UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE CIÊNCIAS BÁSICAS DA SAÚDE
DEPARTAMENTO DE BIOQUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EDUCAÇÃO EM CIÊNCIAS: QUÍMICA DA VIDA E SAÚDE

CHARLES HENRIQUE DE ARAÚJO

ESTUDO SOBRE O IMPACTO DA ADIÇÃO DE VOCABULÁRIOS ESTRUTURADOS DA ÁREA DE CIÊNCIAS DA SAÚDE NO CURRÍCULO LATTES

Porto Alegre
2016

CHARLES HENRIQUE DE ARAÚJO

ESTUDO SOBRE O IMPACTO DA ADIÇÃO DE VOCABULÁRIOS ESTRUTURADOS DA ÁREA DE CIÊNCIAS DA SAÚDE NO CURRÍCULO LATTES

Dissertação de mestrado apresentada ao Programa de Pós-Graduação Educação em Ciências: Química da Vida e Saúde, do Instituto de Ciências Básicas da Saúde, do Departamento de Bioquímica da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Mestre em Educação em Ciências.

Orientadora: Prof.ª Dr.ª Angela Terezinha de Souza Wyse

Banca Examinadora:

Prof.ª Dr.ª Luciana Calabro – UFRGS
Prof.ª Dr.ª Ana Maria Mielniczuk de Moura – UFRGS
Prof. Dr. Ivan Rocha Neto - UCB

Porto Alegre
2016

CIP - Catalogação na Publicação

de Araújo, Charles Henrique
Estudo sobre o Impacto da Adição de Vocabulários Estruturados da Área de Ciências da Saúde no Currículo Lattes / Charles Henrique de Araújo. -- 2016.
78 f.

Orientadora: Angela Terezinha de Souza Wyse.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Ciências Básicas da Saúde, Programa de Pós-Graduação em Educação em Ciências: Química da Vida e Saúde, Porto Alegre, BR-RS, 2016.

1. Expansão de consultas. 2. Mineração de dados. 3. Sistemas de Recomendação. I. Wyse, Angela Terezinha de Souza, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Dedicatória

Dedico este estudo aos meus pais, que sempre estiveram presentes em minha vida, à minha esposa e filhos pelo incentivo, paciência e compreensão da minha ausência durante o período de estudo.

AGRADECIMENTOS

Agradeço a Deus por estar presente em minha vida todos os dias.

À minha orientadora Professora Angela Wyse, pela positividade, paciência, confiança e bons fluidos durante o estudo.

Ao Programa de Pós-Graduação de Educação em Ciências da Universidade Federal do Rio Grande do Sul.

Aos professores do PPG, que se dispuseram a ministrar suas aulas em Brasília.

Ao CNPq, por incentivar e dar a oportunidade de aperfeiçoar os meus conhecimentos para melhor execução das atividades do dia a dia.

À minha família, pela compreensão, paciência e incentivo nessa jornada.

Aos colegas da CGERH, sempre pacientes, disponíveis e atenciosos.

Aos colegas da CGETI, Coordenadores, Chefes de Serviços, demais servidores e colaboradores que apoiaram o meu estudo.

Ao colega Paulo Henrique Assis Santana, pelo grande apoio na avaliação dos resultados dos cálculos obtidos.

RESUMO

A busca de informações em bases de dados de instituições que possuem grande volume de dados necessita cada vez mais de processos mais eficientes para realização dessa tarefa. Problemas de grafia, idioma, sinonímia, abreviação de termos e a falta de padronização dos termos, tanto nos argumentos de busca, quanto na indexação dos documentos, interferem diretamente nos resultados. Diante disso, este estudo teve como objetivo avaliar o impacto da adição de vocabulários estruturados da área de Ciências da Saúde no Currículo Lattes, na recuperação de perfis similares de pesquisadores das áreas de Ciências Biológicas e Ciências da Saúde, utilizando técnicas de mineração de dados, expansão de consultas, modelos vetoriais de consultas e utilização de algoritmo de trigramas. Foram realizados cruzamentos de informações entre as palavras-chaves de artigos publicados registrados no Currículo Lattes e as informações contidas no *Medical Subject Headings* (MeSH) e nos Descritores em Ciências da Saúde (DeCS), bem como comparações entre os resultados das consultas, utilizando as palavras-chaves originais e adicionando-lhes os termos resultantes do processo de expansão de consultas. Os resultados mostram que a metodologia adotada neste estudo pode incrementar qualitativamente o universo de perfis recuperados, podendo dessa forma contribuir para a melhoria dos Sistemas de Informações do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

Palavras-chaves: Expansão de consultas, Mineração de dados, Sistemas de Recomendação.

ABSTRACT

Information retrieval in large databases need increasingly more efficient ways for accomplishing this task. There are many problems, like spelling, language, synonym, acronyms, lack of standardization of terms, both in the search arguments, as in the indexing of documents. They directly interfere in the results. Thus, this study aimed to evaluate the impact of the addition of structured vocabularies of Health Sciences area in Lattes Database, in the recovery of similar profiles of researchers that work in Biological Sciences and Health Sciences, using Query Expansion, Data Mining procedures, Vector Models and Trigram Phrase Matching algorithm. Crosschecking keywords of articles registered in Lattes Database and *Medical Subject Headings* (MeSH) and Health Sciences Descriptors (DeCS) terms, as well as comparisons between the results of queries using the original keywords and adding them to query expansion terms. The results show that the methodology used in this study can qualitatively increase the set of recovered profiles, contributing to the improvement of CNPq Information Systems.

Keywords: Query Expansion, Data Mining, Recommendation Systems**.**

LISTA DE ABREVIATURAS

BVS - Biblioteca Virtual em Saúde
CAPES - Coordenação de Acompanhamento de Pessoal de Nível Superior
CGEE - Centro de Gestão e Estudos Estratégicos
CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico
COOTI - Coordenação de Operação de Tecnologia da Informação
CPLP - Comunidade dos Países de Língua Portuguesa
DOI - *Digital Object Identifier*
DeCS - Descritores em Ciência da Saúde
FNDCT - Fundo Nacional de Desenvolvimento Científico e Tecnológico
IBICT - Instituto Brasileiro de Informação em Ciência e Tecnologia
INPI - Instituto Nacional da Propriedade Industrial
LMPL- Linguagem de Marcação da Plataforma Lattes
MAPA - Ministério da Agricultura, Pecuária e Abastecimento
MCTI - Ministério da Ciência, Tecnologia e Inovação
MESH - *Medical Subject Headings*
NLM - *U.S National Library of Medicine*
PED - Programa Estratégico de Desenvolvimento
PDTI - Plano Diretor de Tecnologia da Informação
PICC - Plataforma Integrada Carlos Chagas
SciELO - *Scientific Electronic Library Online*
SGBDR - Sistema Gerenciador de Banco de Dados Relacional
SISP - Sistema de Administração dos Recursos de Tecnologia da Informação
XML - *Extensible Markup Language*
W3C - World Wide Web Consortium

SUMÁRIO

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ILUSTRAÇÕES

1 INTRODUÇÃO

Um grande problema enfrentado pelos sistemas de buscas é a adaptação da linguagem natural para a linguagem de consulta. No meio acadêmico, quem escreve os artigos científicos associa palavras-chaves à produção que nem sempre são as mesmas palavras fornecidas por um usuário quando este quer realizar uma consulta. Um exemplo clássico para o problema é a busca pelo temo "febre amarela". Quem escreve sobre o assunto e não indexa as suas produções utilizando o termo traduzido para o inglês, dificilmente seria encontrado em um resultado de uma consulta onde o usuário tenha fornecido como parâmetro de busca o termo "*yellow fever*". De forma mais rara ainda, o resultado traria resultados que contenham termos associados à febre amarela, como por exemplo, "malária". Além disso, problemas de grafia, idioma, sinonímia, abreviação de termos e a falta de padronização dos dados informados pelos usuários interferem diretamente nos resultados esperados.

Na literatura podem ser encontradas diversas abordagens sobre o problema em questão. Entre elas existe a técnica denominada expansão de consultas, que consiste em agregar ao conjunto inicial de termos uma série de outros novos termos relacionados, e permite fornecer um maior conjunto de resultados a quem realiza a consulta, dando a possibilidade de se encontrar resultados que de outra forma não seriam obtidos utilizando-se apenas os termos diretos informados como parâmetros de busca**.**

1.1 O Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq

O CNPq, agência do Ministério da Ciência, Tecnologia e Inovação (MCTI), criado em 1951 pela Lei nº 1.310 de 15/01/1951, tem como principais atribuições fomentar a pesquisa científica e tecnológica e incentivar a formação de pesquisadores brasileiros. A lei de criação do Conselho estabelecia como suas finalidades promover e estimular o desenvolvimento da investigação científica e tecnológica, mediante a concessão de recursos para pesquisa, formação de pesquisadores e técnicos,

cooperação com as universidades brasileiras e intercâmbio com instituições estrangeiras. A missão do CNPq era ampla, uma espécie de "estado-maior da ciência, da técnica e da indústria, capaz de traçar rumos seguros aos trabalhos de pesquisas" científicas e tecnológicas do país, desenvolvendo-os e coordenando-os de modo sistemático (CNPq 2016).

O CNPq desempenha papel primordial na formulação e condução das políticas de ciência, tecnologia e inovação. Tem como missão "Fomentar a Ciência, Tecnologia e Inovação e atuar na formulação de suas políticas, contribuindo para o avanço das fronteiras do conhecimento, o desenvolvimento sustentável e a soberania nacional", e visão "ser uma instituição de reconhecida excelência na promoção da Ciência, da Tecnologia e da Inovação como elementos centrais do pleno desenvolvimento da nação brasileira".

Nos anos 50, o CNPq concentrou-se no fomento à concessão de bolsas de estudo para a formação de pesquisadores, apoio à realização de reuniões científicas e intercâmbio científico. Nesse mesmo período surgiram as primeiras modalidades de fomento à bolsa de pesquisa, como a de iniciação científica IC, dentre outras. Influenciado pelo pós-guerra, apoiou-se bolsas nos campos das ciências básicas, ligadas a Física com foco na energia atômica, bem como a aquisição de equipamentos aos institutos aplicados existentes no País. Nesse período o CNPq também apoiou a área de ciências biológicas, que estava entre as mais desenvolvidas do país, e o processo de industrialização brasileira, caracterizado na época pela ênfase na produção de bens de consumo duráveis, importação de bens de capital e aquisição de tecnologia estrangeira (BRASIL, 2016c).

Na década de 60, os governos estaduais e federal deram maior importância à pesquisa cientifica. Destaca-se a criação da FAPESP, no estado de São Paulo e no âmbito nacional, a criação do Fundo de Desenvolvimento Técnico Científico (FUNTEC) (BRASIL, 2016d). Nessa época, o CNPq passou a formular as políticas de C&T nacional em parceira com outras instituições. A alteração dada pela Lei n° 4.533 de 8 de dezembro de 1964, incorporava o desejo do governo militar na formação de profissionais especializados para a indústria, em desenvolvimento, também no

fortalecimento do aparato técnico-científico, vislumbrado pelo regime (BRASIL, 2015f). Em 1965 foi institucionalizado o ensino de mestrado e doutorado no Brasil com a regulamentação e o estabelecimento de conceitos e bases legais para a pós-graduação publicados no Parecer nº 977/65, conhecido como o **Parecer Sucupira.** Esta ficou marcada pela adoção de uma política de ciência e tecnológica no Programa Estratégico de Desenvolvimento (PED), ação que já objetivava o apoio financeiro a pesquisas operacionalizadas pelo CNPq, incrementadas pela criação também do Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT). Os movimentos culminaram na implantação de infraestrutura de pesquisa necessários ao desenvolvimento das décadas seguintes (BRASIL,2016d).

Nos anos 70, o CNPq passou a ser o órgão centralizador do Sistema Nacional de Desenvolvimento Científico e Tecnológico (SNDCT), e tinha como um dos objetivos aumentar o incentivo à pesquisa no setor privado e nas sociedades de economias mistas. Outra alteração importante dada pela lei n° 6129 de 6 de novembro de 1974, na época, foi a transformação do então “Conselho Nacional de Pesquisas” em “Conselho Nacional de Desenvolvimento Científico e Tecnológico”, com a preservação da sigla CNPq (BRASIL,2016e).

Em 1985, O CNPq passou a estar vinculado ao Ministério da Ciência e Tecnologia, e operou na descentralização do gerenciamento da C&T nacional, com a implementação dos Sistemas Estaduais de Ciência e Tecnologia – SECT e a ênfase às ciências humanas/sociais aplicadas, com a introdução de novas áreas de conhecimento nas atividades de fomento, dentre outras ações de impacto (BRASIL,2016f).

Nos anos 90, marcado por intensas atividades e transferências de funções ao MCT, o CNPq foi redirecionado para a missão de promover o desenvolvimento científico e tecnológico e executar pesquisas ao progresso social, econômico e cultural do País. Na década de 90 o CNPq também lançou importantes instrumentos para a gestão do fomento nacional, como a Plataforma Lattes e o Diretório do Grupo de Pesquisas, fundamentais para avaliação, acompanhamento e direcionamento de políticas e diretrizes de incentivo à pesquisa (BRASIL, 2016g).

A partir de 2000, o CNPq incrementou as ações com parcerias internacionais, criação de novos polos de educação superior pelo governo federal, que demandaram novos recursos e a indução para o intercâmbio de estudantes brasileiros com a criação do Programa Ciências sem Fronteiras.

Como órgão de fomento à pesquisa, conforme aprovado pela Portaria nº 816, de 17 de dezembro de 2002, que trata do regimento interno da instituição, compete ao CNPq participar na formulação, execução, acompanhamento, avaliação e difusão da Política Nacional de Ciência e Tecnologia, especialmente: promover e fomentar o desenvolvimento e a manutenção da pesquisa científica e tecnológica e a formação de recursos humanos qualificados para a pesquisa, em todas as áreas do conhecimento; promover e fomentar a pesquisa científica e tecnológica e capacitação de recursos humanos voltadas às questões de relevância econômica e social relacionadas às necessidades específicas de setores de importância nacional ou regional; promover e fomentar a inovação tecnológica; promover, implantar e manter mecanismos de coleta, análise, armazenamento, difusão e intercâmbio de dados e informações sobre o desenvolvimento da ciência e tecnologia; propor e aplicar normas e instrumentos de apoio e incentivo à realização de atividades de pesquisa e desenvolvimento, de difusão e absorção de conhecimentos científicos e tecnológicos; promover a realização de acordos, protocolos, convênios, programas e projetos de intercâmbio e transferência de tecnologia entre entidades públicas e privadas, nacionais e internacionais; apoiar e promover reuniões de natureza científica e tecnológica ou delas participar; promover e realizar estudos sobre o desenvolvimento científico e tecnológico; prestar serviços e assistência técnica em sua área de competência; prestar assistência na compra e importação de equipamentos e insumos para uso em atividades de pesquisa científica e tecnológica, em consonância com a legislação em vigor; e credenciar instituições para, nos termos da legislação pertinente, importar bens com benefícios fiscais destinados a atividades diretamente relacionadas com pesquisa científica e tecnológica (BRASIL, 2016i).

1.2 Tecnologia da Informação

De acordo com Robbins (2005), a Gestão do Conhecimento começa com a identificação de conhecimentos importantes para organização. Depois, é preciso desenvolver redes, de preferência informatizadas, e bancos de dados que tornem as informações rapidamente disponíveis para aqueles que delas necessitam. Nenhum sistema de Gestão de conhecimento será bem sucedido se a cultura da organização não estimular o seu compartilhamento.

Segundo Neto (2011), os Sistemas de Suporte à Decisão - SSD compreendem programas, bancos de dados e dispositivos utilizados para dar suporte à tomada de decisões e para resolver problemas. O foco desses sistemas está na tomada de decisões face aos problemas não estruturados ou semiestruturados. Embora sejam parcialmente destinados aos níveis mais elevados da gestão, podem ser aplicados em todos os ambientes das organizações. O mais importante é que precisam auxiliar à tomada de decisões, incluindo as relacionadas aos processos organizacionais. As principais funcionalidades dos SSD são as seguintes:

a) Obter e processar grandes volumes de dados de fontes diferentes, inclusive externos à organização e integrá-los de forma a facilitar a tomada de decisões;
b) Servir de base para elaboração de relatórios e apresentações em formatos adequados às necessidades e preferências dos tomadores de decisão;
c) Produzir informações que podem ser apresentadas tanto sob a forma textual quanto gráfica (tabelas, desenhos, gráficos, curvas de tendência, entre outras);
d) Realizar avaliações comparativas complexas utilizando softwares específicos integrados;
e) Executar simulações com avaliação adequada pode proporcionar apoio organizacional.

A computação revolucionou a pesquisa científica, sendo hoje reconhecida como o “terceiro pilar” a sustentar tal pesquisa, junto com os pilares da teoria e da experimentação (2005). Desta forma, ela permeia os avanços em todas as áreas do conhecimento. Novas formas de interação entre as ciências, em vários níveis e escalas, são mediadas pela Tecnologia da Informação, que é a simbiose da Ciência da Computação com diferentes domínios do conhecimento. Muitas das grandes descobertas científicas recentes são resultados do trabalho de equipes multidisciplinares que envolvem cientistas da Computação. Finalmente, ela é um componente indispensável para a implementação e o fortalecimento dos objetivos econômicos, tecnológicos e sociais de um país (2006).

No cumprimento de sua missão e na busca pela melhoria contínua de processos institucionais, finalísticos e de apoio, o CNPq apoia-se em recursos da Tecnologia da Informação – TI para fomentar, promover e facilitar a participação de pesquisadores na formação e consolidação das bases cientificas brasileiras. Cabe à TI apoiar o CNPq no cumprimento de suas competências sendo elemento fundamental na promoção, implantação e manutenção de mecanismos de coleta, análise, armazenamento, difusão e intercâmbio de dados e informações sobre o desenvolvimento da ciência e tecnologia e na difusão do conhecimentos científicos e tecnológicos.

A Sociedade Brasileira de Computação (SBC) congrega milhares de professores universitários e estudantes de Ciências da Computação e vem promovendo debates sobre os grandes desafios da computação no país. No relatório "A SBC e os Grandes Desafios da Computação no Brasil: 2006-2016", são citados cinco grandes desafios:

a) Gestão da Informação em grandes volumes de dados multimídia distribuídos;
b) Modelagem computacional de sistemas complexos artificiais, naturais e socioculturais e da interação homem-natureza;

c) Impactos para a área da computação da transição do silício para novas tecnologias;

d) Acesso participativo e universal do cidadão brasileiro ao conhecimento;

e) Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos[1].

1.2.1 Plano Diretor de Tecnologia da Informação - PDTI e Sistemas Estruturantes

O Plano Diretor de Tecnologia da Informação (PDTI), é um instrumento de diagnóstico, planejamento e gestão dos recursos e processos de Tecnologia da Informação de um órgão ou entidade para um determinado período. A partir de 2010, todas as contratações de bens e serviços devem estar vinculadas a elementos existentes no PDTI. Ou seja, se o órgão não elaborou e publicou seu PDTI, não poderá realizar contratação correlata à TI. O PDTI é o instrumento que permite nortear e acompanhar a atuação da área de TI, definindo estratégias e o plano de ação para implantá-las (BRASIL, 2016g).

As atividades fim do CNPq são apoiadas por dois grandes sistemas de informações estruturantes: a Plataforma Integrada Carlos Chagas (PICC) e a Plataforma Lattes.

A PICC é responsável pela automação dos fluxos de trabalho relacionados à execução das atividades de fomento da instituição, acompanhando as etapas de proposição, análise e lançamento de ações de fomento, recebimento de propostas e solicitações de bolsas e auxílios por pesquisadores e estudantes, análise e julgamento dessas solicitações por consultores, analistas de C&T e comitês de assessoramento, divulgação dos resultados finais, contratação dos recursos concedidos, acompanhamento da execução dos projetos, prestação de contas ao final dos projetos e bolsas e, caso necessário, a cobrança e tomada de contas especial de processos cuja prestação de contas ou análise do relatório técnico final foi rejeitada pela

[1] característica de sistemas que coexistem com o homem de forma harmoniosa, de forma a ajuda-lo na realização de tarefas cotidianas, e que estejam em qualquer lugar, embutidos nos mais diversos dispositivos

Instituição. Entre 2006 a 2014, a PICC recebeu mais de 650 mil propostas e pedidos eletrônicos, para os quais foram emitidos cerca de 2,5 milhões de pareceres eletrônicos de consultores ad hoc, membros de comitês julgadores e analistas de C&T.

A Plataforma Lattes é um conjunto de sistemas de informação, bases de dados, portais Web voltados para a gestão de ciência e tecnologia composta pela integração do Diretório de Grupos de Pesquisa, Diretório de Instituições e Currículo Lattes. Sua dimensão atual se estende não só às ações de planejamento, gestão e operacionalização do fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa. Além disso, se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formulação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação (BRASIL, 2016h).

Essa Plataforma conta atualmente com informações de mais de 35 mil grupos de pesquisa em atividade no país e 4 milhões de currículos de pesquisadores, professores, orientadores e estudantes, e é utilizada por diversas instituições públicas e privadas, além do próprio CNPq. Atualmente, cerca de 200 instituições, universidades, institutos de pesquisa, fundações estaduais de amparo à pesquisa e órgãos federais, como os Ministérios da Saúde, Ministério da Educação e Ministério do Meio Ambiente, possuem acesso aos serviços *web* da Plataforma Lattes, que permitem a extração de conteúdos existentes nos currículos e grupos de pesquisa em atividade no país, informações essas que são utilizadas pelas instituições para a avaliação de solicitações de recursos e produtividade de seus pesquisadores e professores.

O reconhecimento da importância da Plataforma Lattes no cenário de C & T lhe conferiu o Prêmio e-Gov[2] 2004 na categoria Governo para Cidadão (G2C). Criado em 2002, o Prêmio e-Gov é uma iniciativa da Associação Brasileira de Entidades Estaduais de Tecnologia da Informação e Comunicação (Abep). Este evento é realizado anualmente e tem como objetivo reconhecer e incentivar o desenvolvimento de projetos e soluções de governo eletrônico na administração pública e divulgar ações que, com o

[2] http://www.premio-e.gov.br/

uso da tecnologia da informação, visem a modernizar a gestão pública em benefício da população.

Lane (2010) afirma em seu artigo publicado na Nature que os países devem dar prioridade à qualidade, transparência e modernização dos bancos de dados acadêmicos a fim de melhorar o desempenho científico mundial e cita a Plataforma Lattes como exemplo de um banco de dados completo e altamente qualificado.

1.2.2 Currículo Lattes

O Sistema de Currículos Lattes, desenvolvido pelo CNPq e disponibilizado em meados de 1999, constitui-se um grande repositório de informações sobre pesquisadores e sua produção científica e tecnológica. Foi concebido para padronizar e consolidar informações curriculares de pesquisadores interessados em obter recursos junto ao CNPq e demais agências de fomento, e unificar uma série de outros diversos formulários de sistemas de currículos que eram mantidos pelas instituições de ensino superior. Este sistema permitiu uma padronização e racionalização no processo de cadastramento, armazenamento e disponibilização de dados curriculares.

Na data de lançamento do sistema, a base de dados contava inicialmente com cerca de 35 mil currículos cadastrados. Atualmente, este número alcança a marca de mais de 4,7 milhões de currículos, com e média de atualizações diárias de cerca de 17 mil currículos[3].

O Currículo Lattes é composto por um módulo de cadastramento, acessado pela Internet através de navegador *web*, sem que haja necessidade de espaço para armazenamento local dos dados registrados, ou instalação de programas, no qual o pesquisador registra os seus dados pessoais, informações para contato, formação acadêmica, prêmios recebidos, proficiência em idiomas, áreas de atuação, atividades profissionais, produções bibliográficas, produções técnicas, produções artísticas, propriedade intelectual, informações sobre eventos, teses orientadas e em andamento,

[3] Dados obtidos da Plataforma Lattes em 26/07/2016

e informações sobre os artigos publicados e disponibilizados em outras bases de dados bibliográficas.

A aplicação Busca de Currículos, baseada em índices textuais, cuja base de dados é resultado de um processo de transformação e consolidação dos dados registrados no formulário de atualização de currículos, e cuja estrutura de dados está baseada em XML, completa a composição do sistema. Na busca de currículos é possível encontrar pesquisadores através de nome, ou parte dele e por assunto (título da produção, palavras-chaves e texto inicial informado pelo pesquisador) e aplicar filtros acerca da formação acadêmica, atuação profissional, nível e categoria de bolsa recebida do CNPq, entre outros.

De acordo com os dados fornecidos pelo site Painel Lattes, cuja atualização dos dados ocorreu em maio de 2016, cerca de 6,48% dos currículos são de doutores, 10,65% são de mestres, e 34,10% são de especialistas e graduados.

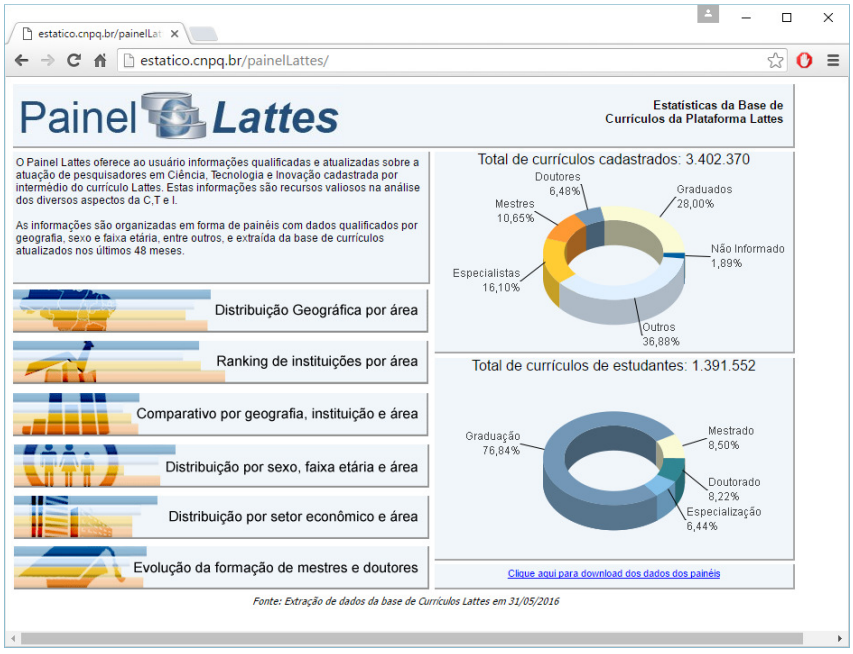


Figura 1 - Painel Lattes, estatísticas da Base de Currículos

As informações registradas no Lattes são declaratórias, e a fim de dar maior confiabilidade a essas informações, o CNPq vêm mantendo acordos de cooperação e integração com outras instituições de governo e outras bases de dados.

Quando o usuário envia o seu Currículo Lattes ao CNPq, os dados pessoais são verificados junto à Secretaria de Receita Federal; durante o preenchimento dos formulários, informações sobre produções bibliográficas são obtidas junto ao Crossref[4], dados sobre o depósito de patentes são obtidas junto ao Instituto Nacional de Propriedade Intelectual (INPI), informações sobre cultivares registradas e protegidas são obtidos junto ao Ministério da Agricultura, Pecuária e Abastecimento (MAPA).

Entre as necessidades registradas no PDTI, está a ampliação dos mecanismos de certificação de dados do Currículo Lattes. As ações previstas para atender essa necessidade são implementar:

a) Validação de dados de teses e dissertações a partir da integração com a base de dados de Teses e Dissertações (BDTD) do IBICT/Ministério da Ciência Tecnologia e Inovação;

b) Funcionalidades para a certificação de currículos de estrangeiros;

c) Validação dos dados de titulação a partir da integração com a base de dados do Coleta[5] da CAPES/MEC;

d) Funcionalidade que permita a certificação interna de publicações que estejam com informações incompletas ou não tenham sido registradas na base de dados do Crossref;

e) Validação de registros de cultivares registrados e protegidos a partir da integração com bases de dados do MAPA;

f) Validação de registros de patentes a partir da integração com bases de patentes internacionais;

[4] http://crossref.org/
[5] https://sucupira.capes.gov.br

g) Funcionalidade para permitir a validação de ISSNs não constantes da tabela de periódicos a partir da integração com a base *da ISSN International Centre.*

Está em andamento cooperação com o Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT, para permitir integração entre o Lattes e a Biblioteca Digital de Teses e Dissertações. Com essa integração será possível recuperar os metadados das teses e dissertações dos alunos o que permitirá o preenchimento automático dessas informações no currículo, além de permitir um acesso direto ao documento que está depositado no Banco de Teses do IBICT.

Outros mecanismos de certificação presentes no sistema de currículo permitem que os próprios pares deem o seu aval sobre a informação registrada no CV. Existe cruzamento de informações entre os dados da formação acadêmica e teses em andamento e orientadas, bem como há validação de dados informados no módulo de projetos, onde é solicitado aos coordenadores citados a certificação sobre a participação de um determinado membro de equipe, ou ainda, envio de mensagem eletrônica ao representante de instituição para certificar a informação sobre a parceria na execução de um determinado projeto.

1.3.3 Integração com outras bases e extração de dados

As integrações entre o Currículo Lattes e outras bases bibliográficas foram iniciadas em 2000, com a Bireme[6]. Foi estabelecido acordo de cooperação para realizar processamento periódico de arquivos do *Scientific Electronic Library Online* (SciELO), a fim de disponibilizar na página do currículo um ícone que remetesse às produções do pesquisador que estão naquela base bibliográfica, e em contrapartida, fornecer àquela instituição o endereço de acesso ao currículo dos autores das produções. Em seguida, a mesma metodologia para identificação de produções e coautores foi aplicada às bases de dados LILACS[7] e MEDLINE[8], conforme descrito por Santana (2001). Em julho

[6] Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde
[7] http://lilacs.bvsalud.org/

de 2016, 206.936 pessoas registradas no Lattes possuíam *link* de acesso às produções do SciELO.

Um importante avanço obtido por este sistema de informação foi a possibilidade de obtenção de metadados de produções bibliográficas junto ao Crossref. O acordo estabelecido em julho de 2007 permitiu inicialmente, a partir do *Digital Object Identifier* (DOI)[9], recuperar as informações sobre o título do artigo, ano de sua produção, ISSN do periódico, título do periódico, páginas inicial e final, edição, e a relação de autores e coautores de artigos completos publicados em periódicos.

Essa ação possibilitou ao usuário realizar o preenchimento automático e correto das informações dos dados dessa produção bibliográfica, bastando para isso informar corretamente o DOI da produção bibliográfica. Além de facilitar o preenchimento dos formulários, a obtenção dos metadados de artigos permitiu realizar estatísticas mais precisas sobre a produção bibliográfica, bem como, ao usuário que consulta o currículo do pesquisador na internet, ensejar a possibilidade de ter acesso imediato ao endereço na internet do periódico e ao artigo publicado, a partir de sua visualização no CV Lattes.

Dessa forma, os artigos registrados no CV Lattes a partir do número DOI passaram a exibir um ícone no sistema de busca de currículos, o que além de facilitar o acesso à produção bibliográfica, agregou valor de certificação à informação. A disponibilização dessas informações no currículo teve como consequência um grande crescimento no interesse no preenchimento correto dessas informações no módulo de artigos do currículo.

Posteriormente, o Lattes passou a obter automaticamente metadados de trabalhos publicados, livros e capítulo de livros. Atualmente existem cerca de 1,3 milhões de registros de artigos, 54 mil registros de trabalhos publicado em anais de evento, 10 mil registros de livros e capítulos de livros no Lattes cujos metadados foram obtidos automaticamente junto ao Crossref.

A inserção da informação do DOI do CV Lattes permitiu ainda a redução de erros de digitação no preenchimento dos formulários e possibilitou que novos acordos

[8] http://goo.gl/7Ki2oh
[9] https://www.doi.org/

de cooperação fossem realizados com outras bases de dados bibliográficas. O DOI é informação fundamental para se obter o número de citações de um artigo, nas bases de dados do *Web of Science*[10], Scopus[11] e SciELO. Atualmente cerca de 570 mil artigos que estão registrados no Lattes possuem citações no *Web of Science*, 101 mil artigos possuem citação no Scopus, e 90 mil possuem citações no SciELO (dados de julho/2016).

O DOI é uma informação imprescindível para a realização deste estudo, tendo em vista permitir a obtenção de metadados de produção bibliográfica junto às diversas bases bibliográficas que estão disponíveis para acesso na Internet.

Adicionalmente, o Currículo Lattes exibe no sistema de busca o *Journal Citation Report* (JCR) do periódico no qual o artigo foi publicado. O JCR fornece recursos para avaliação de títulos de periódicos que compõem a base de dados da *Web of Science*. Serve para o pesquisador identificar títulos de periódicos para publicação de seus trabalhos e para o bibliotecário avaliar coleções de periódicos da biblioteca. Mas serve, sobretudo, para os aspectos bibliométricos ou de avaliação do periódico.

Além das utilidades citadas anteriormente, as informações obtidas junto às bases bibliográficas têm hoje papel fundamental no julgamento de propostas do fomento, pois essas informações são também utilizadas como indicadores da produção científica do pesquisador.

Ainda sobre integração entre bases de dados, está disponível no Currículo Lattes um módulo que permite ao pesquisador registrar informações sobre outras bases de dados bibliográficas onde suas produções podem ser encontradas. Quando selecionado a base de dados *Web of Science*, é permitido ao pesquisador informar o seu ResearchID[12]. Essa informação é utilizada tanto para importar os dados a partir daquela base de dados, quanto para disponibilizar na consulta de currículos um link direto para a página do pesquisador. Além dessas bases, foi disponibilizado no módulo de dados pessoais campo para o pesquisador informar o seu ORCID[13].

[10] https://www.webofknowledge.com/
[11] https://www.scopus.com/
[12] http://www.researcherid.com/
[13] http:// http://orcid.org/

Figura 2 - Integração do Currículo Lattes com bases de dados externas

O sucesso do currículo desenvolvido pelo CNPq fez com que diversos países se interessassem na adoção desse modelo. Em 2001 foi construído o currículo CVLAC, baseado na experiência do currículo brasileiro. Foram realizadas adaptações no sistema para e traduções dos textos para o espanhol para permitir que Chile, Colômbia, México, Peru e Costa Rica, Equador e Venezuela utilizassem uma versão customizada do sistema criado pelo CNPq. Além dos países latinos, uma versão do Currículo Lattes foi adaptada para ser utilizada em Moçambique, onde foi chamado de Currículo Dzowo.

Está registrado no PDTI a necessidade de aprimorar e ampliar o nível de internacionalização da Plataforma Lattes. Esta necessidade está alinhada ao Mapa Estratégico do CNPq[14]. As ações previstas para atendimento dessa necessidade são:

a) Implementar versão do Currículo Lattes em espanhol;

b) Revisar a tradução do Currículo Lattes para o idioma inglês;

c) Implementar módulos na Plataforma Lattes para atender as necessidades dos países da CPLP[15] (tem como objetivo permitir aos países da Comunidade dos Países de Língua Portuguesa, da qual o Brasil é membro, e também composta por Angola, Cabo Verde, Guiné Bissau, Guiné Equatorial,

[14] http://cnpq.br/planejamento-estrategico
[15] Comunidade dos Países de Língua Portuguesa. http://www.cplp.org/

Moçambique, Portugal, São Tomé e Príncipe e Timor Leste, utilizarem o Currículo Lattes e infraestrutura computacional disponível para este sistema de informação via internet, com portais de informações e mensagens de texto adaptados àqueles países;

d) Adaptar a funcionalidade de cadastro e manutenção de Currículo aos tipos de documento de identificação de outros países;

e) Implementar *webservices* para o espelhamento dos currículos Lattes de pesquisadores estrangeiros nas bases nacionais de seus países.

O Currículo Lattes se tornou um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país, e é hoje adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do País. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia (BRASIL, 2016h).

Em 2001, várias instituições de ensino superior solicitaram ao CNPq a abertura tecnológica da Plataforma Lattes. Em atendimento a esse pedido, foi criada a Linguagem de Marcação da Plataforma Lattes (LMPL), bem como o modelo DTD (*Data Type Definition*) do XML do Currículo Lattes para viabilizar a extração de dados a partir da Plataforma Lattes. O padrão que define os tipos de informação existentes em um documento XML é dado pelo seu respectivo DTD. Essa linguagem para modelagem de arquivos XML permite que aplicativos compreendam a estrutura em uma árvore definida no documento XML.

O Sistema de Currículos Lattes disponibiliza às instituições interessadas, *webservices* que permitem a extração parcial ou completa dos dados curriculares no formato XML. Estes dados normalmente são coletados com o objetivo de complementar as informações dos sistemas de informação internos, ou gerar estatísticas acerca da produção em ciência e tecnologia dessas instituições. A importação dos dados junto ao CNPq é um processo vantajoso tanto para a instituição, que coleta dados a partir de um repositório único de informações, quanto para o

pesquisador, que não tem a necessidade de preencher repetidamente vários cadastros. Segundo informações obtidas junto à Coordenação de Operação de Tecnologia da Informação (COOTI), atualmente 200 instituições brasileiras estão autorizadas a extrair informações a partir do Currículo Lattes.

1.3 Ciência da Informação

O processo de indexação de documentos, de acordo com a Ciência da Informação, é um processo de identificação e organização dos itens necessários à posterior recuperação das informações contidas em um documento. Neste processo, é importante a utilização de mecanismos que apoiem a indexação correta dos documentos. A utilização de listas de **cabeçalhos de assuntos**, **vocabulários estruturados**, como **tesauros** permitem dar maior precisão na recuperação da informação. Segundo Cesarino (1978), o objetivo da indexação é a transformação de uma expressão de Linguagem Natural em uma Linguagem de Indexação: a partir de um texto compreendido por palavras e frases organizados de forma linear, há uma condensação dessas informações em uma linguagem documentária, mas a eficiência dos sistemas de recuperação de informações dependem da estrutura e composição dessa linguagem.

De acordo com Leiva (2004) apud Gil Urdician (2004), os cabeçalhos de assuntos constituem um tipo de linguagem pré-coordenada, de estrutura associativa ou combinatória que consiste em uma lista alfabética de palavras ou expressões da linguagem natural, que, normalizadas, são capazes de representar os temas de que trata um documento e por meio dos quais se recuperam os documentos do acervo.

Conforme descrito por Castro (2001), vocabulários estruturados são coleções de termos, organizadas segundo uma metodologia na qual é possível especificar as relações entre conceitos com o propósito de facilitar o acesso à informação. Os vocabulários são usados como uma espécie de filtros entre a linguagem utilizada pelo autor e a terminologia da área e também podem ser considerados como assistentes de

pesquisa, ajudando o usuário a refinar, expandir ou enriquecer suas pesquisas, proporcionando resultados mais objetivos.

Vocabulários estruturados são necessários para descrever, organizar e prover acesso à informação. O uso de um vocabulário estruturado permite ao pesquisador recuperar a informação com o termo exato utilizado para descrever o conteúdo daquele documento científico. Os vocabulários estruturados funcionam também como mapas que guiam os usuários até a informação. Com a expansão da Internet, e o número de potenciais pontos de acesso à informação crescendo exponencialmente, os vocabulários podem ser úteis provendo termos consistentes que permitam ao usuário selecionar a informação que necessita a partir de uma vasta quantidade de dados.

De acordo com Laan (2012), um tesauro é um vocabulário controlado de um domínio específico do conhecimento, sendo que sua organização busca evidenciar as relações conceituais dessa área de especialidade. Dessa forma, entende-se que as unidades lexicais registradas nesses instrumentos de indexação deveriam ser constituídas em conformidade com a terminologia desse domínio específico do conhecimento. Além do que, essas unidades lexicais, mesmo pertencendo a uma linguagem de especialidade, comportam sinonímia e variação, evidenciando-se, assim, a importância do controle do vocabulário.

Segundo Lancaster (2004), os vocabulários controlados, incluindo os cabeçalhos de assuntos e os tesauros, são um tipo de linguagem de indexação na qual a terminologia está controlada.

Conforme descrito por Colepicolo (2006), a importância do tesauro se fundamenta no potencial de auxílio ao usuário da informação em encontrar documentos de acordo com suas necessidades ou expectativas. Diferentes usuários podem expressar suas necessidades de informação, ainda que seja a mesma, usando uma linguagem diferente, por exemplo, sinônimos, abreviações, acrônimos, etc. O tesauro surge como uma alternativa para resolver estes problemas característicos do uso da linguagem natural mapeando, por exemplo, os termos que representam o mesmo conceito, selecionando um termo apenas como padrão e os restantes como sinônimos, além de estabelecer relações entre os termos e outros a estes relacionados.

Segundo a norma de terminologia (Norma ISO – 1087), **termo** é a designação, por meio de uma unidade linguística, de um conceito definido em uma língua de especialidade. Para Vogel (2007) “um termo existe somente em seu próprio campo de aplicação, isto é, dentro do contexto de uma língua de especialidade, na qual seu significado adquire certa particularidade e assume uma carga semântica própria”. Segundo as teorias clássicas, dentro de uma especialidade, a um termo deve corresponder somente um conceito, ou seja, deve haver univocidade, meio de garantir uma comunicação mais rigorosa do que a língua geral. A polissemia pode representar sérios inconvenientes, como o risco de conflito entre os significados do interior de uma mesma especialidade ou dentro de suas subáreas.

1.3.1 A importância das palavras-chaves

As palavras-chaves são um instrumento de representação da informação contida nos documentos, úteis para a indexação em mecanismos de pesquisa ou categorização do texto. A sua utilização potencializa o acesso ao conteúdo dos documentos, para além da informação que é representada pelo título e resumo; traduz o pensamento dos autores, e mantém o contato com a realidade da prática quotidiana, acompanhando a evolução científica e tecnológica, que é refletida pelos documentos, conforme descrito por Miguéis et al (2013).

De acordo com Ercan (2007), palavras-chaves podem ser consideradas como um breve resumo de um texto. Por isso, é possível considerá-las como um conjunto de frases semanticamente cobrindo a maior parte do texto. Muito embora um resumo de um texto seja capaz de proporcionar mais informações sobre o texto, o resumo pode não ser adequado para algumas aplicações devido à complexa estrutura de frases. Palavras-chaves não são substitutos para sumarização, mas representações alternativas de resumos que podem ser consumidos por outros aplicativos com maior facilidade.

Representar apropriadamente informações contidas em um determinado documento é de fundamental importância para a recuperação informacional. Os termos

de indexação, somados aos resumos, são os principais produtos dessa atividade. Juntos, eles descrevem o conteúdo de um registro, indicando seus pontos principais. Nesse sentido, são elementos que facilitam a comunicação do conhecimento, já que funcionam como "ferramentas de representação", necessárias em um processo inicial de filtragem, permitindo assim que a informação flua entre o universo dos documentos originais e o dos usuários de informação, de acordo com Pinto (1999).

1.3.2 *Medical Subject Heading* - MeSH

O MeSH [16]é o dicionário controlado da *U.S National Library of Medicine (NLM)*, cuja versão inicial foi disponibilizada em 1953, que consiste em um conjunto de descritores em estrutura hierárquica que possibilita consultas em vários níveis de especificidade. Estes descritores estão organizados tanto em ordem alfabética quanto em ordem hierárquica. É constituído por cerca de 28.000 descritores na versão 2016 e mais de 232.000 termos suplementares. O MeSH é utilizado para indexar os artigos que estão disponíveis no banco de dados MEDLINE/PubMed[17], e para indexar o banco de dados da NLM.

Este dicionário está disponível sem custos para *download* através do site NLM, arquivo contendo todos os descritores do MeSH, de forma a possibilitar que outras bases de dados possam utilizá-la para indexação e para estudos. O MeSH é um importante instrumento para indexação, classificação e recuperação de informações da área de ciências das saúde e é amplamente utilizado em ferramentas de aplicação da informática em saúde.

No MeSH, um descritor representa uma classe de conceitos, enquanto um conceito representa uma classe termos sinônimos. A organização do MeSH se dá em 16 categorias de assuntos, sendo que cada uma se divide em subcategorias, nas quais os descritores subordinados são organizados hierarquicamente numa relação do mais genérico para o mais específico.

[16] http://www.ncbi.nlm.nih.gov/mesh
[17] http://www.ncbi.nlm.nih.gov/pubmed

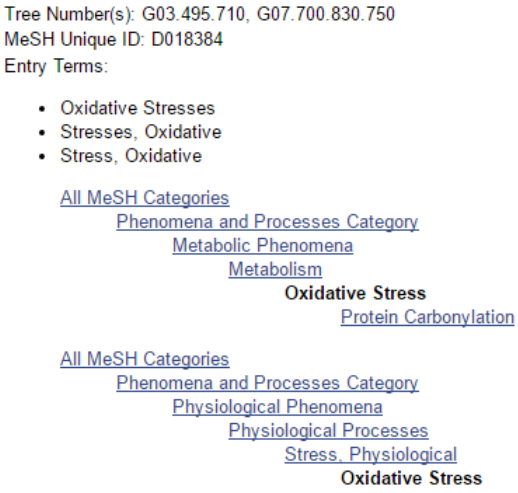


Figura 3 - Exemplo do termo *Oxidative Stress* no MeSH

1.3.3 Descritores em Ciência da Saúde - DeCS

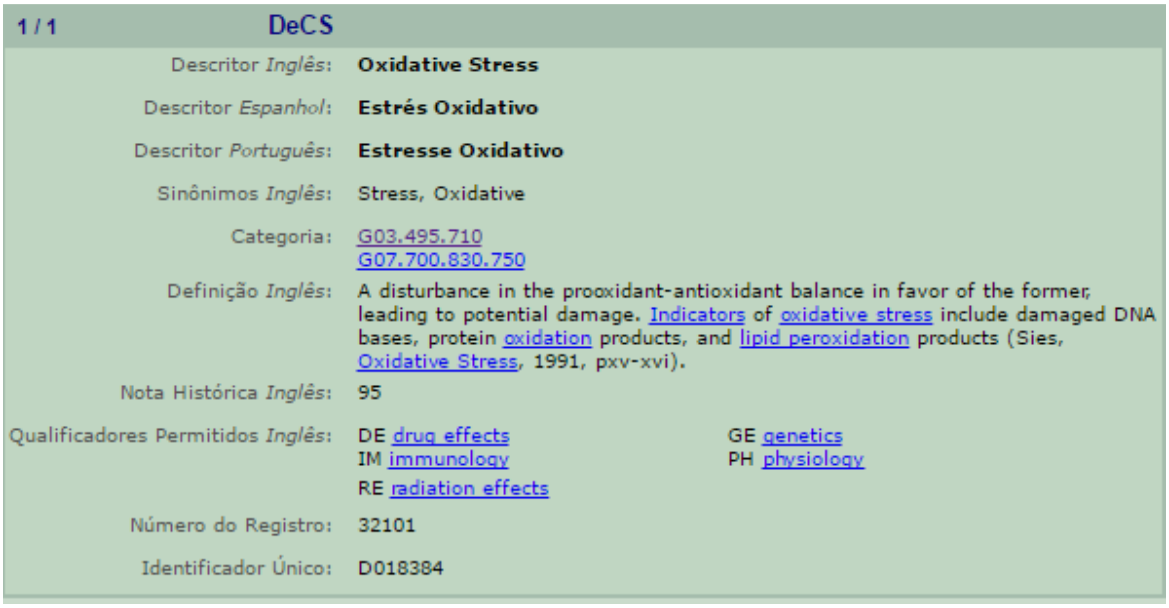
O DeCS[18] é um vocabulário estruturado criado pela Bireme para servir como uma linguagem única na indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos, e outros tipos de materiais, assim como para ser usado na pesquisa e recuperação de assuntos da literatura científica nas fontes de informação disponíveis na Biblioteca Virtual em Saúde (BVS)[19] como LILACS, MEDLINE e outras. Foi desenvolvido a partir do MeSH com o objetivo de permitir o uso de terminologia comum para pesquisa em três idiomas, proporcionando um meio consistente e único para a recuperação da informação independentemente do idioma.

Além dos termos médicos originais do MeSH foram desenvolvidas as áreas específicas de Saúde Pública, Homeopatia, Ciência e Saúde e Vigilância Sanitária. Os conceitos que compõem o DeCS são organizados em uma estrutura hierárquica permitindo a execução de pesquisa em termos mais amplos ou mais específicos ou todos os termos que pertençam a uma mesma estrutura hierárquica.

De acordo com a Bireme, o DeCS é um vocabulário dinâmico totalizando 32.481 descritores, sendo destes 27.883 do MeSH e 4600 exclusivamente do DeCS.

[18] http://decs.bvs.br/
[19] http://bvsalud.org/

Existem 2.123 códigos hierárquicos de categorias DeCS em 1534 descritores MeSH. As seguintes são categorias DeCS e seus totais de descritores: Ciência e Saúde (217), Homeopatia (1.948), Saúde Pública (3.477) e Vigilância Sanitária (825). O número é maior que o total, pois um descritor pode ocorrer mais de uma vez na hierarquia. Por ser dinâmico, registra processo constante de crescimento e mutação registrando a cada ano um mínimo de 1000 interações na base de dados dentre alterações, substituições e criações de novos termos ou áreas.



1 / 1 DeCS

Descritor *Inglês*:	**Oxidative Stress**
Descritor *Espanhol*:	**Estrés Oxidativo**
Descritor *Português*:	**Estresse Oxidativo**
Sinônimos *Inglês*:	Stress, Oxidative
Categoria:	G03.495.710 G07.700.830.750
Definição *Inglês*:	A disturbance in the prooxidant-antioxidant balance in favor of the former, leading to potential damage. Indicators of oxidative stress include damaged DNA bases, protein oxidation products, and lipid peroxidation products (Sies, Oxidative Stress, 1991, pxv-xvi).
Nota Histórica *Inglês*:	95
Qualificadores Permitidos *Inglês*:	DE drug effects GE genetics IM immunology PH physiology RE radiation effects
Número do Registro:	32101
Identificador Único:	D018384

Figura 4 - Exemplo do termo *Oxidative Stress* no DeCS

1.4 Conceitos da área de Tecnologia da Informação

1.4.1 XML - *Extensible Markup Language*

XML é uma especificação de formato de texto que foi originalmente concebido para enfrentar os desafios da publicação eletrônica em larga escala, mas que está desempenhando um papel cada vez mais importante na troca de uma ampla variedade de dados na Web. Tem como vantagens ser um formato aberto e extensível, independente do sistema operacional, da linguagem de programação, ou da fonte de

dados utilizados, além de possibilitar uma autodescrição dos dados. De uma forma geral, o formato XML é o padrão predominante na integração de informações.

1.4.2 Mineração de Dados

Fayyad et al. (1996) define a Mineração de Dados como "um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob limitações computacionais aceitáveis, produzem um conjunto de padrões de certos dados."

Já Cabena et al. (1998), a define como "um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados".

1.4.3 Descoberta de Conhecimento em Bancos de Dados

A descoberta de conhecimento em banco de dados um processo de identificação em dados padrões que sejam válidos, novos (previamente desconhecidos), potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão conforme descrito por Fayyad et al. (1996). Este processo é composto pelas seguintes fases, descritas por Silva (2004), conforme mostra a figura a seguir:

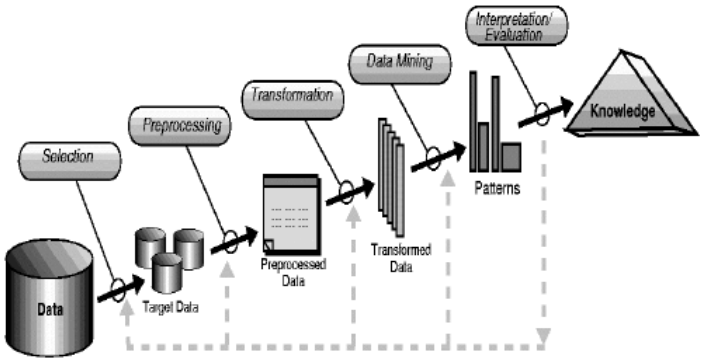


Figura 5 - Etapas da Descoberta do Conhecimento

a) Definição do tipo de conhecimento a descobrir, o que pressupõe uma compreensão do domínio da aplicação bem como do tipo de decisão que tal conhecimento pode contribuir para melhorar;

b) Criação de um conjunto de dados alvo (*selection*): selecionar um conjunto de dados, ou focar num subconjunto, onde a descoberta deve ser realizada;

c) Limpeza de dados e pré-processamento (*preprocessing*): operações básicas tais como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, formatação de dados de forma a adequá-los à ferramenta de mineração;

d) Integração de dados: permite realizar a integração de fontes de dados de diversas origens, além de permitir a consistência e coerência dos dados integrados;

e) Redução de dados e projeção (*transformation*): localização de características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem consideradas para o conjunto de dados, bem como o enriquecimento semântico das informações O volume de dados usado na mineração costuma ser alto. Em alguns casos, este volume é tão grande que torna o processo de análise dos dados e da própria mineração impraticável. Nestes casos, as técnicas de redução de dados podem ser aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, porém, sem perder a representatividade dos dados originais. Isto permite que os algoritmos de mineração sejam executados com mais eficiência, mantendo a qualidade do resultado;

f) Mineração de dados (*Data Mining*): selecionar os métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação ou conjunto de representações; busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão;

g) Interpretação dos padrões minerados (*interpretation/evaluation*), com possibilidade de retorno aos passos citados anteriormente, para posterior iteração;

h) Implantação do conhecimento descoberto (*knowledge*): incorporar este conhecimento ao sistema de informações.

1.4.3.1 Agrupamento de dados

A tarefa de descobrimento de conhecimento denominada Agrupamento (c*lustering*) tem como objetivo identificar e aproximar os registros similares. Um agrupamento é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados.

1.4.3.2 Associação

Uma das tarefas do processo de descoberta do conhecimento é a tarefa de associação, que consiste em identificar quais atributos estão relacionados. Apresentam a forma: SE atributo X ENTÃO atributo Y. É uma das tarefas mais conhecidas devido aos bons resultados obtidos.

1.4.4 Recuperação de Informações

Existem diversos modelos de Recuperação de Informações, entre os quais o modelo Booleano, o Modelo Espaço Vetorial, e os Modelos Probabilísticos são os modelos clássicos de recuperação, segundo Baeza et al (1999). Dentre esses modelos, o mais conhecido e utilizado é o modelo espaço vetorial.

Apesar de ser o modelo mais difundido, Recio-Garcia et al (2008) considera que este modelo tem pouco poder de expressão semântica e apresenta dificuldades para

explicar os resultados, mas concordam que esta técnica apresenta bons resultados, principalmente quando combinada com outras técnicas, como a Indexação Semântica Latente (LSI).

Neste modelo de recuperação cada documento é representado por um vetor de termos e cada termo possui um valor associado, o qual indica grau de importância desse documento. Dessa forma, cada documento é representado por um vetor associado, constituído por pares de elementos na forma { (termo1, peso1), (termo2, peso2), (termo3, peso3) }. O peso do termo em um documento pode ser calculado de formas diversas, entretanto, geralmente é utilizada a frequência em que o termo aparece no documento. Cada elemento do vetor é considerado uma coordenada dimensional. Assim, os documentos podem ser colocados em um Espaço Euclidiano de **n** dimensões, onde **n** é o número de termos, e a posição do documento em cada dimensão é dada pelo seu peso. Nesse espaço, a consulta do usuário também é representada por um vetor. Assim, pode-se comparar os vetores dos documentos com o vetor de consulta e o pode-se calcular o grau de similaridade entre eles, assim descreve Igarashi (2005). Existem várias formas de se calcular essa similaridade. Uma das formas mais utilizadas é a similaridade por cosseno, devido ao seu grau de estabilidade, conforme descrito por Egghe (2002) e pode ser representada a partir da seguinte fórmula:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \qquad \sigma(D,Q) = \frac{\sum_k (t_k \times q_k)}{\sqrt{\sum_k (t_k)^2} \times \sqrt{\sum_k (q_k)^2}}$$

Equação 1 - Cálculo de Similaridade por Cosseno

O grau de similaridade obtido através da medida do cosseno representa as distâncias entre os documentos que estão sendo comparados. Ao invés de adotar um critério binário, os documentos são ordenados com base no grau de similaridade. Dessa forma, um documento pode ser recuperado ainda que satisfaça a consulta parcialmente. O grau de similaridade varia entre 0 e 1 e quanto mais próximo de 1, melhor um documento estará ranqueado em relação à consulta. É importante observar

que os cálculos que envolvem cosseno não são lineares, o que significa que diferenças aparentemente pequenas entre os ângulos formados entre os vetores produzem aproximações ou distanciamentos bastante significativos.

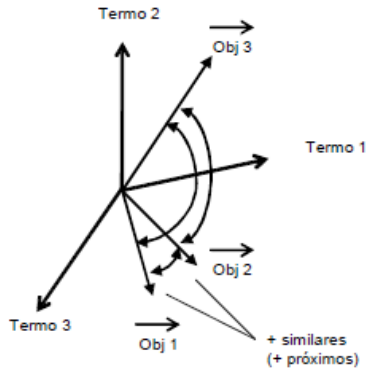


Figura 6 - Modelo Espaço Vetorial (WIVES,2002)

1.4.5 Expansão de Consultas

Ao realizar uma consulta em um sistema de busca, é comum que as palavras utilizadas como argumento sejam diferentes daquelas utilizadas pelos autores dos documentos, ainda que essas palavras estejam relacionadas entre si, ou até possuam o mesmo significado. A expansão de consultas é a técnica que permite contornar esse tipo de problema. O objetivo dos algoritmos que implementam essa técnica é a formulação de uma nova consulta mais elaborada, de forma a melhorar a eficiência na recuperação de informações, a partir de uma consulta realizada inicialmente pelo usuário, descrito por Silva (2000). Essa técnica permite a utilização de palavras ou expressões diferentes, mas que indicam um mesmo objeto, como exemplo, a utilização de acrônimos, abreviações de termos, ou sinônimos. A proposta deste método é realizar uma ampliação do conjunto inicial de palavras, o que aumenta potencialmente a quantidade de resultados relevantes, entretanto, com o desafio de manter o resultado obtido como boa alternativa aos termos fornecidos inicialmente, e de não onerar demasiadamente a infraestrutura tecnológica que suporta o sistema de informações,

uma vez que as consultas modificadas podem ser longas e exigirão maior grau de processamento. A expansão de consultas pode se basear em estatísticas de ocorrência dos termos nos documentos, em tesauros e vocabulários estruturados.

1.4.6 Algoritmo *Trigram Phrase Matching*

Conforme descrito por Tardelli (2008), trigramas são cadeias de três caracteres extraídos de um texto. *Trigram Phrase Matching* é um método de identificação de frases de texto que possuem uma alta probabilidade de serem sinônimas. Baseia-se na representação de cada frase por um conjunto de trigramas que são extraídos dessa frase. Os caracteres do trigrama são utilizados como termos chaves de uma frase, enquanto as palavras são usadas como termos chaves que representam um documento. A similaridade das frases são então calculadas utilizando-se a medida de similaridade do cosseno do vetor formado pelas frases. Para realizar a extração dos trigramas de um frase, é necessário :

- transformar o texto para letras minúsculas;
- a frase é separada em várias palavras e cada palavra produzirá trigramas.
- K+1 trigramas sobrepostos são produzidos de frases com K+3 caracteres;
- um caractere ! é anexado ao primeiro trigrama produzido de cada palavra, o qual é contado duplamente;
- a primeira letra de cada palavra também é extraída, e um caractere # é anexado;
- um trigrama adicional é produzido para cada duas palavras adjacentes, com suas
- primeiras letras separadas por um espaço;
- um único “trigrama” é produzido de “frases” com até três caracteres.

Exemplo:

o texto ***DNA sequence selectivity*** produz o seguintes trigramas:

dna seq equ que uen enc nce sel ele lec ect cti tiv ivi vit ity

dna! dna! seq! seq! sel! sel!

d# s# s#

d s s s (trigramas de junção)

Para cada trigrama da coleção é atribuído um peso global $\sqrt{\log\left(\frac{N}{nt}\right)}$, onde **N** é o tamanho da coleção (número de textos), e **nt** o número de textos da coleção contendo o trigrama. Para cada trigrama é atribuído um peso local $\log(1 + ft)$, onde **ft** é a frequência do trigrama no texto.

Um dado texto é representado pela representação vetorial normalizada ("comprimento" do vetor igual a 1) e de dimensão igual ao número de trigramas distintos que ocorrem na coleção, com o valor dos elementos calculados pelos respectivos pesos locais multiplicados pelos pesos globais. A similaridade entre dois textos varia entre 0 e 1, sendo calculada pelo cosseno do ângulo formado pelas respectivas representações vetoriais normalizadas, conforme Tardelli (2008) descreve. Este algoritmo foi publicado pela NLM, no âmbito no projeto *Indexing Initiative*[20], e foi amplamente utilizado na indexação de documentos dessa biblioteca.

Um exemplo prático de utilização do algoritmo neste estudo foi a identificação de erros de grafia dos termos "**Estresse Oxidativo**" e "**Oxidative Stress**". Foram identificadas as seguintes variações e erros:

[20] https://ii.nlm.nih.gov/

Quadro 1- Exemplo de uso do algoritmo *Trigram Phrase Matching*

ESTRESSE OXIDATIVO	**Similaridade**	**OXIDATIVE STRESS**	**Similaridade**
STRESS OXIDATIVO	0,72563336	OXIDATIVE ESTRESS	0,74353152
STRESS OXIDTAIVE	0,47562369	OXIDATIVE ESTRESSE	0,72563336
ESTRESE OXIDAIVO	0,80124159	OXIDATIVE STRES	0,97336294
ESTRESS OXIDATIVO	0,97592817	OXIDADATIVE STRESS	0,95124739
		OXIDATICE STRESS	0,89487082
		OXIDLATIVE STRESS	0,87223993

O cálculo de similaridade entre os termos utilizando-se o algoritmo *Trigram Phrase Matching* permite identificar possíveis erros de grafia. Normalmente utiliza-se o valor 0,7 como similaridade **mínima** aceitável para a correlação entre os termos comparados. Este valor, denominado *threshold*, é determinado pelo cosseno de 45 graus (0,7071). Similaridades menores que este valor denotam ângulos muito abertos entre os vetores de consulta e resultados.

Dessa forma, na tabela anterior, o termo "STRESS OXIDTAIVE" é considerado similar ao termo "OXIDATIVE STRESS". Este algoritmo foi também utilizado no estudo com o objetivo de avaliar a similaridade entre vetores de trigramas dos currículos.

Para realizar este estudo, foi necessário adaptar o algoritmo original de similaridade baseado em trigramas. A aplicação direta do algoritmo sobre o agrupamento de informações (título dos artigos) geraria vetores muito pouco semelhantes, tendo em vista que a lógica do algoritmo leva em conta os trigramas de junção, os quais são formados pela primeira letra da palavra anterior, mais o caracter "espaço", acrescidos da primeira letra da próxima palavra. Chamando de **n** a quantidade de títulos a serem agrupados, a "concatenação" direta dos títulos geraria **n-1** trigramas de junção entre os títulos das produções. Para evitar este problema, o algoritmo foi adaptado, para extrair as sequências distintas de trigramas de cada um dos títulos, mas não levando em conta as junções existentes entre um título e outro.

1.4.9 Métricas de avaliação de qualidade

Métricas de avaliação de qualidade são medidas utilizadas para avaliar o quão preciso é o conjunto resposta de um sistema de busca. Baeza-Yates et al (2009), define este tipo de avaliação como Avaliação de Desempenho de Recuperação. As medidas mais conhecidas são **precisão** e **cobertura**. Essas medidas avaliam a qualidade de um conjunto de documentos retornados para uma determinada consulta. Precisão é a proporção dos documentos recuperados que são relevantes para uma dada consulta em relação ao total de documentos recuperados. Cobertura é a razão entre o número de documentos recuperados que são relevantes para uma consulta e o total dos documentos na coleção que são relevantes para a consulta.

$$precisao = \frac{|relevantes \cap recuperados|}{|recuperados|} \qquad cobertura = \frac{|relevantes \cap recuperados|}{|relevantes|}$$

Equação 2 - Cálculos de Precisão e Cobertura

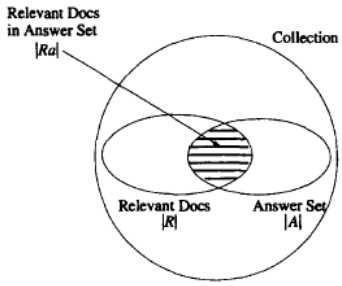


Figura 7 - Exemplo de *precisão* e *cobertura* (BAEZA,2009)

1.5 OBJETIVOS

1.5.1 OBJETIVO GERAL

O objetivo deste estudo foi verificar o impacto da adição de vocabulários estruturados da área de Ciências da Saúde na recuperação de perfis de pesquisadores a partir dos dados de produções científicas registradas no Currículo Lattes.

1.5.2 OBJETIVOS ESPECÍFICOS

Elaborar uma tabela de termos relacionados à área de Ciências da Saúde e Ciências Biológicas, baseada nas palavras-chaves das produções científicas registradas nos Currículos Lattes dos pesquisadores e na importação de dados de vocabulários estruturados dessas áreas do conhecimento, utilizando técnicas de mineração de dados.

Aplicar o método de expansão de consultas, baseado nos termos citados anteriormente, em uma amostra de dados do Currículo Lattes, composta por currículos de doutores que atuam nas áreas de Ciências da Saúde e Ciências Biológicas, e avaliar os resultados utilizando o cálculo de similaridade por cosseno.

2 METODOLOGIA

Neste estudo, foram realizados cruzamentos de informações entre as palavras-chaves dos artigos publicados em periódicos com as informações disponibilizadas pelo MeSH, DeCS e PubMed. O período utilizado no estudo foi 2009 a 2014. Foram incluídos na amostra de dados os artigos publicados que aparecem nos currículos dos membros do Comitê Assessor BF (CA-BF), que tiveram mandato nesse período, bem como nos currículos dos pareceristas que emitiram pareceres favoráveis para as propostas encaminhadas a esse comitê, nas Chamadas Universais[21] realizadas no mesmo período.

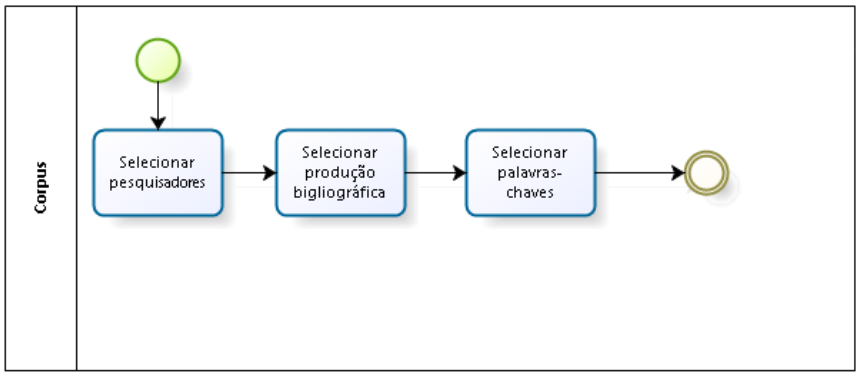


Figura 8 - Seleção do corpus

O CA-BF é o comitê responsável pelo o julgamento das propostas das áreas de Biofísica, Bioquímica, Farmacologia, Fisiologia e Neurociências. É composto por 14 membros, dos quais 4 são suplentes, distribuídos na composição de cinco representantes na subárea de Bioquímica, três representantes em Farmacologia e Fisiologia, dois em Biofísica e um em Neurociências. Os pesquisadores membros e ex-membros desse comitê, bem como os avaliadores, foram escolhidos para compor a amostra de dados pela provável aderência de seus currículos aos termos que compõem o MeSH e DeCS.

[21] Disponível em http://www.cnpq.br/web/guest/chamadas-publicas

Os artigos completos foram escolhidos para fazer parte do estudo devido a sua representatividade no volume de informações registradas na base de dados do Currículo Lattes. Outro fator que influenciou na escolha desse tipo de produção foi a possibilidade de obtenção de metadados através do Crossref, PubMed e outras bases de dados disponíveis na Internet.

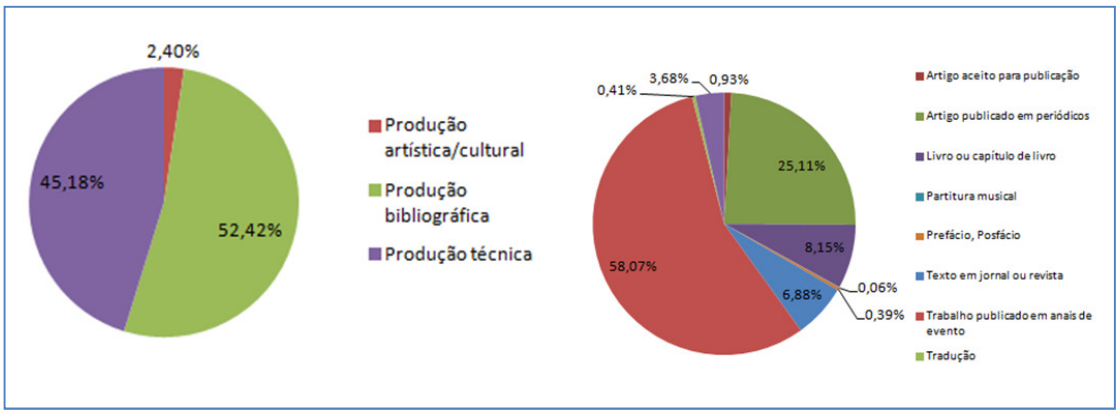


Figura 9 - Distribuição da produção C&T no Currículo. À esquerda, distribuição geral; à direita, distribuição da produção bibliográfica.

Fonte: Plataforma Lattes

O módulo de Sugestão de pareceristas, integrante da PICC, ferramenta baseada em estudo de metodologia para extração de perfis, cujo desenvolvimento foi baseado em estudo que teve como objetivo a proposição de uma metodologia para extração de perfis de pareceristas, na qual os nomes dos pareceristas são sugeridos aos técnicos do CNPq, é uma ferramenta de apoio à equipe técnica e é baseada em informações que estão registradas no Lattes dos pareceristas. Foram considerados para efeito de estudo apenas os currículos de pareceristas que emitiram ao menos um parecer favorável em alguma proposta de fomento no período citado anteriormente. Essa restrição foi necessária diante da inexistência de mecanismo sistêmico que registre de forma apropriada que determinado parecerista tenha sido sugerido e indicado de forma equivocada.

Os passos seguintes envolveram as seguintes atividades:

a) Coleta de dados junto às bases de dados envolvidas no estudo;

b) Agregação das palavras-chaves de produções realizadas em coautoria e identificação dos termos no MeSH e DeCS;

c) Verificação da aderência dos novos termos aos dados obtidos junto ao PubMed e Currículos Lattes;

d) Realização de consultas no Currículo Lattes com as palavras-chaves originais dos currículos e comparação entre os vetores dos currículos obtidos.

e) Realização de consultas no Currículo Lattes com as palavras-chaves originais, acrescidas dos novos termos obtidos por meio da expansão de consulta e comparação entre os vetores dos currículos obtidos;

f) Avaliação dos resultados por meio da similaridade entre os currículos obtidos nas consultas.

2.1 Coleta de dados

O dados iniciais utilizados neste estudo foram obtidos a partir dos bancos de dados dos sistemas de informação do CNPq (Plataforma Integrada Carlos Chagas, Currículo Lattes), e de outras fontes externas, como Crossref, MeSH, DeCS e PubMed.

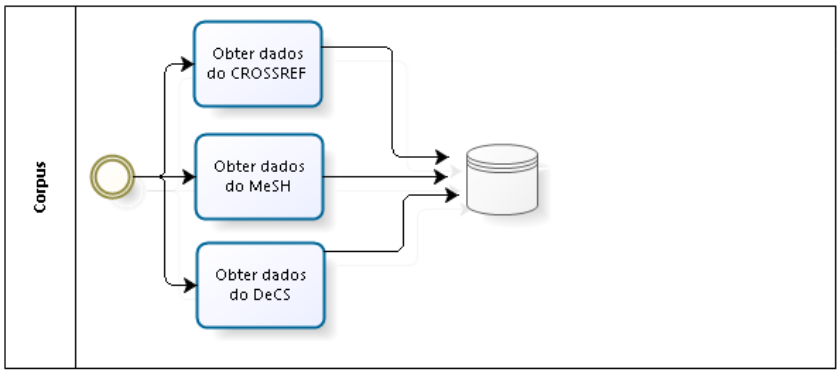


Figura 10 - Diagrama de obtenção de dados externos

As informações sobre os pesquisadores foram obtidas a partir do banco de dados da PICC e Currículo Lattes. A amostra de dados é composta por 1.670 pesquisadores. As informações sobre os artigos foram obtidas a partir do banco de dados do Currículo Lattes. Foram utilizados todos os artigos publicados pelo grupo escolhido de pesquisadores, que possuem o campo DOI preenchido no Currículo Lattes. Essa extração de dados gerou inicialmente uma lista com 39.447 artigos distintos.

Foram realizadas consultas automatizadas junto ao Crossref utilizando-se o DOI das produções bibliográficas como parâmetro de busca para recuperação de metadados de dos artigos no formato XML. Para ter acesso aos dados fornecidos pelo site, é necessário a realização de convênio.

As informações sobre os descritores do MeSH foram obtidas através de *download* do arquivo completo dos termos, fornecido pela NLM, através do seu portal na internet. As informações do MeSH são disponibilizadas no formato XML, e diante disso, foi necessário realizar a importação dessas dados em uma estrutura adequada em banco dados, para que se pudesse realizar os cruzamentos das informações. Foram importados ao todo 229.522 registros.

Para complementar as informações necessárias ao estudo, foi utilizado o DeCS, disponibilizado pela Bireme através do seu portal de internet, e através de endereço próprio para obtenção da informação também em formato XML, o padrão de arquivo utilizado nas trocas de informações entre instituições e empresas, na integração de dados. Foram importados ao todo, 63.968 registros. A construção do DeCS é baseada nos descritores do MeSH, e as informações estão disponíveis em inglês, português e espanhol. Para realização deste estudo, foram utilizados apenas os descritores em inglês e as suas respectivas traduções para o português.

O PubMed também disponibiliza informações sobre as produções bibliográficas pela internet, ou através de ferramentas próprias para extração de dados, denominadas *E-utilities*, as quais permitem ao interessado construir programas de computadores que automatizem a realização de consultas, *download* de sumários ou dos metadados completos dos artigos em diversos formatos, estando o XML entre os formatos

disponíveis. Além de fornecer informações sobre título, autoria, data de publicação, ISSN, páginas, está disponível também o texto relativo resumo do artigo, palavras-chaves e descritores MeSH associadas à produção bibliográfica.

Foram realizadas ainda extrações de dados para recuperação de informações complementares, junto ao PubMed, utilizando-se o DOI das produções bibliográficas como parâmetro de busca. Em resposta a essas consultas, foram recuperados metadados de 27.092 artigos, ou cerca de 69% da amostra dos dados iniciais.

Apesar de não terem sido coletados metadados de artigos a partir da base dados SciELO, está disponível em seu portal *link* que permite realizar a extração em formato XML.

Para a obtenção dos dados acima, foram utilizados as seguintes tecnologias: *scripts shell* do Windows para gerar as listas de arquivos que foram extraídas, Oracle SQL*Loader para realizar a importação dos arquivos gerados, *scripts* Oracle PL/SQL para realizar a limpeza, padronização e processamento dos dados, linguagem Java para construir e calcular a similaridade dos vetores de consultas e de currículos, utilização do *software* Talend Data Integration no apoio da obtenção e transformação dos dados, CURL - ferramenta e biblioteca para transferir dados da internet, por linhas de comando e Microsoft Excel para construção dos gráficos e apoio na análise dos resultados e algoritmo de similaridade por trigramas, implementados na linguagem Java e em Oracle PL/SQL.

Os dados obtidos, tanto do CNPq, quanto das fontes externas de informação, foram persistidos em tabelas de banco de dados Oracle para que pudessem ser armazenados e processados de forma adequada. Para a realização dos cálculos de similaridade foi necessário a extração dos dados para processamento externo ao ambiente do Banco de Dados.

2.2 Agregação de palavras-chaves e Identificação dos termos

O processamento dos dados foi realizado em várias etapas. Cada conjunto de dados resultante do processamento de cada uma das etapas foi agrupado para permitir

uma comparação entre os resultados obtidos. Para o reconhecimento das palavras-chave equivalentes aos descritores foram definidos os seguintes critérios: termos simples com a mesma grafia. Nesse estudo não foram consideradas a relações semânticas entre os termos.

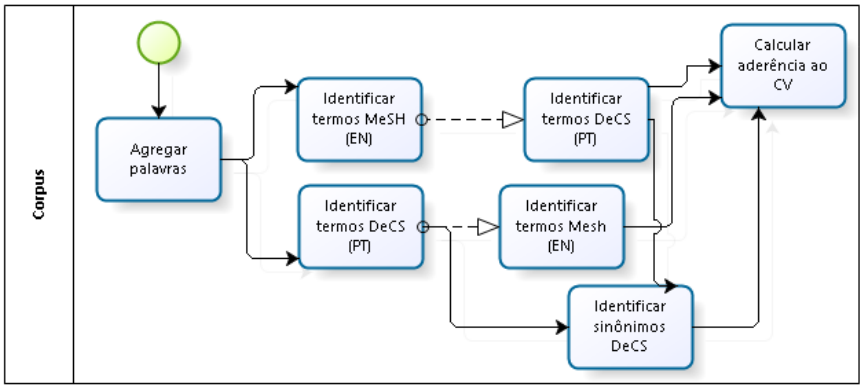


Figura 11 - Diagrama da metodologia de processamento dos dados

2.2.1 Agregação de palavras

No Currículo Lattes, estão disponíveis para preenchimento os campos apropriados para a inserção de palavras-chaves da produção científica. Para cada produção é permitido a associação de até 6 palavras, independentes do idioma. Cerca de 8% dos artigos completos registrados no Lattes não possuem palavras-chaves. Na amostra de dados utilizada no estudo, este percentual é de apenas 0,8 %.

Na metodologia utilizada neste estudo, para cada artigo existente na tabela, foi realizado um agrupamento das palavras-chaves distintas registradas no Currículo Lattes pelos coautores dessa produção bibliográfica. O resultado do agrupamento de palavras permitiu identificar 69.034 palavras-chaves distintas.

Para exemplificar o resultado dessa agregação, será utilizado como exemplo o artigo "***Effects of sirolimus alone or in combination with cyclosporine A on renal ischemia/reperfusion injury***", DOI:10.1590/S0100-879X2010007500058 .

- <ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="667" ORDEM-IMPORTANCIA="">
<DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="Effects of sirolimus alone and association or in combination with cyclosporine A on renal ischemia/reperfusion injury" ANO-DO-ARTIGO="2010" PAIS-DE-PUBLICACAO="" IDIOMA="Inglês" MEIO-DE-DIVULGACAO="VARIOS" HOME-PAGE-DO-TRABALHO="" FLAG-RELEVANCIA="NAO" DOI="10.1590/S0100-879X2010007500058" TITULO-DO-ARTIGO-INGLES="" FLAG-DIVULGACAO-CIENTIFICA="NAO" />
<DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="Brazilian Journal of Medical and Biological Research (Impresso)" ISSN="0100879X" VOLUME="43" FASCICULO="" SERIE="" PAGINA-INICIAL="737" PAGINA-FINAL="744" LOCAL-DE-PUBLICACAO="" />
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Ciclosporina A" PALAVRA-CHAVE-2="Isquemia/reperfusão" PALAVRA-CHAVE-3="Lesão renal aguda" PALAVRA-CHAVE-4="Nefrotoxicidade aguda" PALAVRA-CHAVE-5="Sirolimus" PALAVRA-CHAVE-6="" />

Figura 12 - Extrato do XML do currículo, identificando os metadados do artigo publicado

Esta produção faz parte do currículo de 4 pesquisadores. Cada um deles registrou palavras-chaves diferentes em seus respectivos currículos, conforme pode ser observado na tabela seguinte:

Quadro 2 - Exemplo de lista inicial de palavras chaves, por pesquisador

Pesquisador 1	Pesquisador 2	Pesquisador 3	Pesquisador 4
estudo experimental	ciclosporina A	Ciclosporina A	acute renal failure
insuficiência renal aguda	isquemia renal	Isquemia/reperfusão	Nephrotoxicity
nefrotoxicidade	Nephrotoxicity	Lesão renal aguda	Renal Failure
Sirolimus	túbulos renais	Nefrotoxicidade aguda	Renal tubules
transplante renal		Sirolimus	Reoxigenação
			Reperfusion

No estudo, todas as palavras distintas citadas anteriormente foram consideradas palavras-chaves desse artigo. Para realização das comparações, foi necessário padronizar as palavras, removendo acentos, espaços em brancos desnecessários e transformando as letras em maiúsculas. O resultado desse pré-processamento da informação pode ser visto no quadro a seguir :

Quadro 3 - Exemplo de lista de palavras distintas do artigo

ACUTE RENAL FAILURE	NEPHROTOXICITY
CICLOSPORINA A	RENAL FAILURE
ESTUDO EXPERIMENTAL	RENAL TUBULES
INSUFICIENCIA RENAL AGUDA	REOXIGENACAO
ISQUEMIA RENAL	REPERFUSION
ISQUEMIA/REPERFUSAO	SIROLIMUS
LESAO RENAL AGUDA	TRANSPLANTE RENAL
NEFROTOXICIDADE	TUBULOS RENAIS
NEFROTOXICIDADE AGUDA	

2.2.2 Identificação das palavras diretamente do MeSH (primeira parte)

Para cada uma das palavras-chaves citadas anteriormente, foi realizado um cruzamento de dados com os descritores, conceitos e termos que compõe o MeSH, utilizando-se a operação de SELEÇÃO no banco de dados. Ao todo, foram identificadas 9.157 palavras entre os termos do MeSH, ou cerca de 13,3% do total de palavras. O exemplo a seguir, baseado no mesmo artigo, mostra as palavras que foram encontradas no MeSH, e seus respectivos códigos:

Quadro 4 - Exemplo de palavras identificadas no MeSH

ACUTE RENAL FAILURE	NEPHROTOXICITY
CICLOSPORINA A	RENAL FAILURE **(T051039)**
ESTUDO EXPERIMENTAL	RENAL TUBULES
INSUFICIENCIA RENAL AGUDA	REOXIGENACAO
ISQUEMIA RENAL	REPERFUSION **(D015424)**
ISQUEMIA/REPERFUSAO	SIROLIMUS **(D020123)**
LESAO RENAL AGUDA	TRANSPLANTE RENAL
NEFROTOXICIDADE	TUBULOS RENAIS
NEFROTOXICIDADE AGUDA	

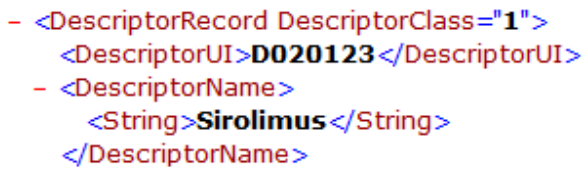


Figura 13 - Exemplo do termo Sirolimus no MeSH

2.2.3 Identificação das palavras no DeCS, a partir do MeSH

O passo seguinte foi identificar a tradução em português do termo encontrado anteriormente, entre os termos que compõem o DeCS. Uma vez que este vocabulário estruturado baseia-se no MeSH, está disponível entre os dados fornecidos no XML o código que permite relacionar o termo DeCS ao termo equivalente no MeSH. O resultado do processamento desses dados identificou 3.955 novas palavras, número este que incrementa em 43% o percentual de palavras identificadas nos vocabulários estruturados.

Quadro 5 - Exemplo de palavras identificadas no DeCS

ACUTE RENAL FAILURE	NEPHROTOXICITY
CICLOSPORINA A	RENAL FAILURE (T051039)
ESTUDO EXPERIMENTAL	RENAL TUBULES
INSUFICIENCIA RENAL AGUDA	REOXIGENACAO
ISQUEMIA RENAL	REPERFUSION (D015424) **REPERFUSÂO**
ISQUEMIA/REPERFUSAO	SIROLIMUS (D020123) **SIROLIMO**
LESAO RENAL AGUDA	TRANSPLANTE RENAL
NEFROTOXICIDADE	TUBULOS RENAIS
NEFROTOXICIDADE AGUDA	

Figura 14 - Exemplo da tradução em português do termo Sirolimus no DeCS

No exemplo anterior, foram identificadas as traduções em português dos termos "***reperfusion***" e "**sirolimus**".

2.2.4 Identificação das palavras diretamente no DeCS

Nessa etapa do processamento, as palavras que não foram identificadas na etapa de identificação de termos MeSH foram comparadas diretamente aos termos em português do DeCS utilizando-se também a operação de SELEÇÃO no banco de dados. Esta comparação resultou na identificação de 1.096 novas palavras e elevou para 55,1% o percentual de palavras identificadas. Nesse exemplo, foram identificados os códigos dos termos "**lesão renal aguda**" e "**túbulos renais**".

Quadro 6 - Exemplo de termos identificados diretamente no DeCS

ACUTE RENAL FAILURE **LESAO RENAL AGUDA (D058186)**	NEPHROTOXICITY
CICLOSPORINA A	RENAL FAILURE (T051039)
ESTUDO EXPERIMENTAL	RENAL TUBULES **TUBULOS RENAIS (D007684)**
INSUFICIENCIA RENAL AGUDA	REOXIGENACAO
ISQUEMIA RENAL	REPERFUSION (D015424) REPERFUSÂO

ISQUEMIA/REPERFUSAO	SIROLIMUS (D020123) SIROLIMO
NEFROTOXICIDADE	TRANSPLANTE RENAL
NEFROTOXICIDADE AGUDA	


```
- <decsws_response service="" tree_id="C12.777.419.780.050">
  + <tree>
  - <record_list>
    - <record lang="pt" db="decs" mfn="53982">
      - <descriptor_list>
          <descriptor lang="en">Acute Kidney Injury</descriptor>
          <descriptor lang="es">Lesión Renal Aguda</descriptor>
          <descriptor lang="pt">Lesão Renal Aguda</descriptor>
        </descriptor_list>
        <synonym_list />
      - <tree_id_list>
          <tree_id>C12.777.419.780.050</tree_id>
          <tree_id>C13.351.968.419.780.050</tree_id>
        </tree_id_list>
      + <definition>
        <indexing_annotation>lesão traumática do rim: indexe RIM /les</indexing_annotation>
        <entry_combination />
        <pharmacological_action_list />
        <see_related_list />
        <unique_identifier_nlm>D058186</unique_identifier_nlm>
```


Figura 15 - Exemplo do termo Lesão Renal Aguda identificado no DeCS

2.2.5 Identificação das palavras no MeSH (segunda parte)

Esta etapa da comparação dos termos envolveu a identificação da tradução em inglês do termo encontrado no passo anterior, entre os descritores, conceitos e termos que compõem o MeSH. Ao final do processamento, foram adicionadas 25.809 novas palavras à lista, o que elevou o percentual de palavras identificadas para 392%. Para exemplificar o resultado desse processo, para termo o "**lesão renal aguda**" foram identificados os seguintes termos a partir de seu código MeSH:

Quadro 7 - Exemplo de termos encontrados no MeSH, a partir dos termos DeCS em português

Acute Kidney Failure	Acute Kidney Failures	Acute Kidney Injuries
Acute Kidney Injury	Acute Kidney Insufficiencies	Acute Kidney Insufficiency
Acute Renal Failure	Acute Renal Failures	Acute Renal Injuries
Acute Renal Injury	Acute Renal Insufficiencies	Acute Renal Insufficiency
Kidney Failure, Acute	Kidney Failures, Acute	Kidney Injuries, Acute
Kidney Injury, Acute	Kidney Insufficiencies, Acute	Kidney Insufficiency, Acute
Kidney Tubule	Kidney Tubules	Renal Failure, Acute
Renal Failures, Acute	Renal Injuries, Acute	Renal Injury, Acute
Renal Insufficiencies, Acute	Renal Insufficiency, Acute	Tubule, Kidney
Tubules, Kidney		

2.2.6 Identificação dos sinônimos dos termos identificados no DeCS

Entre os metadados fornecidos pelo XML do DeCS existe uma estrutura de dados denominada "sinônimos". O processo de identificação desses termos foi deixado por último devido ao fato dessas palavras não estarem identificadas com código MeSH. Essa estrutura de dados foi identificada em cerca de 60% dos termos que compões o DeCS. O total de sinônimos extraídos somam cerca de 73,2 mil termos distintos. Os dados necessários para a execução vieram dos resultados dos processamentos realizados nas etapas 2.2.4 e 2.2.5. O resultado desse processamento elevou de 17 para 49 o número de termos associados ao artigo citado como exemplo.

Quadro 8 - Exemplo da lista final de palavras, adicionados os termos sinônimos do DeCS

ACUTE KIDNEY FAILURE	ACUTE KIDNEY FAILURES	ACUTE KIDNEY INJURIES	ACUTE KIDNEY INJURY
ACUTE KIDNEY INSUFFICIENCIES	ACUTE KIDNEY INSUFFICIENCY	ACUTE RENAL FAILURE	ACUTE RENAL FAILURES
ACUTE RENAL INJURIES	ACUTE RENAL INJURY	ACUTE RENAL INSUFFICIENCIES	ACUTE RENAL INSUFFICIENCY
CICLOSPORINA A	ESTUDO EXPERIMENTAL	INSUFICIENCIA RENAL AGUDA	ISQUEMIA RENAL

ISQUEMIA/REPERFUSAO	KIDNEY FAILURE, ACUTE	KIDNEY FAILURES, ACUTE	KIDNEY INJURIES, ACUTE
KIDNEY INJURY, ACUTE	KIDNEY INSUFFICIENCIES, ACUTE	KIDNEY INSUFFICIENCY, ACUTE	KIDNEY TUBULE
KIDNEY TUBULES	LESAO RENAL AGUDA	NEFROTOXICIDADE	NEFROTOXICIDADE AGUDA
NEPHROTOXICITY	RAPAMICINA	RAPAMYCIN	RENAL FAILURE
RENAL FAILURE, ACUTE	RENAL FAILURES, ACUTE	RENAL INJURIES, ACUTE	RENAL INJURY, ACUTE
RENAL INSUFFICIENCIES, ACUTE	RENAL INSUFFICIENCY, ACUTE	RENAL TUBULES	REOXIGENACAO
REPERFUSAO	REPERFUSION	SIROLIMO	SIROLIMUS
SIRROLIMO	TRANSPLANTE RENAL	TUBULE, KIDNEY	TUBULES, KIDNEY
TUBULOS RENAIS			

Figura 16 - Exemplo de sinônimos do termo Sirolimus no XML do DeCS

2.3 Verificação da aderência dos novos termos com MeSH aos dados recuperados no PubMed.

Ao ser verificado o grau de aderência das informações extraídas a partir do PubMed às informações obtidas após a realização do processamento dos dados, foi possível identificar que em cerca de 41% dos artigos utilizados na amostra houve batimento entre os dados, com pelo menos um termo MeSH identificado. Foi possível observar ainda que as palavras-chaves registradas pelos pesquisadores para as quais não foi possível realizar a identificação no MeSH, estão presentes em 7% dos seus respectivos artigos obtidos junto ao PubMed.

2.4 Verificação da aderência dos novos termos ao Currículo Lattes.

O resultado do processamento das informações utilizando apenas a operação de SELEÇÃO do banco de dados indica que :

a) 40% das novas palavras identificadas estão presentes nos artigos dos currículos dos pesquisadores que fazem parte da amostra de dados;

b) 74% das novas palavras identificadas estão presentes nos artigos dos currículos de todos os pesquisadores que possuem mestrado ou doutorado;

c) 19% das novas palavras identificadas estão presentes nos artigos dos currículos de todos os pesquisadores que possuem graduação ou especialização.

Ao realizar a comparação dos termos incluindo-se os demais tipos de produção científica e tecnológica registradas no currículo, esses números sobem para 45%,74,5% e 31% respectivamente.

3 ANÁLISE DOS RESULTADOS

Para realizar a análise dos resultados obtidos, foram escolhidos os currículos de três pesquisadores que atuam nas áreas de Ciência da Saúde ou em Ciências Biológicas. Dois dos pesquisadores, denominados **pesquisador 1** e **pesquisador 2**, fazem parte da amostra inicial de dados, que é composta por um conjunto de pesquisadores que fazem, ou fizeram parte dos processos de julgamento do CA-BF. O terceiro pesquisador, denominado **pesquisador 3**, não faz parte da amostra inicial de dados.

Para cada um dos três pesquisadores escolhidos, foi calculada a similaridade entre os vetores de seus currículos e os vetores de currículos da amostra inicial, através do algoritmo de trigramas adaptado. Esses vetores foram criados a partir da junção dos títulos das produções científicas (artigos completos, trabalhos, livros e capítulos de livros publicados) e dos títulos das teses orientadas pelo pesquisador.

O resumo das similaridades obtidas entre os currículos dos pesquisadores e os currículos da amostra de dados está descrito no quadro a seguir:

Tabela 1 - Similaridade entre o currículo do pesquisador e a amostra de dados

	Pesquisador 1	**Pesquisador 2**	**Pesquisador 3**
Similaridade			
Mínima	0,51274	0,48803	0,50794
Máxima	0,82461	0,92207	0,92349
Média	0,66855	0,71445	0,73960

Fonte: Elaborado pelo autor – dados da pesquisa

O passo seguinte consistiu em realizar uma consulta direta no Currículo Lattes, utilizando as palavras-chaves originais de cada um dos pesquisadores, e comparar a similaridade entre os vetores de currículos dos pesquisadores utilizados como exemplo e os vetores dos currículos obtidos através da consulta. Uma vez que não foram objeto de estudo as relações semânticas entre as palavras-chaves e termos, e com o intuito de

evitar a obtenção de resultados cujos currículos pertencessem a áreas do conhecimento diferentes daquelas escolhidas na amostra inicial dos dados, a pesquisa das informações foi realizada sobre o conjunto de currículos de doutores que atuam nas áreas de Ciências da Saúde ou em Ciências Biológicas. A tabela a seguir mostra um resumo do resultado das consultas utilizando-se apenas as palavras-chaves originais:

Tabela 2 - Resultado das consultas utilizando-se as palavras originais

	Pesquisador 1		Pesquisador 2		Pesquisador 3	
Similaridade	**%**	**Qtd CV**	**%**	**Qtd CV**	**%**	**Qtd CV**
0,90 - 1,00	0,38%	1	0,12%	1	0,02%	2
0,70 - 0,89	18,80%	50	30,59%	249	22,95%	2.055
0,10 - 0,69	81,20%	215	69,41%	564	77,05%	6.896
Total	100%	266	100%	814	100%	8.953

Fonte: Elaborado pelo autor – dados da pesquisa

O próximo passo consistiu em realizar uma consulta direta no Currículo Lattes utilizando as palavras-chaves originais de cada um dos pesquisadores, acrescida dos termos identificados no MeSH e DeCS e comparar a similaridade entre os vetores de currículos dos mesmos pesquisadores citados no passo anterior, e os vetores dos currículos obtidos através da consulta expandida. Essa consulta foi realizada sobre o mesmo subconjunto de Currículos Lattes utilizado no passo anterior. A tabela a seguir mostra um resumo do resultado das consultas expandidas:

Tabela 3 - Resultado das consultas utilizando-se a expansão de consulta

	Pesquisador 1		Pesquisador 2		Pesquisador 3	
Similaridade	**%**	**Qtd CV**	**%**	**Qtd CV**	**%**	**Qtd CV**
0,90 - 1,00	0,05%	1	0,01%	3	0,01%	2
0,70 - 0,89	6,28%	118	9,39%	2.279	10,78%	3.793
0,10 - 0,69	93,72%	1.759	90,61%	21.986	89,22%	31.383
Total	100%	1.878	100%	24.268	100%	35.178

Fonte: Elaborado pelo autor – dados da pesquisa

Através dos dados da tabela anterior, percebe-se um esperado aumento na quantidade de currículos recuperados através da expansão dos termos originais. Além disso, analisando-se a representatividade dos valores percentuais dos resultados obtidos, verifica-se uma diminuição desses valores relativos, em relação à consulta original. Entretanto, apesar do grande número de novos currículos recuperados, ao realizar um ranqueamento dos resultados baseado no modelo vetorial, por faixas de similaridade, observa-se também um grande aumento em números absolutos de currículos similares, sobretudo na faixa de similaridade compreendida entre **0,70** e **0,89**. Estes valores são compatíveis com as médias de similaridade calculadas inicialmente com os currículos dos pesquisadores que fazem parte da amostra inicial de dados (quadro 9). Nota-se ainda que para o **pesquisador 2** foram encontrados mais 2 currículos com alto grau de similaridade, na faixa entre **0,90** e **1,00**, que não seriam encontrados através da consulta usando-se apenas as palavras-chaves originais de seus currículos.

Os gráficos seguintes exibem de forma comparativa os resultados obtidos utilizando-se consultas com termos originais, e consultas com termos expandidos para os **pesquisadores 1, 2** e **3.** Foi utilizado como ponto de corte o cosseno de 45 graus (0,7071). Nos três gráficos é possível notar a recuperação de currículos melhor ranqueados na consulta expandida, bem como a adição dos novos currículos recuperados: quando o limite mínimo de similaridade da consulta normal é atingido, ainda existem currículos recuperados através da consulta expandida com similaridade maior do que 0,7071.

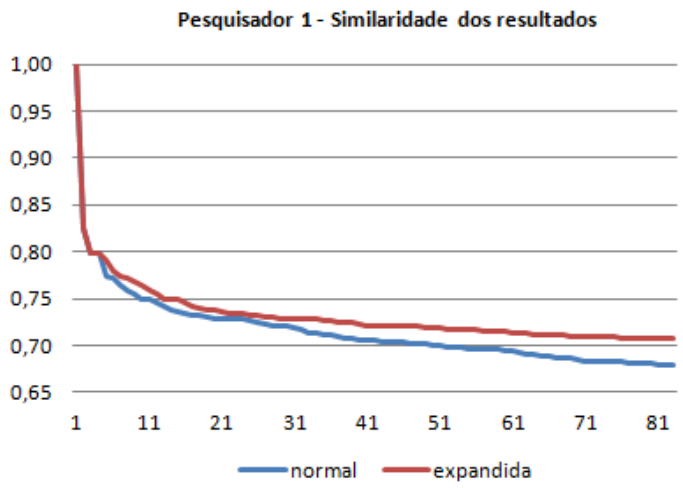


Figura 17 - Recuperação de currículos similares ao do Pesquisador 1

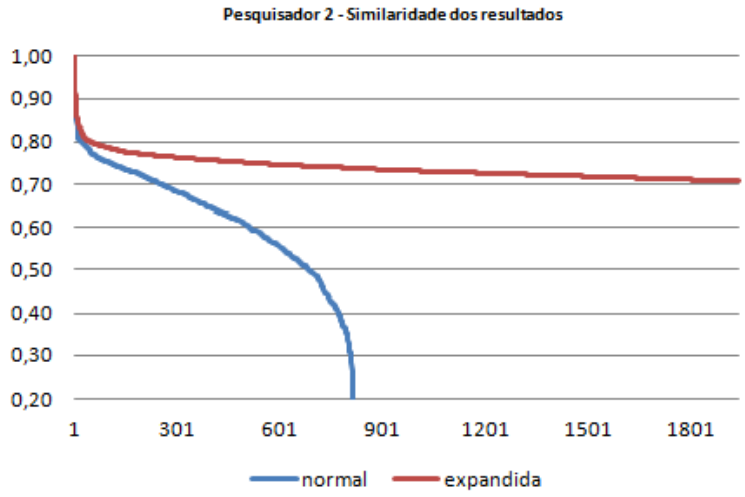


Figura 18 - Recuperação de currículos similares ao do Pesquisador 2

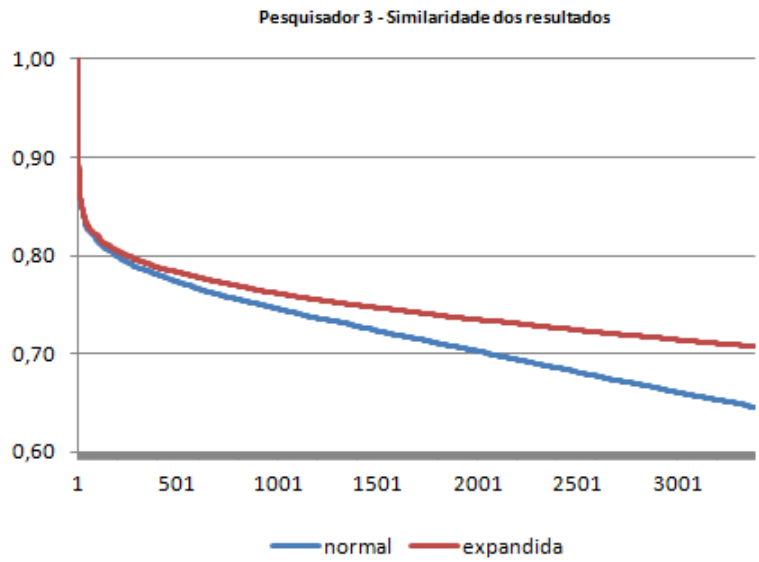


Figura 19 - Recuperação de currículos similares ao do Pesquisador 3

O Anexo I deste documento é um extrato dos números obtidos, onde constam, ordenadas por similaridade, as comparações realizadas entre os currículos dos pesquisadores 1, 2 e 3 e amostra de dados inicial. Constam ainda as similaridades resultantes das consultas que utilizaram apenas as palavras-chaves originais dos currículos, e as consultas que utilizaram palavras-chaves originais acrescidas da expansão de consulta.

4 CONSIDERAÇÕES GERAIS

O CNPq, para o desempenho de suas funções de promoção e apoio ao desenvolvimento, manutenção da pesquisa científica e tecnológica e formação de recursos humanos qualificados, utiliza como subsídios para tomada de decisões, pareceres de Comitês de Assessoramento, de consultores *ad hoc* e de técnicos especializados, que atuam de forma separada ou coordenadamente, conforme estruturação e modo de funcionamento definidos pelo Conselho Deliberativo:

a) **Comitês de Assessoramento (CA):** compõem os CA's mais de 300 pesquisadores, entre titulares e suplentes, selecionados de acordo com sua área de atuação e conhecimento. Eles são escolhidos periodicamente pelo Conselho Deliberativo (CD) , com base em consulta feita à comunidade científico-tecnológica nacional e têm a atribuição, entre outras, de julgar as propostas de apoio à pesquisa e de formação de recursos humanos (BRASIL, 2016a).

b) **Membros dos Comitês:** para analisar, julgar, selecionar e acompanhar os pedidos de projetos de pesquisa e de formação de recursos humanos, o CNPq conta com o apoio de milhares de pesquisadores que constituem sua Assessoria Científico-Tecnológica. Esses pesquisadores, individualmente ou em grupos, têm atribuições específicas e atuam de acordo com suas especialidades.

c) **Comissão de Assessoramento Técnico-Científico (CATC):** esta comissão é formada por 15 integrantes, distribuídos entre três diretores técnicos do CNPq, três representantes da comunidade científica e tecnológica e nove coordenadores de comitês de assessoramento, a CATC é um órgão colegiado criado para auxiliar científica e tecnologicamente a Diretoria Executiva (DEX) e o Conselho Deliberativo (CD).

d) **Núcleo de Assessores em Tecnologia e Inovação (Nati):** é um banco de pesquisadores, os quais são convocados para assessorar o CNPq em suas ações relacionadas com tecnologia e inovação. Essa assessoria poderá se dar na consulta individual ou em grupo sobre um determinado assunto ou tema, bem como na emissão de pareceres ou, especialmente, na formação de comitês avaliadores, quando do julgamento de chamadas, sobretudo as referentes a bolsas DT e as financiadas pelos Fundos Setoriais.

e) **Consultores ad hoc:** são especialistas de alto nível, responsáveis por analisar o mérito científico e a viabilidade técnica dos projetos de pesquisa e das solicitações de bolsas enviadas ao CNPq. Em sua maioria, são bolsistas de Produtividade em Pesquisa que, para o desempenho dessa atividade, são escolhidos pela Diretoria Executiva (BRASIL, 2016a).

O CNPq fomenta a pesquisa científica no país por intermédio de chamadas públicas, que são instrumentos jurídicos, regidos pela Lei nº 8.666, de 21 de junho de 1993, destinados a assegurar a publicidade dos atos da Administração Pública para contratação e financiamento de projetos de pesquisa científica e tecnológica (BRASIL, 1993).

A agência de fomento utiliza como diferencial na apreciação das propostas de financiamento, a avaliação de mérito técnico científico. Essa avaliação é realizada por pesquisadores com expertise no tema de interesse do projeto submetido. A assessoria científica utilizada nas avaliações de mérito técnico-científico dos projetos pode atuar em dois tipos de consultoria, a *ad hoc* e o CA. Esse modelo suporta as decisões do CNPq quanto à distribuição do fomento à pesquisa científica e na aplicação de políticas no âmbito da ciência e tecnologia brasileira (BRASIL, 2016a).

Para terem acesso aos recursos fornecidos pela Agência, os pesquisadores enviam propostas ao CNPq, e para que possam ser contratadas, conforme descrito por Bastos (2009), as propostas passam por um processo de avaliação composto por várias etapas:

a) **Pré-seleção**: consiste na verificação realizada pelo corpo técnico da Agência se o proponente e o objeto da proposta estão de acordo com os requisitos estabelecidos pelo edital ou chamada ao qual foi submetido;

b) **Parecer de consultor ad hoc**: consiste na avaliação de mérito científico e tecnológico da proposta realizada por especialistas nos domínios do conhecimento relacionados à proposta. O resultado dessa análise é uma recomendação de aceitação, ou rejeição, da proposta em função de seu mérito tecnológico e científico, da viabilidade de sua execução e de outros aspectos como inovação, relevância e capacidade da equipe de projeto para realizá-lo. A sugestão da lista de pareceristas é apoiada por um sistema de recomendação automática.

c) **Avaliação por Comitê de Assessoramento**: os comitês de assessoramento são órgãos colegiados cujos membros são nomeados dentre listas de especialistas escolhidos por votação pelos seus pares, possuem mandato fixo e delegação para realizar análise de mérito científico e tecnológico das propostas. Os pareceres dos consultores ad hoc são utilizados como subsídios para a avaliação realizada pelos comitês de assessoramento. Os resultados finais da análise dos comitês resultam em duas listas: uma com as propostas sem mérito para aprovação e, outra, com as propostas com mérito para aprovação em ordem de prioridade de atendimento;

d) **Deliberação final por Diretoria**: ratificação dos pareceres desfavoráveis emitidos pelos comitês e aprovação final das propostas aprovadas, respeitando a disponibilidade de recursos. As propostas com parecer favorável dos comitês são classificadas em uma lista única, de acordo com as prioridades estabelecidas por cada comitê. As propostas que estiverem dentro da disponibilidade orçamentária recebem parecer final de aprovação e são encaminhadas para contratação, ao passo que, aquelas que não alcançarem

prioridade suficiente para atendimento, recebem parecer desfavorável, mas não de mérito.

Os sistemas de recomendação são sistemas de informação que auxiliam o usuário a recuperar informações através da previsão de seus interesses, informando-lhe conteúdo, fontes de consulta ou outras informações, Marques (2007). No CNPq, a indicação de consultor para avaliação de proposta é apoiada por um sistema de recomendação automática, que extrai uma aproximação dos perfis dos consultores, dos proponentes e das propostas para comparação entre eles a partir de informações textuais como palavras-chave e títulos das propostas, e da produção científica registrada no Currículo Lattes dos pesquisadores envolvidos. A base de dados do Currículo Lattes é composta atualmente, por mais de 4,7 milhões de currículos. Neste cenário, a busca de informações baseada nos métodos tradicionais dificulta a descoberta do conhecimento. A transposição dessa dificuldade pode ser atingida através da adição de novos modelos e metodologias de consulta que sejam capazes de satisfazer às perguntas de quem faz uso do sistema de buscas, sejam pessoas que tenham interesse na busca de pessoas que possuam perfis semelhantes para compor uma equipe de trabalho, seja um sistema de recomendação automática em busca de pareceristas para avaliarem propostas de fomento enviadas ao CNPq.

Neste estudo buscamos avaliar como o agrupamento de palavras-chaves da produção científica de coautores, adicionado aos termos MeSH e DeCS poderiam contribuir na melhoria da recuperação de perfis de especialistas. Os resultados confirmam que a metodologia adotada pode permitir a identificação mais precisa de perfis, melhor ranqueados do que o método tradicional de consultas. O método pode ser adaptado ao atual sistema de recomendação existente na PICC, e ao Sistema de Busca de Currículos Lattes, para permitir que outros usuários, instituições e agências de fomento possam se beneficiar das melhorias decorrentes do uso da metodologia, na busca por perfis de especialistas.

5 CONCLUSÃO

Os resultados obtidos neste estudo permitem concluir que a expansão de consultas no Currículo Lattes, através do agrupamento de palavras de produções realizadas em coautoria e a introdução de termos similares, sinônimos ou traduções, provenientes dos vocabulários estruturados MeSH e DeCS, podem contribuir na descoberta de perfis similares de pesquisadores das áreas de Ciências da Saúde e Ciências Biológicas, e dão a possibilidade de obtenção de currículos que não seriam encontrados através de uma consulta comum, que utiliza apenas as palavras-chaves originais das produções científicas. Dessa forma, este estudo pode servir como insumo para o aprimoramento e evolução dos Sistemas de Busca de Currículos Lattes e do módulo de Recomendação de Pareceristas da PICC.

6 PERSPECTIVAS

Os resultados positivos motivam a realização de estudo sobre a possibilidade de se construir *fingerprints*[22] de pesquisadores baseados em trigramas da produção científica registrada no Currículo Lattes. Além disso, as estruturas de dados criadas para a realização do estudo (tabelas de banco de dados com termos MeSH, DeCS, e sinônimos) podem ser adaptadas para apoiar o preenchimento de palavras-chaves nos formulários de produção científica no Currículo Lattes, visando padronizar o registro das informações, bem como diminuir os problemas de grafia das palavras escolhidas na indexação das produções.

[22] https://www.elsevier.com/solutions/elsevier-fingerprint-engine

7 REFERÊNCIAS

ARONSON, A. R.; BODENREIDER, O.; CHANG, H. F.; HUMPHREY, S. M.; MORK, J. G.; et al. The indexing initiative: a report to the board of scientific counselors of the Lister Hill National Center for Biomedical. 1999.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc. Boston, USA, 1999.

BASTOS, W. M. Metodologia para Recomendação de Consultores Ad hoc Baseada na Extração de Perfis do Currículo Lattes. Dissertação. Universidade de Brasília, 2009.

BRASIL. (1993). LEI Nº 8.666, DE 21 DE JUNHO DE 1993. *Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências.* Brasília. Acesso em 26/07/2016, disponível em <http://www.planalto.gov.br/ccivil_03/Leis/L8666cons.htm>

BRASIL. (2016a). *Membros dos Comitês*. Acesso em 24/07/2016, disponível em: Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/membros-dos-comites>

BRASIL. (2016b). *A Criação*. Acesso em 24/07/2016, disponível em Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/a-criacao>

BRASIL. (2016c). *Anos 50*. Acesso em 24/07/2016, disponível em Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/anos-50>

BRASIL. (2016d). *Anos 60*. Acesso em 24/07/2016, disponível em Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/anos-60>

BRASIL. (2016e). *Anos 70*. Acesso em 24/07/2016, disponível em Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/anos-70>

BRASIL. (2016f). *Anos 80*. Acesso em 24/07/2016, disponível em Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/anos-80>

BRASIL. (2016g). *Anos 90*. Acesso em 24/07/2016, disponível em Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/anos-90>

BRASIL. (2016h). *Guia de PDTI do SISP.*. Acesso em 25/07/2016, disponível em Portal do Sistema de Administração dos Recursos de Tecnologia da Informação - SISP: <http://www.sisp.gov.br/guiapdti/wiki/Apresentacao>

BRASIL. (2016i). Portaria nº816, de 17 de dezembro de 2002: <http://www.cnpq.br/web/guest/regimento-interno-po-816/>

BRASIL. (2016j). *Planejamento Estratégico*. Acesso em 24/07/2016, disponível em Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq: <http://www.cnpq.br/web/guest/planejamento-estrategico/ >

CABENA, P; HADJINIAN, P; STADLER, R; JAAPVERHEES; ZANASI, A. Discovering Data Mining: From Concept to Implementation. Prentice Hall, 1998.

CASTRO, E. Terminologia, palavras-chaves, descritores em saúde: qual a sua utilidade? . Jornal Brasileiro de AIDS, v.2, n.1, p.51-61, 2001.

CESARINO, M. A. da N.; PINTO, M. C. M. F. Cabeçalho de assunto como linguagem de indexação. Revista da Escola de Biblioteconomia da UFMG, Belo Horizonte, v. 7, n. 2, p. 268-88, set. 1978.

COLEPICOLO, E.; HOLANDA, A. J. ; RUIZ, E. E. S. ; WAINER, J. ; PISA, I. T. MESH: de cabeçalho de assunto a tesauro. In: Congresso Brasileiro de Informática em Saúde, 10, 2006, Florianópolis - SC. Anais, 2006.

Computational Science: Ensuring America´s Competitiveness. PITAC Report to the President, EUA, 2005.

EGGHE, L.; MICHEL, C. Strong similarity measures for ordered sets of documents in information retrieval. Information Processing & Management, v. 38, p. 823-848, 2002.

ERCAN, G.; CICECKLI, I. Using lexical chains for keyword extraction. Information Processing & Management, v. 43, p. 1705-1714, 2007.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

GIL URDICIAÍN, B. Manual de lenguajes documentales. Gijón: Trea, 2004.

GREINGER, T.; POTTER, T. Solr in Action. Manning Publications Co., USA, 2014.

Grandes Desafios da Pesquisa em Computação no Brasil – 2006 – 2016. Relatório. 2006.

IGARASHI, W. Construção automática de vocabulários temáticos e cálculo de aderência curricular: uma aplicação aos fundos setoriais. Dissertação. Universidade Federal de Santa Catarina, 2005.

LAAN, R. H. Terminologia: uma inter-relação lógica. Tese. Universidade Federal do Rio Grande do Sul, 2002.

LEIVA, I . G.; SPOTTI, M. **Política de indexação**. São Paulo : Cultura Acadêmica; Marília: Oficina Universitária, 2012.

LOPES, G. R.; SOUTO, M. A. M.; de OLIVEIRA, J. P. M. Sistema de recomendação para bibliotecas digitais sob a perspectiva da web semântica. II Workshop de Bibliotecas Digitais, WDL; SBBD/SBES, p. 21–30. 2006.

KORTH, H. F.; SILBERSCHATZ, A. **Sistema de Bancos de Dados**. 2a. ed. São Paulo: MAKRON Books, 1993.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília, 2004.

LANE, J. Let's make science metrics more scientific. Nature. v.464, p. 488-489, 2010.
MARQUES, T. M. Abordagens de Recomendação para a Recuperação de perfis de usuário : uma proposta para o Currículo Lattes, Dissertação. Universidade de Brasília, 2006.

LIU, Y.; THOMAS, P.; SCHMAKEIT, J.; GEDEON, T. Do users benefit from controlled vocabularies in search interfaces ? In Proceedings of the 2nd European Workshop on Human-Computer Interaction and Information Retrieval, 2012.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to Information Retrieval. Cambridge University Press, New York, USA, 2008.

MARQUES, T. M. Abordagens de Recomendação para a Recuperação de Perfis: uma proposta de modelo. Dissertação. Universidade de Brasília, 2007.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; CESAR-JR, R. M. Caracterizando as redes de coautoria de Currículos Lattes. In Brazilian Workshop on Social Network Analysis and Mining, 2012.

MIGUÉIS, A.; NEVES, B.; SILVA, A. L.; TRINDADE, A.; BERNARDES, J. A. A importância das palavras-chaves dos artigos científicos da área das Ciências Farmacêuticas, depositados no Estudo Geral: estudo comparativo com os termos atribuídos na MEDLINE. R. Ci. Inf. e Doc., Ribeirão Preto, v. 4, n. 2, Ed. esp., p. 112-125, 2013.

NETO, I. R.; ALONSO, L. B. N. **Complexus:** tecendo juntos. Brasília: Paralelo 15, 2011.

NETO, I. R.; ALONSO, L. B. N. **Gestão do conhecimento. O olhar da complexidade**. Brasília: Paralelo 15, 2011.

PRESSMAN, R. S. **Engenharia de Software**. São Paulo: Makron Books, 1995.

PINTO, M. Paradigms for Abstracting Systems. Journal of Information Science, v.25, p. 365-380, 1999.

RECIO-GARCÍA, J. A.; DÍAZ-AGUDO, B.; González-Calero, P. . jCOLIBRI 2 Tutorial. Case-Based Reasoning Framework. Universidad Complutense de Madrid, 2008.

ROBBINS, S. **Comportamento Organizacional**. São Paulo: Pearson Prentice Hall, 2005.

SANTANA, P.H; PACKER, A. L.; BARRETTO, M. Y.; SORTE, G. Servidor de enlaces: motivação e metodologia. Ci. Inf., Brasília, v. 30, n. 3, p. 48-55, 2001

SEDGEWICK, R.; WAYNE, K. **Algorithms**. Boston: Addison-Wesley, 2011.

SILVA,I.; RIBEIRO-NETO, B.; CALADO, P.; MOURA, E.; ZIVIANI, N. Link-based and content based evidential information in a belief network model. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*,pages, 2000.

SILVA, M. P. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka. Livro da Escola Regional de Informática Rio de Janeiro - Espírito Santo. Porto Alegre: Sociedade Brasileira de Computação, v. 1, p. 1-20, 2004.

TARDELLI, A. O. Identificação de artigos relacionados e citações na coleção SciELO de revistas eletrônicas através de algoritmo de similaridade de textos por trigramas. Dissertação. Universidade Federal de São Paulo, 2008.

TOMAEL, M. I. Redes de Conhecimento: O Compartilhamento da Informação e do Conhecimento em Consórcio de Exportação do Setor Moveleiro. Tese. Universidade Federal de Minas Gerais. 2005.

VELOSO, P. et al. **Estruturas de Dados**. Rio de Janeiro: Campus, 1986.

VOGEL, M. J. A Noção de estrutura linguística e de processo de estruturação e sua influência no conceito e na elaboração de linguagens documentárias. Dissertação. Universidade de São Paulo, 2007.

WIVES, L. K,. Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos. Tese. Universidade Federal do Rio Grande do Sul, 2004.

YU, Q.; LONG, C.; LV, Y.; SHAO, H.; HE, P., et al. Predicting Co-Author Relationship in Medical Co-Authorship Networks. PLoS ONE 9(7): e101214. 2014.

ANEXOS

Anexo I - Tabela parcial de similaridade entre os currículos dos pesquisadores 1, 2 e 3 e a amostra de dados, e entre os currículos dos pesquisadores 1,2 e 3 e as consultas normais e expandidas

	Pesquisador 1			Pesquisador 2			Pesquisador 3		
	amostra	normal	expandida	amostra	normal	expandida	amostra	normal	expandida
1	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	0,92349	1,00000	1,00000
2	0,82461	0,82461	0,82461	0,92207	0,88843	0,91256	0,89173	0,92349	0,92349
3	0,79884	0,79884	0,79884	0,91256	0,86197	0,90696	0,88492	0,89218	0,89218
4	0,79029	0,79861	0,79861	0,90696	0,85760	0,88843	0,88128	0,88492	0,88492
5	0,77866	0,77358	0,79029	0,88843	0,85519	0,86197	0,87631	0,88128	0,88128
6	0,76832	0,77248	0,77866	0,85985	0,85177	0,85760	0,86491	0,87631	0,87631
7	0,75546	0,76347	0,77358	0,85760	0,84661	0,85519	0,86413	0,86491	0,86491
8	0,74826	0,75917	0,77248	0,85177	0,83360	0,85177	0,86183	0,86413	0,86413
9	0,74176	0,75546	0,76832	0,84895	0,83346	0,84661	0,86142	0,86183	0,86183
10	0,73761	0,74970	0,76347	0,84661	0,82415	0,84555	0,85760	0,86142	0,86142
11	0,73394	0,74929	0,75917	0,84555	0,81395	0,84020	0,85651	0,85760	0,85760
12	0,73393	0,74586	0,75546	0,84020	0,80758	0,83905	0,85614	0,85691	0,85691
13	0,73357	0,74176	0,74970	0,83905	0,80525	0,83360	0,85542	0,85596	0,85596
14	0,73247	0,73761	0,74929	0,83360	0,80449	0,83346	0,85453	0,85453	0,85453
15	0,73239	0,73534	0,74826	0,83346	0,80290	0,83329	0,85264	0,85264	0,85264
16	0,73051	0,73393	0,74586	0,83329	0,80285	0,83090	0,85154	0,85154	0,85154
17	0,72975	0,73247	0,74176	0,83122	0,80209	0,82999	0,85044	0,84962	0,84962
18	0,72964	0,73136	0,73955	0,83090	0,80166	0,82930	0,84962	0,84833	0,84833
19	0,72841	0,73039	0,73782	0,83059	0,79979	0,82767	0,84833	0,84807	0,84807
20	0,72795	0,72902	0,73761	0,82999	0,79942	0,82415	0,84807	0,84791	0,84791
21	0,72765	0,72879	0,73534	0,82930	0,79834	0,82340	0,84791	0,84780	0,84780
22	0,72711	0,72841	0,73465	0,82767	0,79819	0,82019	0,84659	0,84662	0,84662
23	0,72634	0,72795	0,73393	0,82340	0,79735	0,81911	0,84542	0,84542	0,84542
24	0,72604	0,72765	0,73357	0,81911	0,79683	0,81467	0,84512	0,84512	0,84512
25	0,72464	0,72720	0,73247	0,81395	0,79678	0,81395	0,84394	0,84394	0,84394
26	0,72431	0,72413	0,73136	0,81392	0,79642	0,81385	0,84237	0,84237	0,84237
27	0,72418	0,72213	0,73051	0,81385	0,79614	0,81314	0,84195	0,84195	0,84195
28	0,72413	0,72133	0,73039	0,81314	0,79518	0,81117	0,84173	0,84173	0,84173
29	0,72409	0,72103	0,72902	0,81117	0,79428	0,81065	0,84091	0,84091	0,84091
30	0,72372	0,72035	0,72879	0,81065	0,79195	0,80758	0,84004	0,84004	0,84004
31	0,72358	0,71904	0,72841	0,80758	0,79118	0,80674	0,83954	0,83856	0,83856
32	0,72295	0,71778	0,72817	0,80684	0,79076	0,80651	0,83912	0,83825	0,83825

33	0,72269	0,71372	0,72795	0,80651	0,78999	0,80525	0,83732	0,83639	0,83684
34	0,72206	0,71284	0,72765	0,80525	0,78834	0,80503	0,83699	0,83555	0,83639
35	0,72161	0,71209	0,72720	0,80503	0,78692	0,80500	0,83639	0,83505	0,83580
36	0,72139	0,71130	0,72634	0,80479	0,78658	0,80449	0,83580	0,83429	0,83555
37	0,72133	0,70981	0,72418	0,80449	0,78640	0,80330	0,83492	0,83352	0,83505
38	0,72121	0,70839	0,72413	0,80330	0,78496	0,80290	0,83362	0,83190	0,83492
39	0,72103	0,70780	0,72372	0,80290	0,78276	0,80285	0,83352	0,83082	0,83457
40	0,72103	0,70605	0,72213	0,80285	0,78214	0,80245	0,83247	0,83074	0,83429
41	0,72080	0,70594	0,72139	0,80245	0,78184	0,80209	0,83190	0,82955	0,83352
42	0,72035	0,70509	0,72133	0,80205	0,78154	0,80166	0,83047	0,82910	0,83310
43	0,71995	0,70458	0,72121	0,80175	0,78042	0,80115	0,82992	0,82841	0,83247
44	0,71904	0,70368	0,72103	0,80166	0,77935	0,80098	0,82955	0,82800	0,83190
45	0,71903	0,70291	0,72101	0,80133	0,77923	0,80068	0,82910	0,82743	0,83165
46	0,71888	0,70291	0,72035	0,80098	0,77866	0,80054	0,82898	0,82719	0,83091
47	0,71800	0,70218	0,72015	0,80068	0,77436	0,79979	0,82841	0,82703	0,83074
48	0,71792	0,70218	0,71995	0,80054	0,77366	0,79942	0,82800	0,82694	0,82982
49	0,71778	0,70105	0,71904	0,79948	0,77343	0,79926	0,82790	0,82664	0,82955
50	0,71777	0,70015	0,71888	0,79926	0,77292	0,79834	0,82773	0,82661	0,82910
51	0,71725	0,69967	0,71792	0,79834	0,77244	0,79819	0,82752	0,82649	0,82898
52	0,71706	0,69891	0,71778	0,79735	0,77060	0,79765	0,82719	0,82640	0,82854
53	0,71702	0,69776	0,71725	0,79683	0,77039	0,79735	0,82707	0,82630	0,82841
54	0,71686	0,69772	0,71706	0,79614	0,77033	0,79718	0,82703	0,82549	0,82808
55	0,71657	0,69642	0,71702	0,79579	0,77020	0,79683	0,82694	0,82540	0,82800
56	0,71655	0,69617	0,71657	0,79518	0,76952	0,79678	0,82688	0,82498	0,82752
57	0,71642	0,69600	0,71514	0,79504	0,76945	0,79642	0,82668	0,82475	0,82743
58	0,71633	0,69570	0,71514	0,79498	0,76832	0,79614	0,82664	0,82427	0,82726
59	0,71568	0,69522	0,71512	0,79426	0,76813	0,79579	0,82661	0,82350	0,82719
60	0,71514	0,69502	0,71461	0,79422	0,76699	0,79538	0,82640	0,82349	0,82707
61	0,71512	0,69493	0,71372	0,79410	0,76648	0,79518	0,82549	0,82347	0,82703
62	0,71508	0,69254	0,71284	0,79388	0,76597	0,79506	0,82475	0,82345	0,82694
63	0,71470	0,69117	0,71271	0,79366	0,76594	0,79429	0,82464	0,82330	0,82664
64	0,71461	0,69052	0,71209	0,79290	0,76532	0,79428	0,82379	0,82326	0,82649
65	0,71455	0,68889	0,71182	0,79264	0,76431	0,79426	0,82350	0,82310	0,82640
66	0,71453	0,68831	0,71147	0,79229	0,76373	0,79422	0,82347	0,82289	0,82630
67	0,71404	0,68585	0,71130	0,79215	0,76361	0,79388	0,82345	0,82277	0,82549
68	0,71402	0,68583	0,71054	0,79195	0,76336	0,79340	0,82326	0,82237	0,82540
69	0,71294	0,68582	0,70998	0,79149	0,76314	0,79327	0,82325	0,82229	0,82498
70	0,71271	0,68470	0,70981	0,79076	0,76309	0,79305	0,82310	0,82149	0,82475
71	0,71269	0,68358	0,70977	0,79075	0,76234	0,79290	0,82256	0,82140	0,82427
72	0,71255	0,68348	0,70967	0,79051	0,76200	0,79264	0,82237	0,82124	0,82379
73	0,71243	0,68298	0,70898	0,78960	0,76086	0,79229	0,82229	0,82108	0,82350

74	0,71235	0,68293	0,70875	0,78879	0,76072	0,79215	0,82214	0,82086	0,82349
75	0,71226	0,68279	0,70847	0,78876	0,75994	0,79195	0,82140	0,82081	0,82347
76	0,71224	0,68195	0,70839	0,78862	0,75983	0,79149	0,82124	0,82077	0,82345
77	0,71215	0,68175	0,70833	0,78860	0,75942	0,79118	0,82081	0,82072	0,82330
78	0,71203	0,68170	0,70800	0,78858	0,75818	0,79076	0,82077	0,82070	0,82326
79	0,71176	0,68123	0,70799	0,78834	0,75811	0,79051	0,82072	0,82068	0,82310
80	0,71147	0,68025	0,70780	0,78831	0,75798	0,78999	0,82072	0,82066	0,82289
81	0,71140	0,67988	0,70768	0,78825	0,75784	0,78990	0,82070	0,82063	0,82277
82	0,71105	0,67958	0,70745	0,78814	0,75761	0,78960	0,82068	0,82010	0,82256
83	0,71084	0,67919	0,70734	0,78810	0,75667	0,78879	0,82066	0,81988	0,82237
84	0,71082	0,67817	0,70715	0,78658	0,75628	0,78876	0,82063	0,81965	0,82229
85	0,71054	0,67751	0,70647	0,78652	0,75620	0,78870	0,82010	0,81947	0,82149
86	0,71029	0,67742	0,70605	0,78640	0,75589	0,78868	0,81988	0,81921	0,82140
87	0,71016	0,67678	0,70594	0,78496	0,75578	0,78862	0,81967	0,81889	0,82124
88	0,71011	0,67649	0,70574	0,78462	0,75496	0,78860	0,81965	0,81828	0,82108
89	0,70998	0,67472	0,70571	0,78419	0,75480	0,78834	0,81947	0,81797	0,82086
90	0,70977	0,67424	0,70566	0,78396	0,75477	0,78825	0,81921	0,81795	0,82081
91	0,70974	0,67423	0,70558	0,78390	0,75435	0,78814	0,81889	0,81712	0,82077
92	0,70964	0,67418	0,70512	0,78356	0,75430	0,78810	0,81828	0,81642	0,82072
93	0,70964	0,67322	0,70509	0,78334	0,75404	0,78692	0,81797	0,81623	0,82070
94	0,70951	0,67320	0,70495	0,78326	0,75390	0,78658	0,81795	0,81533	0,82068
95	0,70909	0,67315	0,70458	0,78283	0,75349	0,78652	0,81728	0,81516	0,82066
96	0,70907	0,67269	0,70444	0,78282	0,75342	0,78640	0,81635	0,81500	0,82063
97	0,70906	0,67263	0,70408	0,78261	0,75312	0,78632	0,81634	0,81490	0,82062
98	0,70900	0,67243	0,70401	0,78238	0,75271	0,78621	0,81623	0,81472	0,82010
99	0,70875	0,67211	0,70384	0,78183	0,75170	0,78523	0,81533	0,81462	0,81988
100	0,70858	0,67142	0,70368	0,78154	0,75093	0,78505	0,81485	0,81431	0,81965
101	0,70833	0,66999	0,70366	0,78154	0,75078	0,78496	0,81445	0,81394	0,81947
102	0,70824	0,66957	0,70337	0,78127	0,75067	0,78462	0,81404	0,81343	0,81921
103	0,70823	0,66825	0,70320	0,78122	0,75034	0,78453	0,81343	0,81280	0,81889
104	0,70815	0,66813	0,70291	0,78080	0,75033	0,78447	0,81277	0,81226	0,81828
105	0,70807	0,66780	0,70291	0,78060	0,74968	0,78412	0,81215	0,81225	0,81797
106	0,70804	0,66699	0,70290	0,78047	0,74919	0,78390	0,81195	0,81215	0,81795
107	0,70800	0,66651	0,70218	0,78042	0,74861	0,78340	0,81190	0,81214	0,81763
108	0,70799	0,66602	0,70218	0,78039	0,74818	0,78334	0,81185	0,81190	0,81758
109	0,70783	0,66507	0,70190	0,78035	0,74810	0,78326	0,81160	0,81179	0,81712
110	0,70780	0,66502	0,70185	0,78031	0,74745	0,78283	0,81150	0,81162	0,81642
111	0,70776	0,66436	0,70177	0,77984	0,74659	0,78282	0,81143	0,81161	0,81634
112	0,70772	0,66381	0,70161	0,77974	0,74644	0,78276	0,81113	0,81150	0,81623
113	0,70768	0,66341	0,70124	0,77969	0,74581	0,78261	0,81102	0,81108	0,81533
114	0,70734	0,66294	0,70105	0,77954	0,74565	0,78238	0,81076	0,81076	0,81516

115	0,70720	0,66283	0,70102	0,77935	0,74493	0,78214	0,81074	0,81074	0,81500
116	0,70715	0,66259	0,70019	0,77923	0,74474	0,78194	0,81035	0,81074	0,81496
117	0,70710	0,66059	0,70015	0,77889	0,74397	0,78184	0,81023	0,81056	0,81490
118	0,70668	0,66012	0,70014	0,77866	0,74391	0,78154	0,81010	0,81043	0,81472
119	0,70661	0,65954	0,70013	0,77758	0,74388	0,78127	0,81005	0,81035	0,81462
120	0,70645	0,65773	0,69972	0,77748	0,74358	0,78122	0,81000	0,81015	0,81454
121	0,70638	0,65656	0,69967	0,77742	0,74357	0,78083	0,80990	0,81005	0,81445
122	0,70619	0,65631	0,69963	0,77722	0,74343	0,78080	0,80967	0,81003	0,81431
123	0,70606	0,65468	0,69959	0,77718	0,74263	0,78058	0,80960	0,81000	0,81394
124	0,70605	0,65462	0,69942	0,77693	0,74238	0,78047	0,80956	0,80990	0,81343
125	0,70584	0,65259	0,69937	0,77692	0,74214	0,78042	0,80936	0,80973	0,81315
126	0,70571	0,65199	0,69902	0,77679	0,74190	0,78039	0,80921	0,80967	0,81280
127	0,70571	0,65197	0,69891	0,77644	0,74176	0,78035	0,80914	0,80960	0,81277
128	0,70568	0,65171	0,69865	0,77632	0,74175	0,77974	0,80884	0,80956	0,81232
129	0,70558	0,65139	0,69857	0,77615	0,74154	0,77969	0,80882	0,80914	0,81226
130	0,70554	0,65091	0,69833	0,77611	0,74108	0,77954	0,80873	0,80882	0,81225
131	0,70539	0,65031	0,69822	0,77570	0,74038	0,77937	0,80870	0,80878	0,81215
132	0,70537	0,64949	0,69791	0,77565	0,73957	0,77935	0,80857	0,80873	0,81214
133	0,70523	0,64868	0,69776	0,77526	0,73952	0,77923	0,80809	0,80871	0,81190
134	0,70512	0,64776	0,69772	0,77525	0,73933	0,77866	0,80786	0,80857	0,81185
135	0,70509	0,64674	0,69772	0,77520	0,73927	0,77851	0,80783	0,80844	0,81179
136	0,70501	0,64646	0,69771	0,77487	0,73916	0,77805	0,80751	0,80799	0,81162
137	0,70496	0,64592	0,69758	0,77466	0,73880	0,77758	0,80710	0,80786	0,81161
138	0,70495	0,64563	0,69729	0,77462	0,73817	0,77748	0,80709	0,80768	0,81160
139	0,70491	0,64460	0,69717	0,77443	0,73816	0,77742	0,80703	0,80721	0,81150
140	0,70480	0,64365	0,69710	0,77438	0,73779	0,77730	0,80676	0,80709	0,81143
141	0,70479	0,64214	0,69671	0,77436	0,73754	0,77722	0,80664	0,80678	0,81108
142	0,70446	0,64194	0,69654	0,77432	0,73742	0,77693	0,80663	0,80676	0,81076
143	0,70439	0,64121	0,69650	0,77427	0,73677	0,77692	0,80650	0,80663	0,81074
144	0,70428	0,64080	0,69646	0,77366	0,73565	0,77679	0,80647	0,80660	0,81074
145	0,70427	0,64047	0,69642	0,77357	0,73520	0,77672	0,80635	0,80650	0,81061
146	0,70425	0,63998	0,69624	0,77346	0,73455	0,77661	0,80631	0,80635	0,81056
147	0,70408	0,63924	0,69617	0,77343	0,73430	0,77648	0,80628	0,80628	0,81043
148	0,70397	0,63745	0,69600	0,77310	0,73423	0,77634	0,80611	0,80605	0,81035
149	0,70388	0,63741	0,69596	0,77308	0,73397	0,77620	0,80605	0,80583	0,81015
150	0,70384	0,63585	0,69575	0,77304	0,73353	0,77611	0,80548	0,80579	0,81010
151	0,70366	0,63585	0,69571	0,77292	0,73348	0,77600	0,80525	0,80564	0,81005
152	0,70352	0,63397	0,69570	0,77277	0,73301	0,77593	0,80515	0,80548	0,81003
153	0,70350	0,63359	0,69547	0,77263	0,73278	0,77587	0,80500	0,80548	0,81000
154	0,70337	0,63246	0,69540	0,77244	0,73272	0,77579	0,80499	0,80518	0,80990
155	0,70291	0,63169	0,69539	0,77239	0,73259	0,77570	0,80490	0,80515	0,80973

156	0,70290	0,63168	0,69536	0,77221	0,73214	0,77568	0,80479	0,80500	0,80967
157	0,70283	0,62995	0,69522	0,77203	0,73196	0,77565	0,80438	0,80490	0,80967
158	0,70256	0,62977	0,69504	0,77190	0,73194	0,77553	0,80408	0,80479	0,80960
159	0,70254	0,62873	0,69502	0,77184	0,73170	0,77550	0,80406	0,80474	0,80956
160	0,70253	0,62852	0,69500	0,77168	0,73143	0,77550	0,80406	0,80469	0,80921
161	0,70251	0,62852	0,69493	0,77159	0,73122	0,77526	0,80393	0,80461	0,80914
162	0,70222	0,62810	0,69450	0,77153	0,73103	0,77526	0,80361	0,80444	0,80882
163	0,70220	0,62722	0,69435	0,77092	0,73100	0,77525	0,80352	0,80438	0,80880
164	0,70218	0,62536	0,69352	0,77080	0,73069	0,77487	0,80320	0,80412	0,80873
165	0,70218	0,62431	0,69345	0,77072	0,73045	0,77466	0,80287	0,80408	0,80871
166	0,70214	0,62363	0,69293	0,77072	0,73039	0,77462	0,80246	0,80406	0,80857
167	0,70212	0,62310	0,69288	0,77060	0,73019	0,77438	0,80237	0,80397	0,80844
168	0,70190	0,62035	0,69272	0,77043	0,72990	0,77436	0,80217	0,80393	0,80838
169	0,70189	0,61824	0,69263	0,77041	0,72975	0,77436	0,80208	0,80320	0,80799
170	0,70185	0,61812	0,69254	0,77039	0,72947	0,77432	0,80203	0,80316	0,80786
171	0,70183	0,61625	0,69248	0,77033	0,72927	0,77427	0,80189	0,80316	0,80783
172	0,70177	0,61436	0,69222	0,77024	0,72925	0,77383	0,80176	0,80246	0,80768
173	0,70150	0,61436	0,69204	0,76998	0,72883	0,77366	0,80172	0,80246	0,80721
174	0,70143	0,61428	0,69198	0,76989	0,72868	0,77366	0,80161	0,80237	0,80710
175	0,70140	0,61190	0,69192	0,76986	0,72866	0,77357	0,80159	0,80217	0,80709
176	0,70140	0,61131	0,69188	0,76984	0,72815	0,77349	0,80137	0,80208	0,80678
177	0,70140	0,61016	0,69186	0,76952	0,72790	0,77346	0,80125	0,80203	0,80676
178	0,70136	0,60912	0,69175	0,76951	0,72781	0,77343	0,80123	0,80184	0,80670
179	0,70128	0,60874	0,69135	0,76945	0,72757	0,77311	0,80111	0,80176	0,80663
180	0,70124	0,60864	0,69117	0,76935	0,72748	0,77310	0,80099	0,80172	0,80660
181	0,70120	0,60740	0,69112	0,76928	0,72747	0,77308	0,80098	0,80161	0,80650
182	0,70108	0,60658	0,69100	0,76928	0,72701	0,77304	0,80092	0,80159	0,80647
183	0,70106	0,60625	0,69052	0,76903	0,72669	0,77292	0,80092	0,80144	0,80635
184	0,70105	0,60482	0,69047	0,76885	0,72599	0,77277	0,80041	0,80137	0,80628
185	0,70102	0,60386	0,69028	0,76874	0,72538	0,77244	0,80010	0,80125	0,80626
186	0,70085	0,60329	0,69021	0,76860	0,72511	0,77241	0,79998	0,80111	0,80605
187	0,70079	0,60297	0,69005	0,76853	0,72507	0,77225	0,79985	0,80101	0,80583
188	0,70053	0,59901	0,69002	0,76851	0,72476	0,77221	0,79977	0,80098	0,80579
189	0,70053	0,59659	0,68966	0,76840	0,72439	0,77203	0,79944	0,80041	0,80564
190	0,70041	0,59499	0,68941	0,76832	0,72437	0,77194	0,79942	0,80026	0,80548
191	0,70033	0,59478	0,68924	0,76825	0,72369	0,77190	0,79927	0,80020	0,80548
192	0,70030	0,59465	0,68911	0,76823	0,72279	0,77184	0,79918	0,80010	0,80518
193	0,70014	0,59302	0,68911	0,76822	0,72257	0,77168	0,79906	0,79998	0,80515
194	0,70013	0,59023	0,68906	0,76822	0,72230	0,77166	0,79883	0,79985	0,80500
195	0,69999	0,58993	0,68889	0,76807	0,72177	0,77159	0,79872	0,79977	0,80499
196	0,69995	0,58942	0,68879	0,76797	0,72141	0,77153	0,79864	0,79965	0,80490

197	0,69985	0,58883	0,68870	0,76786	0,72106	0,77111	0,79856	0,79942	0,80484
198	0,69977	0,58730	0,68860	0,76777	0,71980	0,77096	0,79854	0,79934	0,80479
199	0,69970	0,58580	0,68843	0,76765	0,71949	0,77092	0,79847	0,79930	0,80474
200	0,69969	0,58483	0,68831	0,76761	0,71934	0,77080	0,79828	0,79927	0,80469