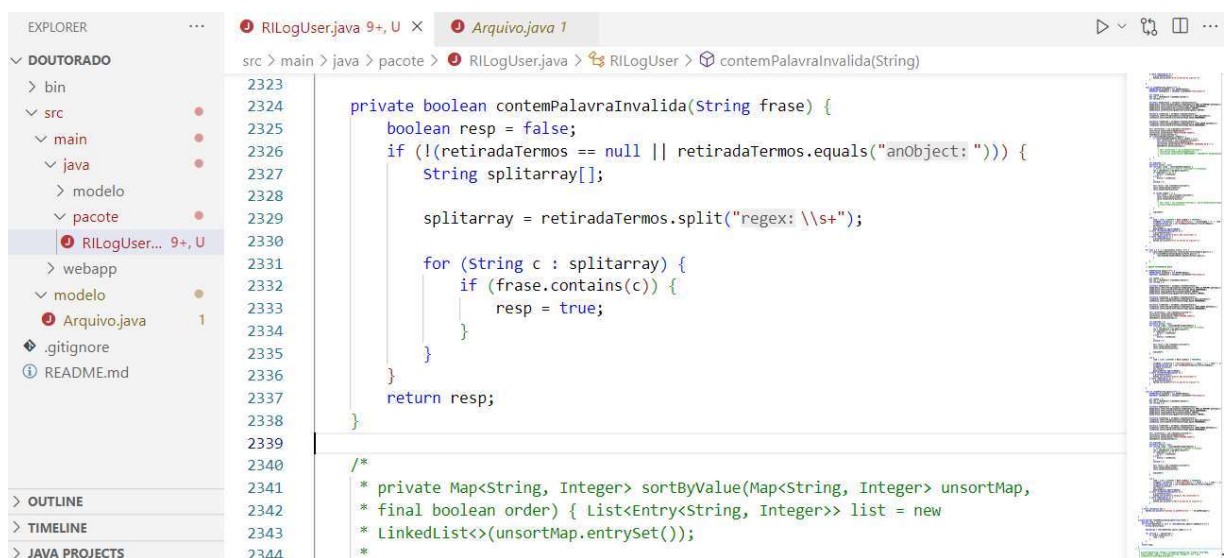
Figura 59. Métricas das listagens dos vocabulários



Fonte: elaborada pelo autor.

Dentro dessa classe, linha 361 da Figura 59, são retiradas as linhas que contêm palavras inválidas através da chamada do método exposto na Figura 60:

Figura 60. Limpeza de palavras inválidas dos vocabulários

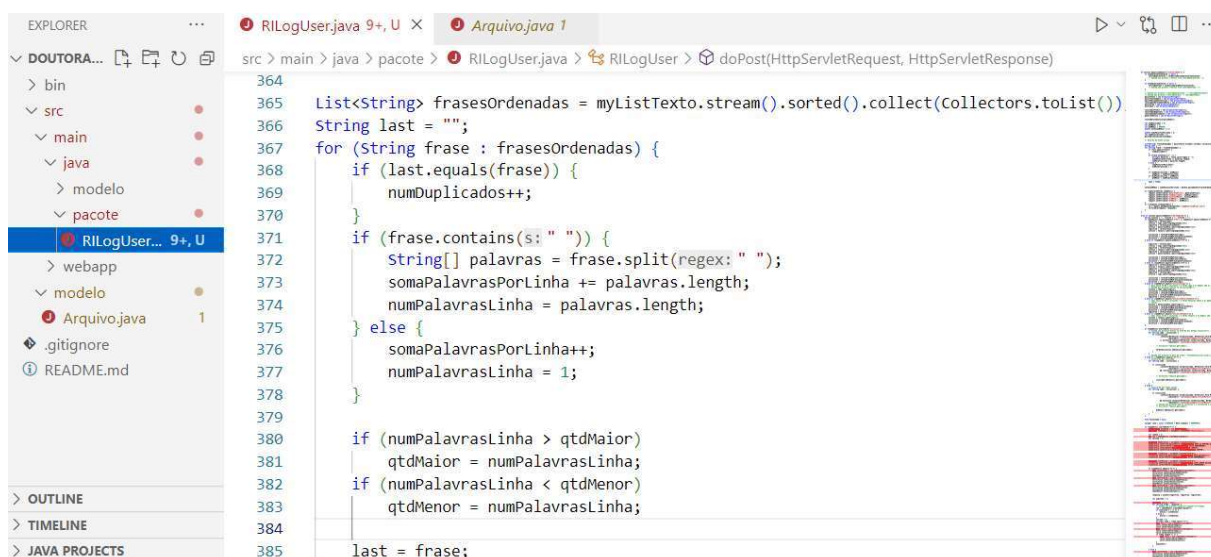


Fonte: elaborada pelo autor.

Uma vez retornado *true* (verdadeiro), conforme exibido na Figura 60, entendemos que existe uma palavra inválida na frase, logo, o método *gerarRefinacoes* passa por ela sem gravá-la em sua lista interna. Também por este método é possível indicar a porcentagem de similaridade escolhida pelo usuário, sendo 100% o máximo para o caso do usuário desejar uma lista idêntica.

Inicialmente, são ordenadas todas as frases por ordem alfabética e comparadas as frases contíguas para identificar se possuem a porcentagem mínima de similaridade indicada. Na sequência, desse processo, dá-se a geração das estatísticas, conforme trecho exposto pela Figura 61:

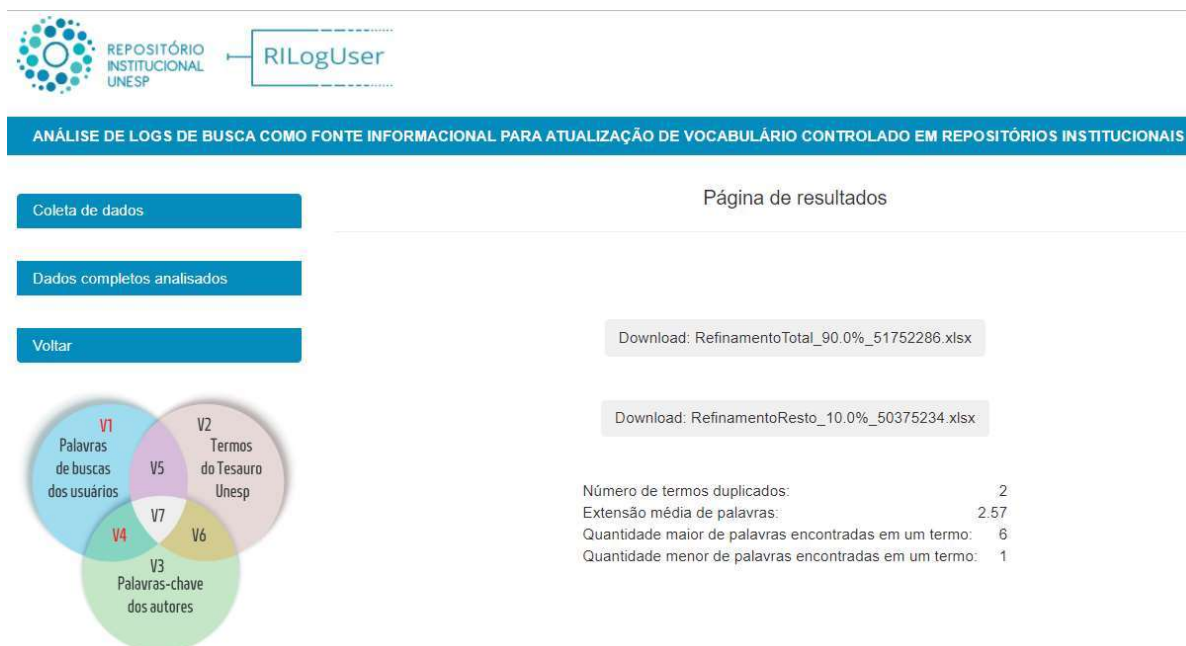
Figura 61. Processo de verificação textual das palavras dos vocabulários



Fonte: elaborada pelo autor.

Por fim, o usuário é redirecionado para a página de resultados do processo de preparação, limpeza e depuração dos dados. Gera-se assim, dois arquivos: os dados do refinamento, respeitando a porcentagem indicada e o que ficou de fora do refinamento quando da mesclagem de palavras. A Figura 62 exhibe exemplo com o resultado do processo com definição de similaridade de 90%:

Figura 62. Resultado do processo de preparação dos dados das listagens dos vocabulários



Fonte: elaborada pelo autor.

Mesmo com as diferentes funcionalidades demonstradas e os resultados obtidos, identificamos duas limitações para o processamento *on-line* dos dados. A primeira é a limitação de 300 linhas para os arquivos de análise de dados e de 5000 linhas para os arquivos da preparação de dados. Essa diferença se dá pela natureza de cada seção e seu consequente número de instruções e uso de memória.

Os arquivos usados para o processamento *off-line* tinham a seguinte quantidade de linhas: Palavras-chave-dos-autores.xlsx (94.216 linhas); Termos-do-tesouro-da-Unesp.xlsx (89.588 linhas); Palavras-de-busca-dos-usuarios.xlsx (7.362 linhas).

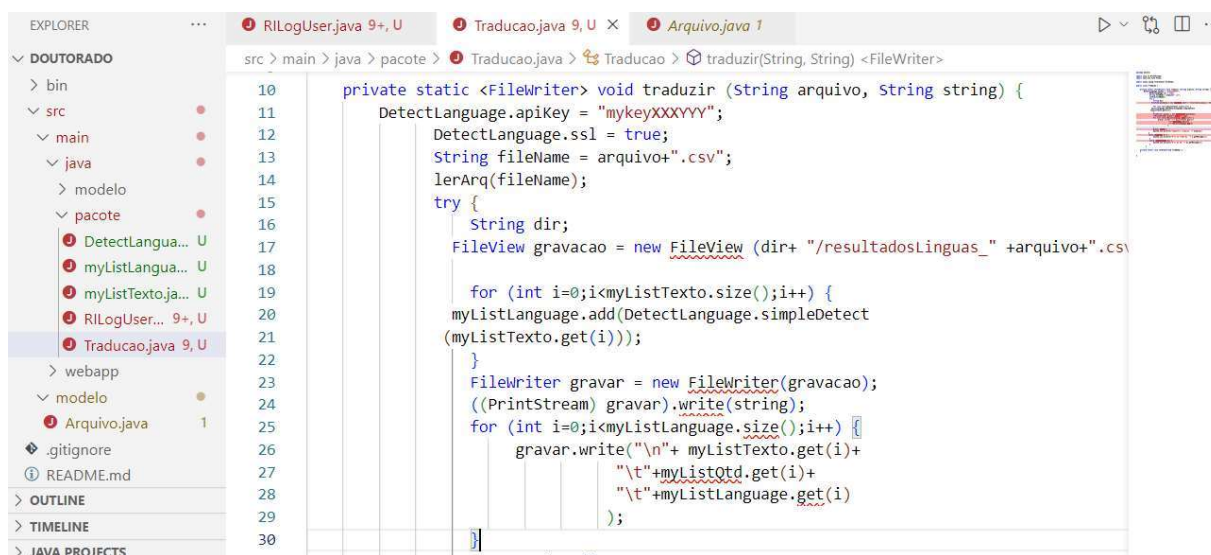
Nesse contexto, ponderando sobre a limitação de linhas, para que pudéssemos fazer uma apresentação do software *on-line* que representasse o mais fielmente possível os arquivos originais, foi feito um levantamento das porcentagens das palavras separadas por línguas (português, inglês, espanhol e outras). A seguir é apresentada a lógica que levou à construção desses arquivos-amostra.

Foi utilizada a API disponível em: <https://detectlanguage.com/> para identificação de idiomas, sendo observada a maior ocorrência de três: inglês, com 172.057 termos (61,8%); português, com 48.732 (17,5%) e espanhol, com 18.520 termos (6,7%). Outros idiomas somados resultaram 38.893 termos restantes (13,9%), no geral. A API *DetectLanguage* na sua versão gratuita permite até mil palavras por dia, como nós tínhamos arquivos com até 90 mil

linhas (e eram três) precisamos da versão paga por um mês apenas, que permitia 100 mil linhas traduzidas por dia.

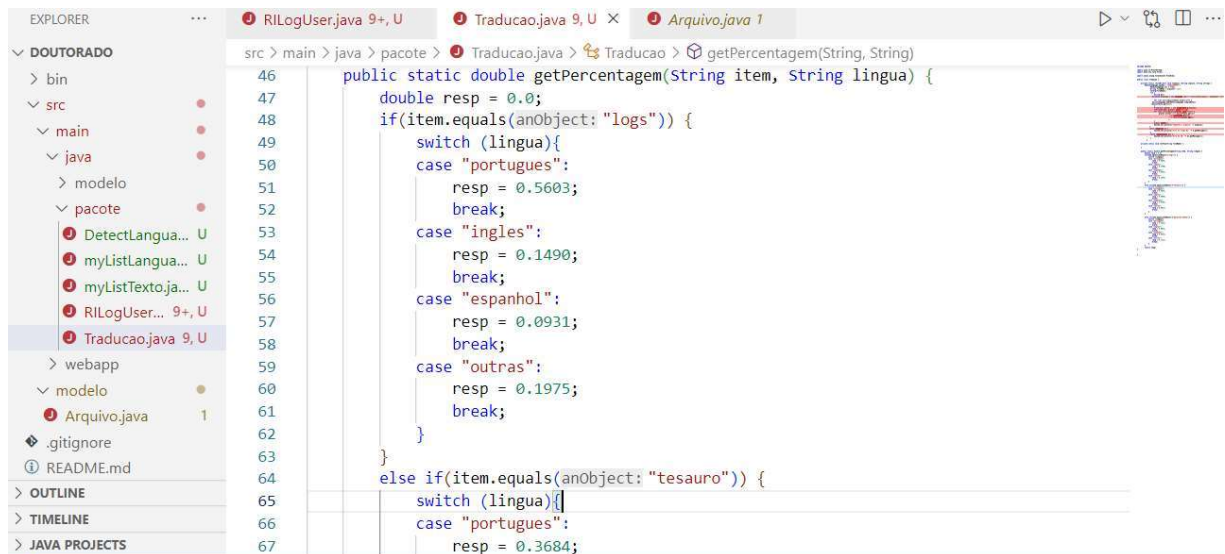
O seguinte método lê os arquivos de entrada, alimentando as listas “myListTexto” e “myListQtd”, que são as frases e seu número de aparições, conforme exposto pela Figura 63:

Figura 63. Verificação dos idiomas nos vocabulários criados



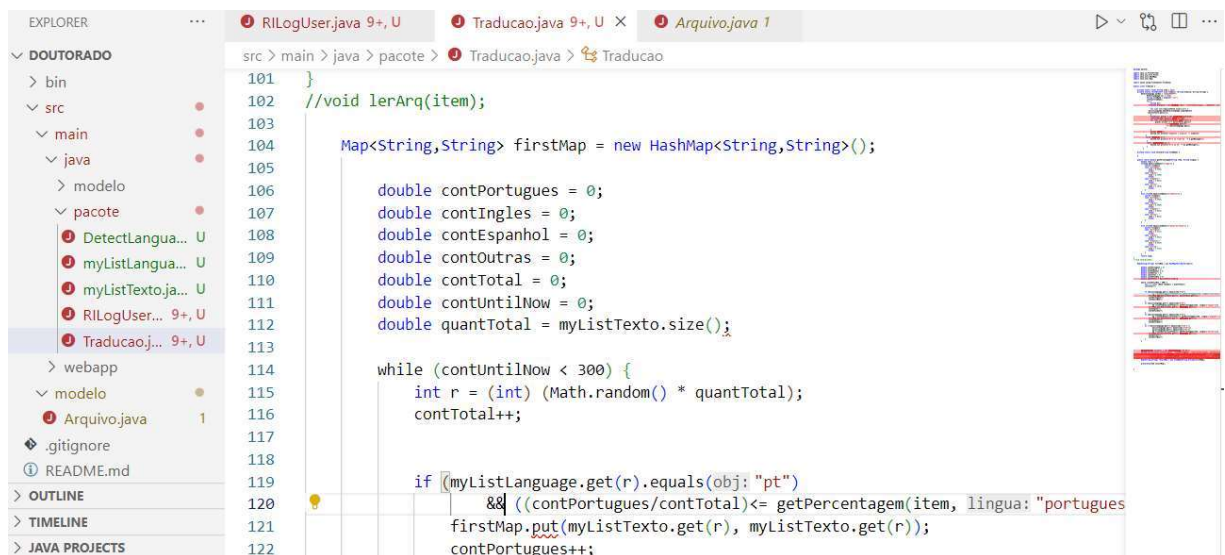
Fonte: elaborada pelo autor.

Toda a classe da Figura 63 foi intitulada “*DetectarLinguas*” que se seguiu da classe abaixo “*ExtractSignificant*”. Para cada arquivo foi determinada a porcentagem de frases em inglês, português, espanhol e a porcentagem da soma das demais línguas, como no código exibido na Figura 64:

Figura 64. Seleção dos idiomas nos vocabulários criados

Fonte: elaborada pelo autor.

Assim foi respeitada a mesma porcentagem para a geração dos arquivos de 300 linhas, fazendo-se um sorteio entre as linhas que tinham as frases dentro da porcentagem esperada, conforme exposto pelo código da Figura 65:

Figura 65. Criação dos arquivos amostrais separados por idiomas dos vocabulários

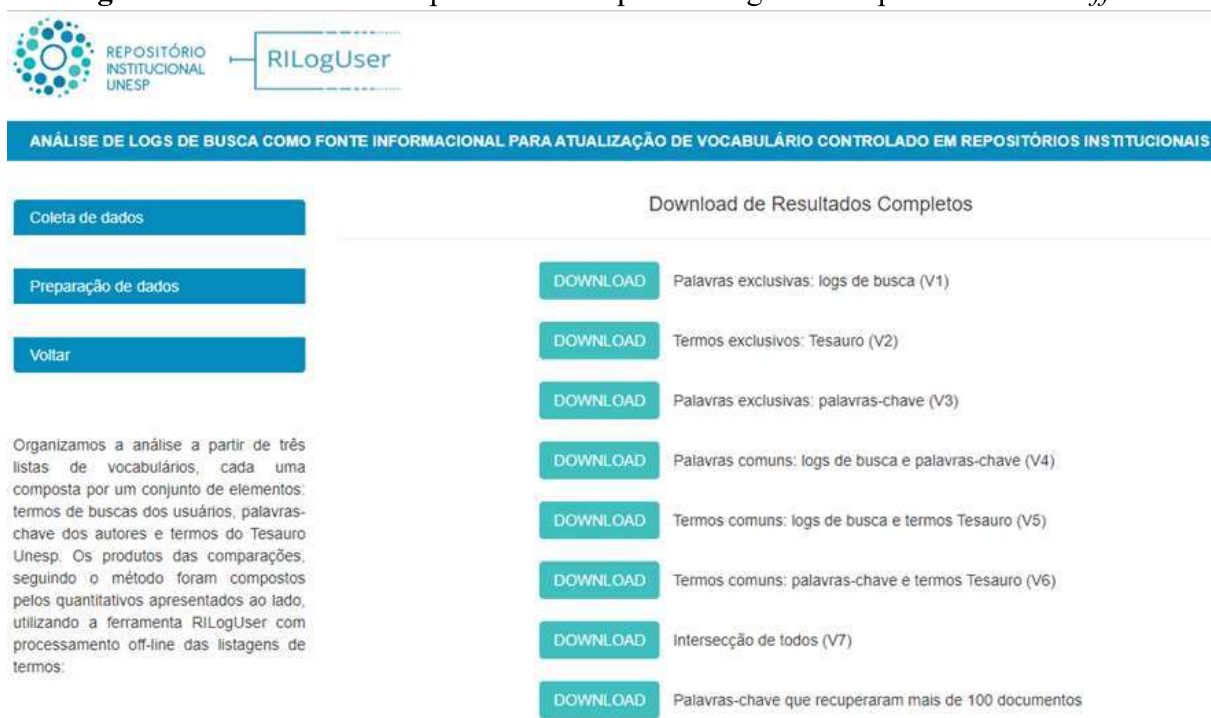
Fonte: elaborada pelo autor.

Pelo código exposto na Figura 65, foram armazenadas as frases e as quantidades representadas em variáveis, as quais foram sorteadas para significar as linhas dos arquivos originais. Assim, o conteúdo das variáveis vai sendo adicionado à amostra desde que esteja dentro da porcentagem daquela língua específica e daquele arquivo de entrada específico.

Algumas estatísticas são geradas na tela. Depois a amostra é ordenada em ordem alfabética crescente e passado como argumento para o método que grava o arquivo.

Por fim, além das funcionalidades já expostas, no protótipo *on-line*, como modo de facilitar o acesso aos dados completos analisados nesta Tese, também estão disponibilizados os *downloads* dos processamentos completos feitos pela ferramenta de forma *off-line*, portanto, sem a limitação de memória do servidor web UOL Host, conforme exposto na Figura 66:

Figura 66. Resultados completos obtidos pela RILogUser em processamento *off-line*



Fonte: elaborada pelo autor.

Assim, na página indicada na Figura 66, o usuário da ferramenta pode fazer os *downloads* completos das palavras exclusivas a cada listagem, palavras coincidentes em duas e/ou três listagens, além de acessar lista com as palavras-chave que geraram recuperação de mais de 100 documentos, compreendidas com de alta representatividade.

Em síntese, a versão *off-line* da RILogUser viabilizou o cumprimento do objetivo geral desta Tese, uma vez que permitiu realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais, assim como evidenciou a análise de logs como fonte informacional relevante e viável. O protótipo *on-line*, por sua vez, demonstrou que é possível implementar a análise em interface simples e amigável, sendo necessário apenas ampliar os investimentos para aumentar a capacidade de armazenamento e processamento.

7 CONSIDERAÇÕES FINAIS

Em meio a extensão de dados disponibilizados diariamente na área da Ciência da Informação, estratégias computacionais para a coleta, preparação e análise de informações relevantes têm ganhado destaque. No cenário das organizações que disponibilizam suas produções científicas por meio de seus repositórios, essa realidade não é diferente, visto que o volume de dados proporcionados pelos indexadores automáticos (coleta automática de dados), ao mesmo tempo que impulsionam avanços na área, também tornam necessário o desenvolvimento de estratégias para analisar os dados produzidos.

Nesse sentido, o uso da análise de logs como fonte informacional mostrou-se viável e produtivo, por permitir o acesso direto a dados dos usuários que pudessem ser convertidos em informações significativas para atualização ou manutenção de vocabulários controlados, em prol da otimização da recuperação de materiais de interesse dos próprios usuários.

Para apresentar as considerações finais desta Tese, é necessário revisitar o problema de pesquisa: como os logs resultantes das pesquisas feitas pelos usuários em um repositório institucional podem auxiliar na atualização de vocabulários controlados no que concerne à padronização terminológica entre os processos de indexação e de busca de materiais? Dessa maneira, os logs podem ser utilizados para compreensão das estratégias de busca de usuários, a fim de subsidiar propostas para a atualização de vocabulários controlados, que viabilizem recuperações mais eficazes e eficientes das informações desejadas.

Desse modo, confirmamos a Tese da pesquisa sobre a análise de logs ser recurso adequado e extensível para, a partir do vocabulário empregado pelos usuários, viabilizar atualizações de vocabulários controlados no RI Unesp.

Ao retomarmos os objetivos específicos como ações essenciais à concretização do objetivo geral de realizar estudo teórico e metodológico sobre a análise de logs, suas funções e sua utilização para atualização de vocabulários controlados para repositórios institucionais, o estudo dos logs de buscas de usuários do RI Unesp para proposta de atualização do vocabulário controlado Tesouro Unesp foi fundamentado na seção 2 desta Tese, a qual trouxe o estado da questão de pesquisa, por meio de revisão da literatura.

Sequencialmente, desenvolvemos a análise experimental, a partir de logs de pesquisa feitas pelos usuários no RI Unesp, da viabilidade metodológica de aplicação de recursos computacionais para realização de rotinas automatizadas para depuração / limpeza de logs de pesquisa para composição de lista de palavras candidatas à inclusão no Tesouro Unesp, subsidiando metodologia específica de análise de logs. Para isso, utilizamos as linguagens de

programação Python e Java e desenvolvemos a ferramenta RILogUser, a qual forneceu listagens de palavras limpas de inconsistências.

Por meio da análise de logs realizada neste estudo, foi possível atingir o terceiro objetivo - criar e investigar a aplicação experimental da análise de logs implementada por meio de ferramenta computacional desenvolvida durante a produção desta Tese, para a atualização de vocabulário controlado do Tesouro Unesp, com vistas à ampliação de seu uso no RI Unesp - com a obtenção de uma lista de expressões de busca contendo palavras simples, palavras compostas ou frases para possibilitar a atualização do vocabulário controlado Tesouro Unesp.

A metodologia aqui apresentada se mostrou eficaz, para realizar uma grande depuração de expressões de busca de usuários e palavras-chave dos autores, que produziu um *corpus* terminológico proveniente diretamente da linguagem utilizada nesse sistema.

Dado o exposto, confirmaram-se os pressupostos da Tese quanto ao grau de interatividade associar-se de maneira diretamente proporcional à efetividade de um repositório. Adicionalmente, a constatação da necessidade contínua de atualização de vocabulários controlados para manutenção de sua função representativa, vincula-se à demanda por vocabulários que se atualizem constantemente por meio do aproveitamento de recursos computacionais, os quais coadunem a velocidade das publicações produzidas e propiciem ambientes adequados à recuperação precisa e intuitiva de materiais. Associativamente, a análise de logs dos usuários se mostrou efetiva para o mapeamento não intrusivo das estratégias de busca de usuários do RI Unesp, a fim de subsidiar ações interventivas pontuais para a melhoria da recuperação de informações.

Com a confirmação dos pressupostos, o problema de pesquisa é respondido, pois a análise de logs se apresentou como fonte confiável de informações, as quais, gerenciadas por estratégias computacionais, têm amplo potencial para otimizar instrumentos de representação, a fim de também beneficiar os processos de busca e, conseqüentemente, de recuperação da informação.

É certo que algumas questões aqui discutidas motivam outras reflexões e demonstram que o processo de indexação de assuntos em repositórios institucionais deve ser ampla e criticamente investigado, a fim de que a Organização do Conhecimento se efetive, de modo que os ciclos informacionais se concretizem ao longo do tempo e pesquisas de grande relevância e impacto científico não sejam negligenciadas por falhas na indexação.

7.1 Limitações da pesquisa

As limitações desta pesquisa são observáveis em três aspectos principais: I) Processamento de linguagem natural; II) Custo para o processamento *on-line*; III) Acesso atualizado aos logs de busca dos usuários.

O processamento de linguagem natural é tópico essencial para otimização das pesquisas, de forma a captar relações semânticas além dos limites estritamente morfológicos, porém, para tanto, é necessária a utilização de processamento com aprendizagem de máquina associada ao emprego de Inteligência Artificial, uma vez que a complexidade envolvida na construção de sentidos requer aprimoramento associativo constante com modelos preditivos que se retroalimentem, na tentativa de acompanharem o desenvolvimento e as transformações naturais das línguas humanas, como utilizam buscadores web como o Google.

A propósito do custo para o processamento *on-line* na ferramenta RILogUser, a hospedagem do servidor, o acesso simultâneo por vários usuários, a contratação de memória RAM e de espaço de armazenamento suficientes para o processamento em uma escala de grande volume de dados, como obtido nas listagens descritas, custaria mensalmente milhares de reais. Dado o custo elevado do investimento, até a defesa desta Tese, o processamento completo de listagens de termos permanece restrito ao processamento *off-line*, no qual a interface é menos plástica, porém o processamento é rápido (cerca de 12 minutos com as listas atuais) e executável em uma máquina com configurações básicas (processador i5, 8GB-RAM e SSD 256GB).

Sobre o acesso atualizado e permanente aos logs, a limitação da coleta pelos pesquisadores se dá, tanto pelas restrições referidas na LGPD, quanto pelo volume diário de dados que requerem grande capacidade de processamento e armazenagem para análise. Destacamos ainda que a pesquisa concretizada nesta Tese está amparada pelo item IV do Art. 7º da LGPD, o qual estabelece que o tratamento de dados pessoais poderá ser executado “para a realização de estudos por órgão de pesquisa, garantida, sempre que possível, a anonimização dos dados pessoais”, como tratado nesta Tese (BRASIL, 2018).

7.2 Trabalhos futuros

No decorrer desta pesquisa, visualizamos algumas direções para realização de trabalhos futuros. A seguir, apresentam-se algumas delas:

1. Busca/desenvolvimento de estratégias para Processamento de Linguagem Natural para enriquecer a RILogUser;
2. Aprimoramento da versão *on-line* da RILogUser, a fim de que tenha maior capacidade de processamento e armazenagem de dados e permaneça como aplicação economicamente viável;

3. Investigação extensiva da aplicação e adaptabilidade da RILogUser a outros RI em escopos mais amplos e diversificados;
4. Análise da possibilidade de integração da RILogUser com outros projetos do RI Unesp, evidenciando, assim, a incorporação do processo de enriquecimento de vocabulários naturais com outros processos de manutenção e atualização de vocabulários controlados;
5. Estudos complementares e parecer técnico, por parte do grupo gestor responsável pela criação e atualização do Tesouro Unesp, sobre a pertinência e potencial das listagens obtidas pela ferramenta RILogUser.

REFERÊNCIAS

ALBRECHTSEN, H. Subject analysis and indexing: from automated indexing to domain analysis. *The Indexer: the international journal of indexing*, London, v.18, n. 4, p. 219-224, 1993. DOI: <https://doi.org/10.3828/indexer.1993.18.4.3>.

ALMEIDA, M. C. B. *Planejamento de bibliotecas e serviços de informação*. 2. ed. rev. e ampl. Brasília: Briquet Lemos, 2005.

ASSUMPÇÃO, F. S *et al.* A conversão de registros na implantação de repositórios institucionais: o caso do Repositório Institucional Unesp. *In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS*, 18., 2014, Belo Horizonte. *Anais....* Belo Horizonte: UFMG, 2014. p. 1-16. Disponível em: <http://hdl.handle.net/11449/123645>. Acesso em: 10 ago. 2020.

AVERY, S.; TRACY, D. G. Using transaction log analysis to assess student search behavior in the library instruction classroom. *Reference Services Review*, v. 42, n. 2, p. 320-335, 2014. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/RSR-08-2013-0044/full/html>. Acesso em: 12 jul. 2020. DOI: <https://doi.org/10.1108/RSR-08-2013-0044>.

BACHA, M. N.; ALMEIDA, M. S.G. Vocabulário controlado e palavras-chave em repositórios digitais: relato de experiência do repositório institucional da FGV. *In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO-FEBAB*, 2013. *Anais...* p. 1439-1447.

BARBOSA, J. *et al.* Introdução ao processamento de linguagem natural usando Python. *III Escola Regional de Informática do Piauí*, v. 1, n. 1, p. 336-360, 2017. Disponível em: http://www.facom.ufu.br/~wendelmelo/terceiros/tutorial_nltk.pdf. Acesso em: 21 jun. 2020.

BARBOSA, P. F. M.; PERICO, D. H. Desenvolvimento de material didático para ensino a distância de Python básico. *In: SIMPÓSIO DE INICIAÇÃO CIENTÍFICA, DIDÁTICA E DE AÇÕES SOCIAIS DA FEI*, 9., 2019, São Bernardo do Campo. *Anais...* São Bernardo do Campo: FEI, 2019. p. 1-2. Disponível em: https://fei.edu.br/sites/sicfei/2019/cc/SICFEI_2019_paper_87.pdf. Acesso em: 27 abr. 2021.

BARITÉ ROQUETA, M. G. El control de vocabulario en la era digital: revisión conceptual. *Scire: representación y organización del conocimiento*, v. 20, n. 1, p. 99-108, 2014. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/168022>. Acesso em: 21 fev. 2021.

BERNSTAM, E. V.; HERSKOVIC, J. R.; HERSH, W. R. Query log analysis in biomedicine. *In: JANSEN, B. J.; SPINK, A.; TAKSA, I. (ed.). Handbook of research on web log analysis*. Hershey, PA: IGI Global, 2008. Chapter 18, p. 359-377. Disponível em: <https://www.igi-global.com/chapter/query-log-analysis-biomedicine/22010>. Acesso em: 21 jul. 2020.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. Sepastopol, CA: O'Reilly Media, 2009.

BLECIC, D. D. *et al.* Using transaction log analysis to improve OPAC retrieval results. *College & Research Libraries*, v. 59, n. 1, p. 39-50, 1998. Disponível em: <https://crl.acrl.org/index.php/crl/article/view/15180/16626>. Acesso em: 15 jul. 2020. DOI: <https://doi.org/10.5860/crl.59.1.39>.

BORKO, H.; BERNIER, C. L. *Indexing concepts and methods*. New York: Academic Press, 1978.

BRASIL. *Lei nº 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, [2018]. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 21 abr. 2022.

BRITO, J. P.; SOUSA, M. R. F. Projetando a experiência do usuário no Laboratório de Tecnologias Intelectuais- LTi. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia*, v. 13, n. 2, 2018. Disponível em: <https://brapci.inf.br/index.php/res/v/109111>. Acesso em: 26 jul. 2020. DOI: 10.22478/ufpb.1981-0695.2018v13n2.43241

CAFÉ, L.; KAFURE MUÑOZ, I. Avaliação de usabilidade no Repositório Institucional da Universidade de Brasília. *Informação & Tecnologia (ITEC)*, v. 3, n. 2, p. 39-61, 2016. (Especial ENANCIB 2016 – Parte 2). Disponível em: <https://periodicos.ufpb.br/index.php/itec/article/view/38431/20165>. Acesso em: 5 jul. 2021.

CARLAN, E.; MEDEIROS, M. B. B. Sistemas de Organização do Conhecimento na visão da Ciência da Informação. *Revista Ibero-Americana de Ciência da Informação*, v. 4, n. 2, 2011. Disponível em: <https://periodicos.unb.br/index.php/RICI/article/view/1675>. Acesso em: 16 maio 2020. DOI: <https://doi.org/10.26512/rici.v4.n2.2011.1675>.

CHING, W. K. *et al.* *Markov chains: models, algorithms, and applications*. New York: Springer, 2013. (International series in operations research & management science, v. 189). DOI: 10.1007/978-1-4614-6312-2.

CHO, Junghoo; GARCIA-MOLINA, Hector. *The evolution of the web and implications for an incremental crawler*. Stanford, 1999. Disponível em: <http://ilpubs.stanford.edu:8090/376/1/1999-22.pdf>. Acesso em: 16 julho 2021.

CLEVELAND, W. S. Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, v. 69, n. 1, p. 21-26, 2001. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2001.tb00477.x>. Acesso em: 16 maio 2020.

CORRÊA, R. F.; BAZÍLIO, L. H. Análise da extração de descritores como sintagmas nominais através do software OGMA. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, v. 22, n. 50, p. 44-58, 2017. DOI:10.5007/1518-2924.2017v22n50p44.

COSTA, M. *Características e contribuições da Via Verde para o acesso aberto à informação científica na América Latina*. 2014. Tese (Doutorado em Ciência da Informação) –

Universidade de Brasília, Brasília-DF, 2014. Disponível em:

http://eprints.rclis.org/28968/1/Michelli%20Costa_disserta%C3%A7%C3%A3o_2014.pdf.

Acesso em: 13 jun. 2020.

CROW, R. The case for institutional repositories: a SPARC position paper. *ARL*, n. 223, p. 1-4, 2002. Disponível em: <http://www.arl.org/newsltr/223/instrepo.html> . Acesso em: 25 de nov. 2020.

DODEBEI, V. Repositórios institucionais: por uma memória criativa no ciberespaço. In: SAYÃO, L. F. *et al.* (org.). *Implantação e gestão de repositórios institucionais: políticas, memória, livre acesso e preservação*. Salvador: EDUFBA, 2009. p. 91-92. Disponível em: http://www.repositorio.ufba.br/ri/bitstream/ufba/473/3/implantacao_repositorio_web.pdf. Acesso em: 11 jun. 2020.

FARIAS, R. A. N.; SANTOS, R. F. Análise do uso de repositórios digitais da Universidade Federal do Rio Grande do Norte. *Revista Informação na Sociedade Contemporânea*, v. 1, p. 1-20, 2017. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/219442/PCIN0243-T.pdf?sequence=-1> . Acesso em: 15 mar. 2020.

FERNANDES, W. M.; VECHIATO, F. L. Encontrabilidade da informação no repositório institucional da Unesp: uma avaliação com dispositivos móveis. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 18., 2017. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/104997>. Acesso em: 11 jun. 2020.

FERNEDA, E. *Introdução aos modelos computacionais de recuperação de informação*. Rio de Janeiro: Ciência Moderna, 2012.

FIDEL, R. User-centered indexing. *Journal of the American Society for Information Science*, Washington, v.45, n. 8, p. 572-576, 1994.

FISCHER, R. K. *et al.* A transaction log analysis of EBSCO Discovery Service using Google Analytics: the methodology. *Library Hi Tech*, v. 39, n. 1, p. 249-262, 2020. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/LHT-09-2019-0199/full/html>. Acesso em: 13 mar. 2021. DOI: <https://doi.org/10.1108/LHT-09-2019-0199>.

FREITAS, R. C.; FREGADOLLI, V. K.; FREITAS, M. C. D. A informação na intermediação de transações em plataformas on-line: a inovação da empresa Airbnb. *P2P E INOVAÇÃO*, v. 6, n. 2, p. 57-72, 2020. Disponível em: <http://revista.ibict.br/p2p/article/view/5091>. Acesso em: 26 jul. 2020. DOI: 10.21721/p2p.2020v6n2.p57-72.

FUJITA, M. S. L. Apresentação. In: GIL LEIVA, I.; FUJITA, M. S. L. (ed.) *Política de indexação*. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2012a. p. 13-16.

Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/politica-de-indexacao_ebook.pdf. Acesso em: 20 jun. 2020.

FUJITA, M. S. L. A política de indexação para representação e recuperação da informação. In: GIL LEIVA, I.; FUJITA, M. S. L. (ed.) *Política de indexação*. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2012b. p. 17- 28.

Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/politica-de-indexacao_ebook.pdf. Acesso em: 20 jun. 2020.

FUJITA, M. S. L. *Controle de vocabulário na representação e recuperação da informação em repositórios universitários*. 2017. Projeto de Pesquisa (Bolsa de Produtividade em Pesquisa) - CNPq, Programa de Pós-Graduação em Ciência da Informação, Unesp, Marília, 2018.

FUJITA, M. S. L. *Controle de vocabulário na representação e recuperação da informação em repositórios universitários*. 2022. Projeto de Pesquisa (Bolsa de Produtividade em Pesquisa) - CNPq, Programa de Pós-Graduação em Ciência da Informação, Unesp, Marília, 2022.

FUJITA, M. S. L.; ALMEIDA, C. C. O leitor indexador: expert e profissional. In: FUJITA, M. S. L.; ALVES, R. C. V.; ALMEIDA, C. C. (org.) *Modelos de leitura documentária para indexação: abordagens teóricas interdisciplinares e aplicações em diferentes tipos de documentos*. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2020. DOI: <https://doi.org/10.36311/2020.978-65-86546-07-1.p43-66>.

FUJITA, M. S. L.; MOREIRA, W. *Manual do planejamento, construção e manutenção do tesouro Unesp para bibliotecas: do conceitual a práxis*. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2021. Disponível em: https://ebooks.marilia.unesp.br/index.php/lab_editorial/catalog/book/234. Acesso em: 27 nov. 2021. DOI: <https://doi.org/10.36311/2021.978-65-5954-069-3>.

FUJITA, M. S. L.; TARTAROTTI, R. C. D. Análise de palavras-chave da produção científica de pesquisadores: o autor como indexador. *Informação & Informação*, Londrina, v. 25, n. 3, p. 332-374, jul./set. 2020. Disponível em: <https://www.uel.br/revistas/uel/index.php/informacao/article/view/41866/pdf>. Acesso em: 13 dez. 2020.

FUJITA, M. S. L.; TOLARE, J. B. Vocabulários controlados na representação e recuperação da informação em repositórios brasileiros. *Informação & Informação*, Londrina v. 24, n. 2, p. 93-125, 2019. Disponível em: <https://www.uel.br/revistas/uel/index.php/informacao/article/view/37985/pdf>. Acesso em: 29 set. 2020.

GARCÍA GUTIÉRREZ, A. L. *Linguística documental: aplicación a la documentación de la comunicación social*. Barcelona: Mitre, 1984.

GIL, A. C. *Métodos e técnicas de pesquisa social*. 3. ed. São Paulo: Atlas, 1999.

GIL LEIVA, I. Aspectos conceituais da indexação. *In: Política de indexação*. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2012. p. 31- 105. Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/politica-de-indexacao_ebook.pdf. Acesso em: 20 jun. 2020.

GIL LEIVA, I. SISA—Automatic Indexing System for Scientific Articles: experiments with Location Heuristics Rules versus TF-IDF Rules. *Knowledge Organization*, v. 44, n. 3, p. 139-162, 2017. Disponível em: [10.5771/0943-7444-2017-3-139](https://doi.org/10.5771/0943-7444-2017-3-139). Acesso em: 23 out. 2020.

GIL LEIVA, I.; FUJITA, M. S. L. (ed.). *Política de indexação*. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2012. Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/politica-de-indexacao_ebook.pdf. Acesso em: 20 jun. 2020.

GIL LEIVA, I.; KATAHIRA, I.; FUJITA, M. S. L. Importância do controle de vocabulário em um repositório institucional: estudo de caso. *In: SILVA, C. G.; REVEZ, J.; CORUJO, L. (ed.) Organização do conhecimento no horizonte 2030: desenvolvimento sustentável e saúde*. Lisboa: Centro de Estudos Clássicos, 2021. (Coleção Ciência Aberta – CA, n. 1). P. 447-459. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=8411194>. Acesso em: 20 jan. 2022.

GONZÁLEZ, M. M.; NIETO, M. M. Thesaurusapi: una api para la manipulación de tesauros. *Scire: representación y organización del conocimiento*, v. 18, n. 1, p. 41-48, 2012. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/167709>. Acesso em: 21 ago. 2021.

GONÇALEZ, P. R. V. A. *et al.* Repositórios de dados na América Latina: a ISO 16363:2012 e a representação da informação. *In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB*, 19., 2018. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/103008>. Acesso em: 21 abr. 2022.

HERRERÍAS, J.; GÓMEZ, R. Log analysis towards an automated forensic diagnosis system. *In: INTERNATIONAL CONFERENCE ON AVAILABILITY, RELIABILITY AND SECURITY*, 2010, Krakow. IEEE, 2010. p. 659-664. Disponível em: <https://ieeexplore.ieee.org/document/5438019>. Acesso em: 23 jul. 2020. DOI: 10.1109/ARES.2010.120.

HUURNINK, B. *et al.* Search behavior of media professionals at an audiovisual archive: a transaction log analysis. *Journal of the American society for information science and technology*, v. 61, n. 6, p. 1180-1197, 2010. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/asi.21327>. Acesso em: 20 jun. 2020. DOI: <https://doi.org/10.1002/asi.21327>.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *ISO 25964-1: Information and documentation – Thesauri and interoperability with other vocabularies*. Genebra, 2011.

JAKOBSON, R. *Linguística e comunicação*. 22.ed. Tradução de Izidoro Blikstein; José Paulo Paes. São Paulo: Cultrix, 2010.

JANSEN, B. J. Search log analysis: what it is, what's been done, how to do it. *Library & information science research*, v. 28, n. 3, p. 407-432, 2006. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0740818806000673>. Acesso em: 27 jul. 2020. DOI: <https://doi.org/10.1016/j.lisr.2006.06.005>.

JANSEN, B. J. The methodology of search Log Analysis. In: JANSEN, B. J.; SPINK, A.; TAKSA, I. (ed.). *Handbook of research on web log analysis*. Hershey, PA: IGI Global, 2008. Chapter 6, p. 100-123. Disponível em: <https://www.igi-global.com/chapter/methodology-search-log-analysis/21998>. Acesso em: 23 jul. 2020. DOI: 10.4018/978-1-59904-974-8.ch006.

JANSEN, B. J.; SPINK, A.; TAKSA, I. (ed.) *Handbook of research on web log analysis*. Hershey, PA: IGI Global, 2008. Disponível em: <https://www.igi-global.com/gateway/book/520>. Acesso em: 20 jul. 2020. DOI: 10.4018/978-1-59904-974-8.

JANSEN, J.; TAKSA, I.; SPINK, A. Research and methodological foundations of transaction log analysis. In: JANSEN, B. J.; SPINK, A.; TAKSA, I. (ed.). *Handbook of research on web log analysis*. Hershey, PA: IGI Global, 2008. Chapter 1, p.1-16. Disponível em: <https://www.igi-global.com/gateway/chapter/21993#pnlRecommendationForm>. Acesso em: 20 jul. 2020. DOI: 10.4018/978-1-59904-974-8.ch001.

JONES, R. Improving Okapi: transaction log analysis of failed searches in an online catalogue. *Vine*, v. 16, n. 1, p. 3-13, 1986. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb040341/full/html>. Acesso em: 18 jul. 2020. DOI: <https://doi.org/10.1108/eb040341>.

JONES, S. *et al.* A transaction log analysis of a digital library. *International journal on digital libraries*, v. 3, n. 2, p. 152-169, 2000. Disponível em: <https://link.springer.com/article/10.1007/s007999900022>. Acesso em: 12 jul. 2020. DOI: <https://doi.org/10.1007/s007999900022>.

KACPRZAK, E. *et al.* A query log analysis of dataset search. In: CABOT, J.; DE VIRGILIO, R.; TORLONE, R. (ed.) *Web Engineering*. Rome: Springer, 2017. p. 429-436. (International Conference on Web Engineering. ICWE 2017. Lecture notes in Computer Science, v. 10360). Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-60131-1_29. Acesso em: 15 jun. 2020. DOI: https://doi.org/10.1007/978-3-319-60131-1_29.

KASKE, N. K. Research methodologies and transaction log analysis: issues, questions, and a proposed model. *Library Hi Tech*, v. 11, n. 2, p. 79-86, 1993. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb047886/full/html>. Acesso em: 14 jul. 2020. DOI: <https://doi.org/10.1108/eb047886>.

KINSELLA, J.; BRYANT, P. Online public access catalog research in the United Kingdom: an overview. *Library Trends*, v. 35, n. 4, p. 619-628, 1987. Disponível em: <http://hdl.handle.net/2142/7508>. Acesso em: 11 jun. 2020.

KITCHENHAM, B. *Procedures for performing systematic reviews*. Keele - Staffordshire: Keele University – Department of Computer Science, 2004. (Joint Technical Report. Keele University Technical Report TR/SE-0401) Disponível em: <https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf>. Acesso em: 10 jul. 2020.

KRIPKA, Rosana Maria Luvezute; SCHELLER, Morgana; BONOTTO, D. de L. Pesquisa documental na pesquisa qualitativa: conceitos e caracterização. *Revista de investigaciones UNAD*, v. 14, n. 2, p. 55-73, 2015.

KURTH, M. The limits and limitations of transaction log analysis. *Library Hi Tech*, v. 11, n. 2, p. 98-104, 1993. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb047888/full/html>. Acesso em: 14 jul. 2020. DOI: <https://doi.org/10.1108/eb047888>.

LAIPELT, R. C. F. A análise de logs como estratégia para a realização da garantia do usuário. *Em Questão*, v. 21, n. 3, p. 150-170, 2015a. Disponível em: <https://seer.ufrgs.br/EmQuestao/article/view/59806/36047>. Acesso em: 12 jun. 2020. DOI: <https://doi.org/10.19132/1808-5245213.150-170>.

LAIPELT, R. C. F. *Metodologia para seleção de termos equivalentes e descritores de tesouros: um estudo no âmbito do Direito do Trabalho e do Direito Previdenciário*. 2015. Tese (Doutorado em Linguística Aplicada) – Escola da Indústria Criativa, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2015b. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/4853>. Acesso em: 13 jun. 2020.

LANCASTER, F.W. *Indexing and abstracting in theory and practice*. London: The Library Association, 1991.

LANCASTER, F. W. *Indexação e resumos*. Tradução de A. A. Briquet de Lemos. Brasília: Briquet de Lemos, 2004.

LEITE, F. C. L. *Como gerenciar e ampliar a visibilidade da informação científica brasileira: repositórios institucionais de acesso aberto*. Brasília-DF: Instituto Brasileiro de Informação em Ciência e Tecnologia, 2009. Disponível em: <https://livroaberto.ibict.br/bitstream/1/775/4/Como%20gerenciar%20e%20ampliar%20a%20visibilidade%20da%20informa%C3%A7%C3%A3o%20cient%C3%ADfica%20brasileira.pdf>. Acesso em: 15 out. 2021.

LI, Y. *et al.* MR-CLOPE: A MapReduce based transactional clustering algorithm for DNS query log analysis. *Journal of Central South University*, v. 22, n. 9, p. 3485-3494, 2015. Disponível em: <https://link.springer.com/article/10.1007/s11771-015-2888-9>. Acesso em: 23 jun. 2020. DOI: <https://doi.org/10.1007/s11771-015-2888-9>.

LIMA, S. M. A. *Avaliação do uso do Repositório Institucional da UNILAB pelos discentes de graduação*. 2021. 140 f. Dissertação (Mestrado em Ciência da Informação) – Centro de Humanidades, Universidade Federal do Ceará, Fortaleza, 2021. Disponível em: <https://repositorio.unilab.edu.br/jspui/handle/123456789/2258>. Acesso em: 8 jan. 2022.

LISBOA, I. V. M. V.; BARROSO, J. M. P.; ROCHA, T. J. V. Acessibilidade dos conteúdos educacionais online na perspectiva da experiência do aluno cego/Accessibility of educational online contents on the perspective of blind student's UX. *Brazilian Journal of Development*, v. 6, n. 6, p. 35623-35628, 2020. Disponível em: <https://www.brazilianjournals.com/index.php/BRJD/article/viewFile/11354/9476>. Acesso em: 26 jul. 2020.

LOPES, I. L. Estratégia de busca na recuperação da informação: revisão da literatura. *Ciência da Informação*, v. 31, n. 2, p. 60-71, 2002. Disponível em: <https://www.scielo.br/j/ci/a/VPbDtPgqnFs5cm8GSLhtsyM/?format=pdf&lang=pt>. Acesso em: 5 jan. 2021

LUCAS, E. O.; LARA, M. L. G. Palavras-chave como elo entre artigos e autores: visualizações possíveis. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 4., 2014, Recife. *Anais...* Recife, 2014. Disponível em: https://www.brapci.inf.br/_repositorio/2014/05/pdf_138a07e357_0014319.pdf. Acesso em: 03 abr. 2022.

LYNCH, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. *Portal: Libraries and the Academy*, v. 3, n. 2, p. 327-336, 2003. Disponível em: <https://muse.jhu.edu/article/42865/pdf>. Acesso em: 11 jun. 2020.

MAHOUI, M.; CUNNINGHAM, S. J. Search behavior in a research-oriented digital library. In: CONSTANTOPOULOS, P.; SOLVBERG, I. (ed.) *Research and advanced technology for digital libraries*. Berlin: Springer, 2001. p. 13-24. (European Conference on Digital Libraries – ECDL 2001. Lecture notes in Computer Science, v. 2163). Disponível em: https://www.researchgate.net/publication/221176181_Search_Behavior_in_a_Research-Oriented_Digital_Library. Acesso em 13 jul. 2020. DOI: 10.1007/3-540-44796-2_2.

MAHOUI, M.; CUNNINGHAM, S. J. A comparative transaction log analysis of two computing collections. In: BORBINHA, J.; BAKER, T. (ed.) *Research and advanced technology for digital libraries*. Berlin: Springer-Verlag, 2000. p. 418-423. (International Conference on Theory and Practice of Digital Libraries. ECDL 2000. Lecture notes in Computer Science, v. 1923). Disponível em: https://link.springer.com/chapter/10.1007/3-540-45268-0_53. Acesso em: 15 jul. 2020. DOI: https://doi.org/10.1007/3-540-45268-0_53.

MAI, J. E. Analysis in indexing: document and domain centered approaches. *Information Processing & Management*, Elmsford, v.41, n.3, p. 599-611, 2005. Disponível em: <https://doi.org/10.1016/j.ipm.2003.12.004>. Acesso em: 23 jun. 2020.

MARCONI, M. A.; LAKATOS, E. M. *Metodologia científica*. 7. ed. São Paulo: Atlas, 2017.

MARQUES, C. A. G.; VECHIATO, F. L. Interoperabilidade entre Repositório Institucional e Sistema Integrado de Gestão de Atividades Acadêmicas. *Informação & Tecnologia (ITEC)*, v. 4, n. 2, p. 4-25, 2017. Disponível em: <https://periodicos.ufpb.br/index.php/itec/article/view/40445/21035>. Acesso em: 8 jul. 2021.

MARTÍNEZ-COMECHÉ, J. A. Determinación de grupos de usuarios de bibliotecas digitales mediante el análisis de ficheros log. *Revista española de Documentación Científica*, v. 40, n. 3, e181, 2017. Disponível em: <http://redc.revistas.csic.es/index.php/redc/article/view/985/1520>. Acesso em: 26 jul. 2020.

MEISTER, D.; SULLIVAN, D. J. *Evaluation of user reactions to a prototype on-line information retrieval system*. Washington, D.C.: National Aeronautics and Space Administration, 1967. (NASA CR – 918). Disponível em: <https://ntrs.nasa.gov/api/citations/19670030754/downloads/19670030754.pdf>. Acesso em: 12 jul. 2020.

MIGUEIS, A. *et al.* A importância das palavras-chave dos artigos científicos da área das Ciências Farmacêuticas, depositados no Estudo Geral: estudo comparativo com os termos atribuídos na MEDLINE. *InCID: Revista de Ciência da Informação e Documentação*, v. 4, n. 2, p. 112-125, jul./dez. 2013. Edição especial. Disponível em: <https://www.revistas.usp.br/incid/article/view/69284>. Acesso em: 20 set. 2020.

MIRANDA, A. M. M.; ALCARÁ, A. R. Busca e uso da informação: uma abordagem sobre as habilidades informacionais de universitários. *Biblionline*, João Pessoa, v. 12, n. 3, p. 94-111, jul./set. 2016. Disponível em: <https://periodicos.ufpb.br/index.php/biblio/article/view/28897>. Acesso em: 10 maio 2021.

MISCHO, W. H.; NORMAN, M. A.; SCHLEMBACH, M. C. Transaction log analysis within a bento discovery system. In: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES (JCDL), 2019, Champaign. IL. IEEE, 2019. p. 420-421. Disponível em: <https://ieeexplore.ieee.org/document/8791111>. Acesso em: 20 jul. 2020. DOI: 10.1109/JCDL.2019.00099.

MOUKDAD, H.; LARGE, A. Users' perceptions of the Web as revealed by transaction log analysis. *Online information review*, v. 25, n. 6, p. 349-359, 2001. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/EUM0000000006534/full/html>. Acesso em: 23 jul. 2020. DOI: <https://doi.org/10.1108/EUM0000000006534>.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia. *Informação & Sociedade: Estudos*, v. 19, n. 2, p. 99-118, 2009. Disponível em: <https://periodicos.ufpb.br/ojs/index.php/ies/article/view/2925/3028>. Acesso em: 14 jul. 2020.

NATIONAL INFORMATION STANDARDS ORGANIZATION (NISO). *ANSI/NISO Z39.19-2005 (R2010)*. Guidelines for the construction, format, and management of monolingual controlled vocabularies. Baltimore, MD: NISO, 2010. Disponível em: <https://bityli.com/e4Sun>. Acesso em: 21 de mar. 2021.

NETO, J. J.; FERNEDA, E. Ontologia como recurso de padronização terminológica no processo de recuperação de informação. *Informação em Pauta*, v. 1, n. 1, p. 30-45, 2016. Disponível em: <https://dialnet.unirioja.es/descarga/articulo/6254130.pdf>. Acesso em: 10 abr. 2021.

NICHOLAS, D.; HUNTINGTON, P.; WATKINSON, A. Scholarly journal usage: the results of deep log analysis. *Journal of documentation*, v. 61, n.2, p. 248-280, 2005. Disponível em: <https://doi.org/10.1108/00220410510585214>. Acesso em: 12 set 2021.

NIWATTANAKUL, S. *et al.* Using of Jaccard coefficient for keywords similarity. In: INTERNATIONAL MULTICONFERENCE OF ENGINEERS AND COMPUTER SCIENTISTS, 2013, Hong Kong. *Proceedings...2013*. V. 1. Disponível em: https://www.researchgate.net/publication/317248581_Using_of_Jaccard_Coefficient_for_Keywords_Similarity. Acesso em: 3 jul. 2020.

OBRIEN, P. *et al.* RAMP—the Repository Analytics and Metrics Portal. A prototype web service that accurately counts item downloads from institutional repositories. *Library Hi Tech*, v. 35, n. 1, p. 144-158, 2017. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/LHT-11-2016-0122/full/pdf?title=ramp-the-repository-analytics-and-metrics-portal-a-prototype-web-service-that-accurately-counts-item-downloads-from-institutional-repositories>. Acesso em: 2 jun. 2021. DOI: <https://doi.org/10.1108/LHT-11-2016-0122>.

OZMUTLU, S.; OZMUTLU, H. C.; SPINK, A. Topic analysis and identification of queries. In: JANSEN, B. J.; SPINK, A.; TAKSA, I. (ed.). *Handbook of research on web log analysis*. Hershey, PA: IGI Global, 2008. Chapter 17, p. 345-358. Disponível em: <https://www.igi-global.com/chapter/topic-analysis-identification-queries/22009>. Acesso em: 20 jul. 2020.

PANDAS. *Python Data Analysis Library*. 2021. Disponível em: <https://pandas.pydata.org/>. Acesso em: 3 jun. 2021.

PARK, M.; LEE, T. Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, v. 31, n. 1, p. 123-140, 2013. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/07378831311303976/full/html>. Acesso em: 21 jun. 2020. DOI: <https://doi.org/10.1108/07378831311303976>.

PEÑA, R.; BAEZA-YATES, R.; RODRÍGUEZ MUÑOZ, J. V. *Gestión digital de la información: de bits a bibliotecas digitales y a la web*. Madrid, 2002.

PENNIMAN, W. D. Historic perspective of log analysis. In: JANSEN, B. J.; SPINK, A.; TAKSA, I. (ed.). *Handbook of research on web log analysis*. Hershey, PA: IGI Global, 2008. Chapter 2, p. 18-38. Disponível em: <https://www.igi-global.com/gateway/chapter/21994#pnlRecommendationForm>. Acesso em: 20 jul. 2020. DOI: 10.4018/978-1-59904-974-8.ch002.

PETERS, T. A. The history and development of transaction log analysis. *Library Hi Tech*, v. 11, n. 2, p. 41-66, 1993. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb047884/full/html>. Acesso em: 20 jul. 2020. DOI: <https://doi.org/10.1108/eb047884>.

PONTES, A. M. *OPAC como recurso para a gestão da informação no contexto da biblioteca central da UFPB*. 2006. Trabalho de Conclusão de Curso (Especialização em Gestão de Unidades de Informação) – Departamento de Biblioteconomia e Documentação, Centro de Ciências Sociais Aplicadas, Universidade Federal da Paraíba, João Pessoa, 2006.

PÖTTKER, L. M. V.; FERNEDA, E.; MOREIRO-GONZÁLEZ, J. A. Mapeamento relacional entre padrões de metadados educacionais. *Perspectivas em Ciência da Informação*, v. 23, n. 3, p. 25-38, 2018. Disponível em: <https://doi.org/10.1590/1981-5344/2843> . Acesso em: 25 jul. 2020.

RAFIQUE, A.; AMEEN, K.; ARSHAD, A. Use patterns of e-journals among the science community: a transaction log analysis. *The Electronic Library*, v. 37, n. 4, p. 740-759, 2019. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/EL-03-2019-0073/full/html>. Acesso em: 13 jul. 2020. DOI: <https://doi.org/10.1108/EL-03-2019-0073>.

RAUTENBERG, S.; SOUZA, L.; KELNIAR, J. P. Uma proposta de tesauro da Ciência da Informação: organização de elementos de conhecimento com SKOS. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, v. 26, 2021. Disponível em: <https://brapci.inf.br/index.php/res/v/160602>. Acesso em: 23 out. 2021.

REIS, C. M. B. *Otimizando a recuperação da informação em Repositórios Institucionais*. 2008. 32 f. Trabalho de Conclusão de Curso (Especialização em Informação Científica e Tecnológica). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz, Rio de Janeiro, 2008. Disponível em: <https://www.arca.fiocruz.br/handle/icict/2950> . Acesso em: 27 out. 2020.

RICE, R. E.; BORGMAN, C. L. The use of computer-monitored data in information science and communication research. *Journal of the American Society for Information Science*, v. 34, n. 4, p. 247-256, 1983. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630340404>. Acesso em: 12 jun. 2020. DOI: <https://doi.org/10.1002/asi.4630340404>.

RIEGER, O. Y. Select for success: key principles in assessing repository models. *D-lib Magazine*, v. 13, n. 7/8, p. 1-8, 2007. Disponível em: <http://www.dlib.org/dlib/july07/rieger/07rieger.html>. Acesso em: 25 de set. 2020.

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. *Ciência da informação*, v. 20, n. 2, 1991. Disponível em: <https://revista.ibict.br/ciinf/article/view/348>. Acesso em: 12 ago. 2020.

ROWLEY, J. Fundamentos da recuperação da informação. In: ROWLEY, J. *A biblioteca eletrônica*. Tradução de Antônio Agenor Briquet de Lemos. 2. ed. Brasília, DF: Briquet de Lemos, 2002.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P.B. Definições dos enfoques quantitativo e qualitativo, suas semelhanças e diferenças. In: SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P.B. *Metodologia de Pesquisa*. 5.ed. Porto Alegre: Penso, 2013. p. 30-48. Disponível em: <https://statics-shoptime.b2w.io/sherlock/books/firstChapter/116717564.pdf>. Acesso em: 15 jul. 2020.

SANCHEZ, F. A.; VECHIATO, F. L.; VIDOTTI, S. A. B. G. Recomendações de encontrabilidade da informação para repositórios institucionais. In: SEMINÁRIO EM CIÊNCIA DA INFORMAÇÃO - SECIN, 8., 2019, Londrina. *Informação social no contexto da Ciência da Informação*. Londrina: Universidade Estadual de Londrina, 2019. Disponível

em: <http://www.uel.br/eventos/cinf/index.php/secin2019/secin2019/paper/viewFile/577/406>. Acesso em: 13 jun. 2020.

SANDORE, B. Applying the results of transaction log analysis. *Library Hi Tech*, v. 11, n. 2, p. 88-97, 1993. Disponível em: <https://doi.org/10.1108/eb047887>. Acesso em: 16 jul. 2020.

SANDORE, B. *et al.* A manifesto: regarding the future of transaction log analysis. *Library Hi Tech*, v. 11, n. 2, p. 105-106, 1993. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb047889/full/html>. Acesso em: 14 jul. 2020. DOI: <https://doi.org/10.1108/eb047889>.

SANTOS, J. C. F.; MOREIRA, W.; FUJITA, M. S. L. TESAURO UNESP: integração do registro de autoridade para o TemaTres. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 19., 2018, Londrina. *Sujeito informacional e as perspectivas atuais em Ciência da Informação*. Disponível em: http://enancib.marilia.unesp.br/index.php/XIX_ENANCIB/xixenancib/paper/view/1495/1433. Acesso em: 5 out. 2020.

SANTOS, R. F.; NEVES, D. A. B. Práticas de indexação em repositórios digitais de acesso aberto: análise do metadado assunto do repositório institucional da UFRN. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 19., 2018, Londrina. *Sujeito informacional e as perspectivas atuais em Ciência da Informação*. Disponível em: http://enancib.marilia.unesp.br/index.php/XIX_ENANCIB/xixenancib/paper/view/1117/1417. Acesso em: 5 out. 2020.

SCHULTHEIB, S. *et al.* Known-item searches and search tactics in library search systems: results from four transaction log analysis studies. *The Journal of Academic Librarianship*, v. 46, n. 5, art. 102202, 2020. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0099133320301051>. Acesso em: 13 nov. 2020. DOI: <https://doi.org/10.1016/j.acalib.2020.102202>.

SHIRI, A. Exploration of interdisciplinarity in nanotechnology queries: the use of Transaction Log Analysis and Thesauri. *Advances in Classification Research Online*, 2009. Disponível em: <https://journals.lib.washington.edu/index.php/acro/article/view/12883/11379>. Acesso em: 15 jul. 2020. DOI: 10.7152/acro.v20i1.12883. v. 20, n. 1, 2009.

SIEBRA, S. A.; OLIVEIRA, J. N. N.; MARCELINO, C. S. Avaliação do acesso e visualização da informação em Repositórios Institucionais. *Informação & Tecnologia (ITEC)*, v.1, n. 2, p. 77-95, jul./dez, 2014. Disponível em: <https://periodicos.ufpb.br/index.php/itec/article/view/21535/12662>. Acesso em: 20 jan. 2021.

SOERGEL, D. *Organizing information: principles of data base and retrieval systems*. New York: Academic Press, 1985.

SOUZA, S. S.; SOUZA, J. E. Repositório institucional como ferramenta de preservação da memória e disseminação do conhecimento no ensino público superior sergipano. *Revista*

Fontes Documentais, v. 1, n. 1, p. 94-110, 2018. Disponível em:
<http://hdl.handle.net/20.500.11959/brapci/134698>. Acesso em: 13 jun.2020.

SMIT, J. W. Vocabulário controlado e controle de vocabulário em arquivos. *Acervo*, v. 31, n. 3, p. 46-56, 2018. Disponível em:
<http://revista.arquivonacional.gov.br/index.php/revistaacervo/article/view/956/982>. Acesso em: 12 set. 2020.

TAKSA, I.; SPINK, A.; JANSEN, B. J. Web log analysis: diversity of research methodologies. In: JANSEN, B. J.; SPINK, A.; TAKSA, I. (ed.). *Handbook of research on web log analysis*. Hershey, PA: IGI Global, 2008. Chapter 25, p. 506-522. Disponível em:
<https://www.igi-global.com/chapter/web-log-analysis/22017>. Acesso em: 22 jul. 2020.
 DOI: 10.4018/978-1-59904-974-8.ch025

TARTAROTTI, R. C. D.; DAL'EVEDOVE, P. R. Avaliação da indexação em repositórios institucionais brasileiros: uma análise comparada entre USP, UNESP e UNICAMP. *Revista Ibero-Americana de Ciência da Informação*, v. 14, n. 2, p. 583-599, 2021. Disponível em:
<https://doi.org/10.26512/rici.v14.n2.2021.37609>. Acesso em: 23 jan. 2022.

TOLLE, J. E. Transaction log analysis online catalogs. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 6., 1983, New York. *SIGIR'83 Proceedings...* New York, NY, USA: ACM, 1983. p. 147-160. Disponível em: <https://dl.acm.org/doi/10.1145/511793.511816>. Acesso em 21 jul. 2020. DOI: <https://doi.org/10.1145/511793.511816>.

UNIVERSIDADE ESTADUAL PAULISTA. Grupo Gestor do Repositório Institucional UNESP. *Regulamento Interno do Repositório Institucional UNESP*. São Paulo: Universidade Estadual Paulista (UNESP), 2016. Disponível em: <http://hdl.handle.net/11449/144653>. Acesso em: 6 jun. 2020.

VÁLLEZ, M. *et al.* Updating controlled vocabularies by analysing query logs. *Online Information Review*, v. 39, n. 7, p. 870-884, 2015. Disponível em: <https://bityli.com/hiatE>. Acesso em: 7 set. 2020. DOI: <https://doi.org/10.1108/OIR-06-2015-0180>.

VIDOTTI, S. A. B. G. *et al.* Coleta automática para povoamento de repositórios digitais: conversão de registros utilizando XSLT. *Tendências da Pesquisa Brasileira em Ciência da Informação*, v. 9, n. 2, p. 1-21, 2016. Disponível em:
<https://revistas.ancib.org/index.php/tpbci/article/view/390/390> Acesso em: 10 set. 2020.

VIDOTTI, S.A.B.G. *et al.* Metadata reuse to populate an institutional repository: procedures applied in Unesp institutional repository. In: INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS – DCMI, 2015, São Paulo. *DC-2015 – The São Paulo Proceedings*. Washington - EUA: DCMI - ASIS&T - Association for Information Science and Technology, 2015. v. 1. p. 234-235. Disponível em:
<https://dcevents.dublincore.org/public/dc-docs/2015-Master.pdf>. Acesso em: 23 set. 2020.

VIEIRA, S. B. Indexação automática e manual: revisão da literatura. *Ciência da Informação*, v. 17, n. 1, p. 43-57, jan./jul 1988. Disponível em:
<https://doi.org/10.18225/ci.inf.v17i1.298>. Acesso em: 20 jan. 2022.

WEBOMETRICS. *Ranking Web of Repositories*. Disponível em: <https://repositories.webometrics.info/en/node/32>. Acesso em: 15 mar. 2022.

WHITE, R.W.; HORVITZ, E. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *Journal of the American Medical Informatics Association*, v. 21, n. 1, p. 49-55, 2014. Disponível em: <https://academic.oup.com/jamia/article/21/1/49/695622>. Acesso em: 7 set. 2020.

WU, S.; FAN, Y. Music literature indexing: comparing users' free-text queries and controlled vocabularies. *Cataloging & classification quarterly*, v. 56, n. 4, p. 330-353, 2018. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/01639374.2017.1422582>. Acesso em: 23 jul. 2020. DOI: <https://doi.org/10.1080/01639374.2017.1422582>.

ZHANG, H.; HUANG, W.; YANG, J. Design and implementation of real-time log analysis system of map world platform. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, v. 42, n. 4, p. 745-749, 2018. Disponível em: <https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-4/745/2018/>. Acesso em: 24 jul. 2020. DOI: <https://doi.org/10.5194/isprs-archives-XLII-4-745-2018>.

ZHOU, X.; ZHANG, P.; WANG, J. Examining task relationships in multitasking consumer search sessions: a query log analysis. *Proceedings of the Association for Information Science and Technology*, v. 53, n. 1, p. 1-5, 2016. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pa2.2016.14505301102>. Acesso em: 13 jul. 2020.

ZINK, S. D. Monitoring user search success through transaction log analysis: the Wolf PAC example. *Reference Services Review*, v. 19, n. 1, p. 49-56, 1991. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb049114/full/html>. Acesso em: 17 jul. 2020. DOI: <https://doi.org/10.1108/eb049114>.