



Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz



PAULO EDUARDO POTYGUARA COUTINHO MARQUES

PADRÃO DE FINANCIAMENTO À PESQUISA EM DENGUE A PARTIR DO DIÁRIO OFICIAL DA UNIÃO

RIO DE JANEIRO

2017

PAULO EDUARDO POTYGUARA COUTINHO MARQUES

PADRÃO DE FINANCIAMENTO À PESQUISA EM DENGUE A PARTIR DO DIÁRIO OFICIAL DA UNIÃO

Tese apresentada ao Programa de Pós-Graduação em Informação, Comunicação e Saúde (Icict), para obtenção do grau de Doutor em Ciências.

Orientadores:

Prof^a. Dr^a. Maria Cristina Soares Guimarães
Prof. Dr. Renato Rocha Souza

Orientador Externo:

Prof. Dr. Fred Popowich

Rio de Janeiro

2017

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

M357 Marques, Paulo| Eduardo Potyguara Coutinho

Padrão de financiamento à pesquisa em dengue a partir do Diário Oficial da União / Paulo Eduardo Potyguara Coutinho Marques. – Rio de Janeiro, 2017.

xii, 142 f. : il. ; 30 cm.

Tese (Doutorado) – Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Pós-Graduação em Informação e Comunicação em Saúde, 2017.

Bibliografia: f. 130-135

1. Financiamento de pesquisa. 2. Política de Ciência e Tecnologia. 3. Diário Oficial da União (DOU). 4. Dengue. 5. Big Data. 6. Análise visual. 7. Ciência da Política Científica. I. Título.

CDD 338.926

PAULO EDUARDO POTYGUARA COUTINHO MARQUES

PADRÃO DE FINANCIAMENTO À PESQUISA EM DENGUE A PARTIR DO DIÁRIO OFICIAL DA UNIÃO

Tese apresentada ao Programa de Pós-Graduação em Informação, Comunicação e Saúde (Icict), para obtenção do grau de Doutor em Ciências.

BANCA EXAMINADORA

Prof^a. Dr^a. Maria Cristina Soares Guimarães
Presidente – (ICICT/PPGICS)

Prof. Dr. Renato Rocha Souza
Coorientador - (EMAP/FGV)

Prof^a. Dr^a. Cicera Henrique da Silva
Membro Interno – (ICICT/Fiocruz)

Prof. Dr. Christovam Barcellos
Membro Interno – (ICICT/PPGICS)

Prof^a. Dr^a. Cláudia Torres Codeço
Membro Externo – (FIOCRUZ)

Prof. Dr. Flávio Codeço Coelho
Membro Externo – (EMAP/FGV)

Dedicatória

Dedico esta tese a Paulo Roberto
(*in memoriam*), meu pai, meu
exemplo.

AGRADECIMENTOS

Várias pessoas a quem agradeço de coração e alma já pertenciam à minha vida; outras surgiram pelo caminho, como se direcionadas por Alguém, no momento certo. Pessoas especiais pelo simples fato de ser GENTE com todas as letras em maiúsculo. Sendo assim, agradeço a:

- Rachel, minha esposa, por estar do meu lado sempre;
- Daniel, meu filho, por ser minha luz e iluminar meus dias com um sorriso ao me acordar de manhã;
- Vera Maria, por me ensinar a buscar meus sonhos, por me apoiar, por demonstrar força, por me dar força, por ser minha mãe;
- Paulo Marcos, meu irmão, por cumprir um papel que não tive como;
- Faisal , Nida, Colin, Claire, Gary e Olga por serem amigos de verdade.

Durante o desenvolvimento desta tese, fui uma pessoa de muita sorte, pois tive três orientadores que foram professores de cunho acadêmico e de vida. Por isso agradeço a:

- Cristina Guimarães, por tentar extrair de todas as formas o melhor de mim;
- Renato Souza, pelos incentivos que me deu nos momentos certo;
- Fred Popowich, por sua paciência em entender o inglês turístico que apresentei no início da jornada e pelas perguntas que me fizeram pensar sempre no melhor caminho.

Não poderia deixar de agradecer a profissionais que me ajudaram durante todo desenvolvimento deste estudo:

- Eder Freire e Angelo Silva pelo time que formamos;
- Ana Maranhão e Jorge Nundes, pelo apoio durante o meu afastamento;
- Carlos Henrique, por ser um profissional exemplar;
- Vinicius Assef, por todas as conversas técnicas que tivemos.

Agradeço ainda ao PPGICS pela vaga para realizar o doutorado sanduíche assim como à CAPES pelo financiamento recebido para a realização daquele.

ΕΠÍΓΡΑΦΕ

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it... (ARIELY, 2013)

RESUMO

A Ciência da Política Científica (CPC) surge como resposta à necessidade de produzir, a partir de estudos interdisciplinares de rigor científico, bases quantitativas sólidas que subsidiem os formuladores de políticas e pesquisadores a analisar a dinâmica da pesquisa e a avaliar seus resultados. A análise destes dados é, portanto, de grande interesse para o Estado no direcionamento de recursos públicos para a pesquisa. Como o volume de dados cresce exponencialmente ao longo do tempo, *big data* se torna uma dimensão potencial para análise no campo. Esta tese se debruça sobre o Diário Oficial da União (DOU) como fonte potencial de dados para análise do financiamento de pesquisa em saúde no Brasil, mais especificamente, na área de dengue. A pesquisa objetivou desenvolver uma metodologia de extração de dados do DOU com a finalidade de explicitar o padrão público de financiamento para pesquisa em dengue no período entre 2005 e 2014, quadro esse que, espera-se, possa auxiliar os tomadores de decisão na análise das políticas públicas no campo. Os dados foram extraídos do DOU utilizando tecnologia desenvolvida exclusivamente para apoiar essa metodologia. Os dados foram categorizados e sistematizados em conjunto com outras duas fontes complementares: a Plataforma Lattes e o Google Acadêmico. A perspectiva usada para dar sentido aos dados coletados foi o *Visual Analytics*. Os resultados apresentam a distribuição do financiamento por pesquisador, região demográfica, agência de fomento, instituição e ano. A metodologia desenvolvida, embora com limites, é um passo importante no manejo do DOU e trazem à tona os dados de financiamento para pesquisa em dengue.

Palavras-chave: Financiamento de pesquisa, Política de Ciência e Tecnologia, Diário Oficial da União (DOU), Dengue, Big Data, Análise Visual, Ciência da Política Científica

ABSTRACT

The Science of Science Policy (SoSP) emerges as a response to the need to produce, from scientific interdisciplinary studies, solid quantitative bases that subsidize policy makers and researchers to analyze the dynamics of the research and evaluate its results. The analysis of these data is, therefore, of great interest for the State in directing public resources for research. As data volume grows exponentially over time, big data becomes a potential dimension for field analysis. This thesis focused on the Federal Official Gazette, in Portuguese *Diário Oficial da União* (DOU) as a potential data source for analysis of health research funding in Brazil, specifically in the area of dengue. The research aimed to develop a methodology with the purpose of, by extracting data from the DOU, explaining the public financing pattern for dengue research in the period between 2005 and 2014, a framework that is expected to assist decision makers in the analysis of Policies in the field. Data were extracted from the DOU using technology developed exclusively for supporting this methodology. The data were categorized and systematized together with two complementary data sources: the Lattes Platform and Google Scholar. Visual Analytics was the perspective used for the data collected makes sense. The results show the distribution of funding per researcher, demographic region, development agency, institution and year. We determined that the methodology developed, although with limits, is an important step in the management of the DOU and brings the funding data for dengue research.

Keywords: Funding for research, Science and Technology Policy, Dengue, Federal Official Gazette, Big Data, Visual Analytics, Science of Science Policy

LISTA DE IMAGENS

Figura 1 – Exemplo de <i>dashboard</i> do site Deep Insights Anytime, Anywhere	20
Figura 2 – Fluxo da ciência	24
Figura 3 – Complexo Industrial da Saúde	40
Figura 4 – Composição do Visual Analytics	55
Figura 5 – Esquema de relacionamento dos elementos deste estudo	58
Figura 6 – Esquema metodológico utilizado nesta pesquisa	65
Figura 7 – Link para acesso de uma página do Diário Oficial da União.....	66
Figura 8 – Exemplos de diagramação das páginas do Diário Oficial da União	68
Figura 9 – Processo de conversão de uma página do Diário Oficial da União para texto	69
Figura 10 – Exemplo de tabelas trabalhadas na conversão de PDF para texto	70
Figura 11 – Exemplo de detecção de linhas verticais e horizontais por Hough Transform	71
Figura 12 – Processo de download e conversão das páginas do DOU.....	71
Figura 13 – Exemplo de texto retirado do DOU de 3/5/1990, Seção 2, página 143	72
Figura 14 – Formação do código de Natureza de Despesa e exemplo.....	88
Figura 15 – Modelagem da base de dados.....	91
Figura 16 – Tela principal do portal.	92
Figura 17 – Tela de busca.....	93
Figura 18 – Resultado da busca.....	94
Figura 19 – Menu de Visualizações.	94
Figura 20 – Financiamento por Agências de Fomento.....	95
Figura 21 – Financiamento por Agências de Fomento - CNPq.....	96
Figura 22 – Financiamento e produção científica por ano.	97
Figura 23 – Financiamento e produção científica por ano – ano 2014.....	98
Figura 24 – Financiamento e produção científica por ano – um pesquisador.	98
Figura 25 – Financiamento e produção científica por região demográfica – 2005 a 2014	99
Figura 26 – Rede de Coautoria – Exemplo: ano 2005, ranking 1.	100
Figura 27 – Rede de parcerias institucionais – Exemplo Fiocruz, ano 2005.....	101
Figura 28 – Modelo do padrão público de financiamento.....	105
Figura 29 – Pesquisadores identificados com financiamento.....	111
Figura 30 – Distribuição geográfica do financiamento e produção científica no Brasil	114
Figura 31 – Parcerias Institucionais formadas por meio dos pesquisadores	115
Figura 32 – Modelo de extração de dados do DOU sobre o financiamento para pesquisa	116
Figura 33 – Notícia do D.O.U. referente a financiamento	125

LISTA DE QUADROS

Quadro 1 – Verba arrecadada e empenhada no CT-SAÚDE de 2002 a 2013.....	43
Quadro 2 – Bibliotecas de funções Python usadas para a coleta dos dados.....	61
Quadro 3 – Formato do arquivo salvo	67
Quadro 4 – Dados recuperados do DOU referentes às agências de financiamento federal	74
Quadro 5 – Base para identificação das agências.....	75
Quadro 6 – Passagem de texto com código de identificação “5” para agência de fomento.....	75
Quadro 7 – Tipos de entidades nomeadas	80
Quadro 8 – Entidades nomeadas	82
Quadro 9 – Quantidade de páginas do D.O.U. – corpus	102
Quadro 10 – Quantidade de passagens recuperadas contendo agências de fomento	103
Quadro 11 – Quantidade de passagens de texto pelas agências de fomento na Seção 3.....	104
Quadro 12 – Distribuição da classificação manual	106
Quadro 13 – Comparação entre os métodos de classificação (Naive Bayes e <i>tf-idf</i>)	108
Quadro 14 – Pares de classificação que contém RES	108
Quadro 15 – Lista das 20 palavras com maior frequência para cada classificação.....	109
Quadro 16 – Financiamento recebido entre 2005 e 2014 pelas instituições (<i>top 10</i>).....	113
Quadro 17 – Auxílio a Pesquisadores pela CAPES nos anos de 2005 a 2014.....	119
Quadro 18 – Financiamento do CNPq para pesquisa segundo CNPq – 2006 a 2014	122
Quadro 19 – Financiamento do CNPq para pesquisa segundo DOU – 2006 a 2014	123
Quadro 20 – Pesquisadores utilizados para conferência dos dados do DECIT.....	124

LISTA DE GRÁFICOS

Gráfico 1 – Financiamento DECIT e parceiros para pesquisa entre 2004 e 2014	45
Gráfico 2 – Financiamento DECIT e parceiros para pesquisa em dengue entre 2004 e 2013	46
Gráfico 3 – Frequência do termo “Aedes Aegypti” No D.O.U. entre 2005 e 2014	102
Gráfico 4 – Frequência do termo “Dengue” No D.O.U. entre 2005 e 2014.....	103
Gráfico 5 – Frequência do termo “Mosquito” No D.O.U. entre 2005 e 2014.....	103
Gráfico 6 – Quantidade de passagens recuperadas contendo agências de fomento nos anos	104
Gráfico 7 – Distribuição da classificação utilizando Naive Bayes	106
Gráfico 8 – Distribuição da classificação utilizando tf-idf	107
Gráfico 9 – Precisão, revocação e F1-Score dos métodos Naive Bayes e TF-IDF	107
Gráfico 10 – Distribuição da classificação normalizada ao longo do tempo.....	109
Gráfico 11 – Produção sobre dengue indexada no Google Scholar ao longo do tempo	111
Gráfico 12 – Financiamento em nível federal para pesquisa em dengue.....	112
Gráfico 13 – Financiamento em nível federal e produção acadêmica.....	112
Gráfico 14 – Auxílio a pesquisadores (CAPES) e financiamento para pesquisa em dengue	119
Gráfico 15 – Auxílio a pesquisadores (CAPES) e financiamento para pesquisa em dengue	120
Gráfico 16 – Financiamento para pesquisa em dengue – DOU x G-Finder.....	120
Gráfico 17 – Financiamento para pesquisa em dengue – Decit e Finep	121
Gráfico 18 – Quantidade de artigos indexados no Google Scholar sobre dengue	122
Gráfico 19 – Representatividade do financiamento para pesquisa em dengue do CNPq por áreas do conhecimento e ano.....	123

LISTA DE SIGLAS

APF	Administração Pública Federal
AWR	Recebimento de Financiamento – classificação dos parágrafos
BNDES	Banco Nacional de Desenvolvimento Econômico e Social
C,T&I	Ciência, Tecnologia e Inovação
CALL	Chamada para financiamento – classificação dos parágrafos
CAPEX	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CAPTCHA	Completely Automated Public Turing Test To Tell Computers and Humans Apart
CIDE	Contribuição de Intervenção no Domínio Econômico
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COHRED	Council on Health Research for Development
CON	Prestação de contas – classificação dos parágrafos
CP2T	Convert PDF to Text
CPC	Ciência da Política Científica
CT-SAÚDE	Fundo Setorial da Saúde
CV	Curriculum Vitae
DECIT	Departamento de Ciência e Tecnologia
DIA2	Deep Insights Anytime, Anywhere
DOU	Diário Oficial da União
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
FAPEAM	Fundação de Amparo à Pesquisa do Estado do Amazonas
FAPEG	Fundação de Amparo à Pesquisa do Estado de Goiás
FAPERGS	Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul
FAPESB	Fundação de Amparo à Pesquisa do Estado da Bahia
FAPs	Fundações de Amparo à Pesquisa
FHD	febre hemorrágica da dengue
FINEP	Financiadora de Estudos e Projetos
Fiocruz	Fundação Oswaldo Cruz
FNDCT	Fundo Nacional de Desenvolvimento de Científico e Tecnológico
FUNASA	Fundação Nacional de Saúde
FVA	Fundo Verde-Amarelo
HTML	HyperText Markup Language
ICTs	Instituições de Ciência e Tecnologia
INPA	Instituto Nacional de Pesquisas da Amazônia
IOB	Inside, Outside, Begin
ISI	Institute for Scientific Information
JSON	JavaScript Object Notation
KDD	Knowledge Discovery and Data Mining
LN	Linguagem Natural
LOA	Lei Orçamentária Anual
MAPA	Ministério da Agricultura, Pecuária e Abastecimento
MCT	Ministério da Ciência e Tecnologia
MCTI	Ministério da Ciência, Tecnologia e Inovação
MEC	Ministério da Educação
MS	Ministério da Saúde
NER	Named Entity Recognition
NLTK	Language Tool Kit
NSF	National Science Foundation
OCC	orçamento para capital e consumo
OCR	optical character recognition

OPAS	Organização Pan-Americana da Saúde
P&D	Pesquisa e Desenvolvimento
PDF	Portable Document Format
PDP	Partnership for Development Product
PIL	Python Image Library
PLN	Processamento de Linguagem Natural
PMCD	Programa Municipal de Controle da Dengue
PNCTIS	Política Nacional de Ciência, Tecnologia e Inovação em Saúde
POS	Part of Speech
PRO	aditivo para o financiamento
RES	publicação de resultado
SGBD	sistema gerenciador de banco de dados
SIOP	Sistema Integrado de Planejamento e Orçamento
SIORG	Sistema de Informações Organizacionais do Governo Federal
SNIS	Sistema Nacional de Inovação em Saúde
SO	Sistema Operacional
SOSP	Science of Science Policy
SUS	Sistema Único de Saúde
tf-idf	term frequency–inverse document frequency
TICs	Tecnologias de Informação e Comunicação
TXT	Texto
UK	Reino Unido (United Kingdom)
Unesco	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
USA	Estados Unidos da América (United States of America)
USP	Universidade de São Paulo
VA	Visual Analytics
VIVA	Vancouver Institute for Visual Analytics
WSL	Web ScriptLattes
XML	Extendable Markup Language

SUMÁRIO

1. Introdução.....	13
2. Um Novo olhar sobre a política científica.....	16
3. Pesquisa em saúde: a perspectiva global.....	26
4. Pesquisa em Saúde no Brasil.....	32
4.1. O caráter estratégico da pesquisa em dengue.....	47
5. Big data: para além de um grande conjunto de dados.....	52
6. Objetivos.....	57
7. Metodologia.....	58
7.1. As fontes de dados e os instrumentais utilizados.....	58
7.2. Passos da Pesquisa.....	64
8. Resultados.....	90
8.1. Ferramentas tecnológicas.....	90
8.1.1. CP2T: convertendo de PDF para TXT.....	90
8.1.2. A base de dados.....	90
8.1.3. FarejaDOU: A ferramenta de busca do DOU e visualização.....	92
8.2. Os dados identificados.....	102
8.3. Modelo de extração de dados do DOU sobre o financiamento para pesquisa.....	115
9. Discussão.....	117
9.1. Financiamento para pesquisa em dengue no Brasil.....	117
9.2. Limites e possibilidades.....	126
REFERÊNCIAS.....	130
ANEXO 1- Texto original para submissão do CV no Canadian Common CV.....	136
ANEXO 2- Dicionário de dados das tabelas do banco de dados.....	138

1. Introdução

O século XXI presencia outro perfil de ciência. De fato, a ciência contemporânea cresceu de tal monta que acaba sendo vítima de sua própria dinâmica, senão do seu avanço. A produção de dados e informação científica alcançou um volume tal que, desde o final do século XX, a sociedade passou a ser denominada como “da informação” e “do conhecimento”, e a ciência passou a ser *e-science*, desterritorializada e colaborativa e, em alguns casos, invisível, uma vez que os canais de e para publicização não têm o mesmo dinamismo. Mineração, ou a busca do valioso em meio à lama e o cascalho, passa a ser uma atividade nobre. Aparentemente a quantidade de dados suscitaria pensar que, em um “estalar de dedos” se consegue a informação ou dado correto para a tomada de decisão em face de um determinado problema. Entretanto, o aumento no volume de dados e a velocidade em que eles ganham visibilidade se tornam mais um desafio para as políticas públicas. Se não há mais linearidade entre investimento em ciência e bem estar social, a política científica tem que buscar novas bases para avaliar o que existe e orientar, de forma mais cuidadosa, quando não pontual, novos investimentos em ciência, tecnologia e inovação.

Frente a isso, o termo Ciência da Política Científica surge como explicitação da nova abordagem da Política Científica que usa de bases científicas para analisar a própria Política Científica, de sorte a instrumentalizar os *policy makers* a estruturarem políticas com base em evidências construídas a partir de métodos científicos. Ciência da Política Científica, *do inglês Science of Science Policy* (SOSP), define um novo campo do conhecimento, em construção que, de acordo com Valdez e Lane (2008), tem o objetivo de prover, a partir de pesquisa interdisciplinar, bases quantitativas de rigor científico para formuladores de políticas e pesquisadores com o intuito de analisar a dinâmica da pesquisa e avaliar seus resultados.

De fato, o Estado tem a preocupação de produzir e analisar dados de qualidade que proporcionem um olhar mais qualificado para o direcionamento dos recursos investidos na pesquisa. Isso se dá não só em um ambiente de recursos financeiros cada vez mais escassos para custear pesquisas cada vez mais grandiosas e dispendiosas, mas também com uma crescente preocupação social sobre os caminhos que a ciência vem tomando, especialmente nas respostas que ela não mais consegue apresentar.

Se são várias as perguntas (especialmente, quão eficiente e efetivo tem sido o financiamento público da ciência?) e poucas ainda são as soluções apresentadas. Algumas iniciativas vêm sendo desenvolvidas, como o *Deep Insights Anytime, Anywhere* (DIA2 – <http://www.dia2.org/>), nos Estados Unidos da América (EUA), para suprir esta necessidade.

O DIA2 é uma iniciativa da National Science Foundation (NSF), agência de financiamento dos EUA em conjunto com quatro grandes universidades daquele país: Arizona State University, Purdue University, George Mason University e Stanford University. A NSF disponibilizou seus dados de financiamento para que as quatro universidades, a partir de um projeto de pesquisa, construíssem uma ferramenta de visualização interativa que auxiliasse os tomadores de decisão assim como os pesquisadores a analisarem os dados sobre financiamento para pesquisa e sua relação com as publicações e resultados proveniente destes financiamentos.

No Brasil, porém, iniciativas dessa natureza estão ainda longe de acontecer. O financiamento é quase majoritariamente realizado pelo governo por meio de suas agências de fomento. Em nível federal, o Brasil conta com o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a Financiadora de Estudos e Projetos (FINEP) e, no caso da saúde, o Departamento de Ciência e Tecnologia (DECIT do Ministério da Saúde). Estas agências apresentam os dados sobre o financiamento realizado isoladamente umas das outras, e em níveis de detalhamento diferentes, o que dificulta agregar e analisar os dados.

O Brasil, no entanto, conta com uma fonte de dados onde são publicados todos os financiamentos para a pesquisa de nível federal, o Diário Oficial da União (DOU). O DOU, criado em 1808 teve como objetivo principal dar transparência às ações da família real. Esta premissa é mantida até hoje, ou seja, dar publicidade sobre tudo o que o Estado faz, entre outros temas, o financiamento das atividades de ciência e tecnologia. O DOU é organizado em três Seções principais numeradas. A Seção 1 apresenta as leis e decretos; a Seção 2 apresenta os atos da Administração Pública Federal (APF) e a Seção 3 publica as chamadas e os resultados de editais de financiamento, além de compras e editais de concursos. O DOU é disponibilizado na internet em formato PDF por páginas isoladas, podendo ser encontrado jornais datados do ano de 1990.

Como todo o financiamento da esfera federal é publicado no DOU, este passa a ser uma fonte de dados valiosa e inequívoca.

Esta tese aposta no Diário Oficial da União (DOU) como fonte qualificada de dados para análise do financiamento de pesquisa em saúde no Brasil, mais especificamente, na área de dengue. Para isso, este trabalho teve como objetivo principal o desenvolvimento de uma metodologia de extração de dados do DOU com a finalidade de explicitar o padrão de

financiamento para pesquisa em dengue no período entre 2005 e 2014 de modo a suportar os tomadores de decisão na análise das políticas públicas.

Para alcançar tal objetivo, o presente trabalho está assim organizado: o capítulo a seguir, capítulo 2, discorre sobre a Ciência da Política Científica, a busca por bases sólidas e confiáveis e como o Diário Oficial da União é potencial fonte de dados para suportar as necessidades dos *policy makers*. A seguir, o capítulo 3, aborda a situação global da pesquisa em saúde no mundo, apresentando a dificuldade enfrentada em agregar dados sobre o financiamento para pesquisa e ressaltando a necessidade de sistemas que monitorem os gastos com pesquisa em saúde. O capítulo 4 segue com o histórico do financiamento para pesquisa em saúde no Brasil, apresentando a criação do Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT), resgatando que a origem do financiamento no Brasil é prioritariamente público e na seção 4.1 deste capítulo aponta para o caráter estratégico da pesquisa em dengue no Brasil. Em seguida, o capítulo 5, descreve as características de *big data* assim como apresenta que o *Visual Analytics* é uma boa perspectiva para trazer sentido aos dados. O capítulo 6 traz os objetivos que guiaram todo o desenvolvimento desta tese. Em sequência, o capítulo 7 descreve as fontes de dados e os instrumentais utilizados na metodologia, assim como todo o percurso trilhado. Os resultados, capítulo 8, foram organizados de sorte a apresentar os produtos tecnológicos frutos da metodologia, os dados identificados e o modelo de extração de dados do DOU sobre o financiamento para pesquisa. Por fim, o capítulo 9, discute os resultados sobre o financiamento para pesquisa em dengue no Brasil, assim como apresenta limites encontrados e possibilidades futuras.

2. Um Novo olhar sobre a política científica

Há muito se discute o papel da ciência e tecnologia no desenvolvimento econômico de um país. A partir de meados do século passado, considerava-se a ciência como uma fábrica em seu conceito “fordista” e linear, onde se injeta recurso no início da linha de produção, em seguida os operários-cientistas desenvolvem suas pesquisas e, como última etapa, e como consequência ao trabalho desenvolvido, eram publicizados os resultados, particularmente sob a forma de artigos científicos, ou em alguns casos, como patentes, ambos os produtos registros de conhecimento fundamentais para orientar a produção de bens e serviços de valor social. Às políticas de ciência e tecnologia caberiam, principalmente, investir em ciência como uma dimensão fundamental para o alcance do bem estar social. À Política Científica, cabia, então, fomentar a ciência, uma vez que ela se encarregaria de, seguindo uma cadeia linear de eventos, conduzir ao desenvolvimento socioeconômico.

A partir das últimas décadas do século passado, esse modelo linear passou a não mais responder às demandas colocadas pelo desenvolvimento dos países, especialmente dentro de um processo de globalização econômico e frente à (re)evolução tecnológica. Guimarães (1998, p.27) aponta várias explicações para a exaustão do modelo linear, e a consequente proposição de um modelo iterativo. A autora relata que, durante décadas, dentre os *policymakers* havia “uma crença que investimentos em ciência redundavam, de forma inequívoca, em crescimento econômico”. A autora ainda acrescenta que vários estudos, alguns clássicos, se debruçaram na busca de um modelo que ligasse a ciência ao desenvolvimento. Especificamente aqueles estudos de caso voltados para as inovações na área militar identificaram que as conquistas científicas pouco ou nada influenciaram as referidas inovações. Entretanto, o *Technology In Retrospect and Critical Events In Science* (TRACES), um estudo com foco em inovações não militares, tidas como econômica e socialmente importantes, chegou a uma conclusão oposta, identificando uma clara influência dos eventos científicos no processo de inovação. A autora conclui que isso não é nenhuma surpresa resultados tão opostos, pois os estudos utilizaram de uma metodologia que, embora semelhante em alguns aspectos, era substancialmente diferente. (GUIMARÃES, 1998). A questão de fundo, para todos os estudos, era a busca pela raiz que garantisse o desenvolvimento tecnológico.

Guimarães (1998) tece um panorama histórico sobre a evolução tecnológica apresentando que a tecnologia **não** é a ciência aplicada; e que modelo linear não tem como ser a estrita explicação da inovação tecnológica. O modelo deveria considerar diversos aspectos e

características que o tornam multifacetado, considerando que as tecnologias compreendem pelo menos três dimensões: conhecimento, *skill* e artefato. Uma dimensão fundamental nesse tripé é o conhecimento tácito, que, segundo Polanyi (1969, p.4), seria melhor traduzido por “(...) sabemos mais do que podemos contar”.

Em contraponto ao modelo linear começou então a ser introduzido um modelo interativo do processo de inovação. Tratava-se, agora, não só do desenvolvimento de tecnologias, mas do processo de inovação, ou seja, da adoção das tecnologias no meio social. Este modelo, chamado de *chain-linked model* foi apresentado por Stephen J. Kline em 1985, e traz à tona a ideia de trabalho cooperativo sem haver necessariamente uma sequência formal entre os atores deste processo produtivo:

a inovação passou a ser vista como um processo onde se integram conhecimentos vindos de várias fontes; um processo que demanda que à uma disponibilidade de conhecimento científico e tecnológico se somem oportunidades tecnológicas e de mercado. (GUIMARÃES, 1998)

Ruivo (1994) apresenta este mesmo conceito como “modelo complexo associando oferta e demanda” e traduz o paradigma da ciência como “fonte de oportunidade estratégica”.

Guimarães (1998) relata que

“as descrições oferecidas por este autor [Edward Constant, 1980] sobre as comunidades de praticantes, e como elas se estruturam internamente à uma organização; de uma prática de resolução de problemas, e do *locus* cognitivo das tecnologias podem ser amplamente utilizadas para entender **uma faceta** do desenvolvimento tecnológico” [grifo adicionado] (GUIMARÃES, 1998, p.19)

A partir daqui, a Política de Ciência, Tecnologia e Inovação abraça toda a complexidade do processo de inovação. Ainda que o fomento à pesquisa continue sendo o grande vetor da dinâmica do empreendimento científico, inúmeras outras dimensões passaram a ocupar a agenda dos tomadores de decisão.

Fealing (2011) aponta que a Ciência da Política Científica não se pode esquivar de prover entendimento acadêmico, mas deve ir além da discussão teórica, levantando e sistematizando um conjunto de dados da pauta da formulação de políticas. O autor ressalta ainda que esta não é uma ciência simples, e tampouco exata, e que políticas possuem impactos intencionais e não intencionais na sociedade. A ciência ocorre em um sistema complexo composto por pesquisadores, grupos de pesquisas (intra e interinstitucionais), instituições de pesquisa públicas e privadas, agências de financiamento, desenvolvedores de equipamentos científicos, unidades de política do governo, editoras científicas e a própria sociedade. Dispor de dados que permitam uma análise mais orgânica desse conjunto de atores e respectivos fluxos é o desafio em foco, em perspectiva internacional.

Entretanto, dispor de um conjunto de dados sobre o financiamento público de ciência está muito longe de ser trivial. Alguns países, como os Estados Unidos da América tentam, por meio de infraestruturas tecnológicas, reunir os dados sobre gastos públicos em ciência e tecnologia, relacionando o investimento com seus resultados (LANE, 2009). Lane (2009) aponta ainda para algumas ferramentas que vêm sendo propostas para auxílio na construção e manutenção das políticas: (i) um portal unificado que informa as bolsas concedidas pelo governo (<http://grants.gov>), (ii) portais que relacionam os resultados da pesquisa com bolsas específicas (<http://research.gov> e <http://science.gov>) e (iii) portais que promovem a participação da sociedade no processo decisório por meio da disponibilização dos dados públicos (<http://data.gov>).

Tais iniciativas tomaram força e forma a partir do compromisso assumido por Barack Obama, presidente dos Estados Unidos, já no início de seu mandato em 2009. A partir de um memorando distribuído para os chefes dos departamentos e agências de seu governo, o presidente Obama assumiu a importância de se ter um governo aberto e transparente. Tal abertura a qual ele se referiu se baseia na participação do cidadão, a qual só se torna possível se este tiver acesso aos dados. É neste momento que governo transparente é fundamental para a colaboração do cidadão, pois tal transparência promove responsabilização por parte do governo. (USA, 2009)

A transparência dos dados não implica somente na participação do cidadão e na responsabilização por parte do Governo. Ela também possibilita o cruzamento de dados oriundos de diversas fontes, suscitando a possibilidade de se saber se o investimento que o Estado tem feito na ciência está sendo capaz de produzir bem estar social. Entretanto, medir o bem estar social é complexo e ainda não se tem um consenso de como se fazer isso. Portanto, vincular financiamento com os produtos clássicos da ciência ainda parece o mais adequado.

Fealing (2011) acrescenta à visão de Lane (2009) que a política é determinada pelo julgamento humano aplicado a um contexto e, para tanto, necessita de um *dashboard* analítico contendo dados sobre o financiamento à pesquisa. Ambas as autoras citadas anteriormente (FEALING, 2011; LANE, 2009) acrescentam ainda que os padrões de financiamento emergem a partir da análise de um conjunto de dados crescente e sofisticado; o termo sofisticado aparece neste contexto para indicar a dificuldade na coleta e na correlação dos dados. Fealing (2011) tece considerações a respeito do uso de planilhas eletrônicas chamando atenção para os riscos dos mesmos como instrumentos para a tomada de decisão, pois estas podem não revelar dados importantes dentro de tantas linhas e colunas. As planilhas

eletrônicas aplicadas ao contexto eram – em muitos casos ainda são – a forma mais comum de ferramenta para a tomada de decisão. Entretanto, este tipo de método prescinde de decisões prévias que definem as variáveis que serão utilizadas nas referidas planilhas, muitas vezes antes de fazer o levantamento. Tal fato pode direcionar os achados e não contribuir com novas descobertas (FEALING, 2011).

A busca por bases teóricas e empíricas que deem conta do embasamento necessário para o processo decisório e, por consequência, a construção e avaliação das políticas públicas em ciência e tecnologia, é um movimento relativamente tímido no mundo. Raramente se encontram bases abertas de avaliação científica como Deep Insights Anytime, Anywhere (DIA2¹). O DIA2, produto desenvolvido por meio da cooperação entre as universidades Arizona State University, Purdue University, George Mason University e Stanford University é uma iniciativa que foi apoiada pela *National Science Foundation* (NSF), nos Estados Unidos da América.

A NSF, uma agência federal independente, é responsável por cerca de 24 % do financiamento à pesquisa realizadas naquele país. Essa agência tem dois papéis neste projeto: a de financiador do projeto e de cliente do projeto. Enquanto cliente do projeto, a NSF tem a necessidade de se vincular o que estava sendo financiado com o que era produzido a partir de seu financiamento. O produto promoverá a unificação dos dados sobre financiamento à pesquisa realizada pelas diversas agências financiadoras naquele país, o que acarretará, entre outros, na diminuição da sobreposição de financiamentos e em uma melhor efetividade na escolha de onde aplicar os recursos financeiros.

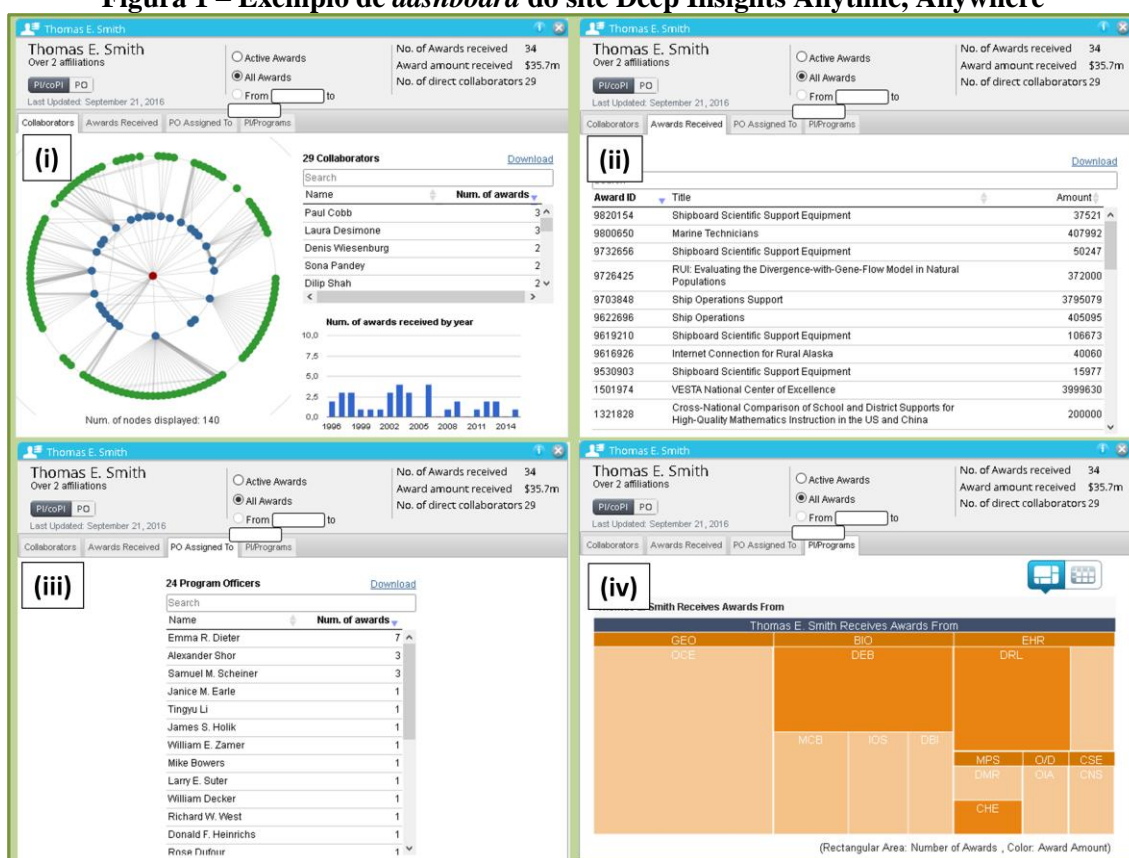
O DIA2 é de acesso aberto, bastando criar uma conta de usuário para iniciar as buscas. O produto permite a criação dinâmica de *dashboards* os quais são representações gráficas de visões distintas dos dados. Os *dashboards* permitem a interação por parte do usuário entre um gráfico e outro, ou seja, ao selecionar um determinado dado, os gráficos relacionados ao dado são automaticamente atualizados. Estes *dashboards* propiciam ao usuário do sistema uma visão ampliada do financiamento no que diz respeito ao quanto foi financiado e quem, instituição e/ou pesquisador, recebeu o financiamento. Este tipo de cruzamento dos dados permite uma visão multidimensional e simplificada do objeto de estudo.

A seguir é apresentada a visão de um *dashboard* do DIA2 iniciada pela visão de um pesquisador. Cada um dos quadros da Figura 1 está numerado de acordo com a descrição abaixo:

¹ DIA2 disponível em <http://dia2.org>

- (i) a quantidade de financiamentos recebidos, o valor total, a rede de colaboração com a quantidade de financiamentos recebidos por cada um dos colaboradores e a distribuição dos financiamentos ao longo do tempo. O *dashboard* possui ainda três outras abas que permitem
- (ii) visualizar a lista de financiamentos recebidos pelo pesquisador contendo ID, título e valor de cada uma,
- (iii) a lista de programas ao qual os projetos estão associados e
- (iv) um *tree map* que apresenta de onde veio o recurso.

Figura 1 – Exemplo de *dashboard* do site Deep Insights Anytime, Anywhere



Fonte: DIA2 (2016)

A construção de uma ferramenta de tal porte não é simples. Para tanto, o primeiro passo necessário é identificar fontes capazes de subsidiar a Política Científica com dados confiáveis e valiosos para que os tomadores de decisão possam se embasar. É preciso trazer estas fontes para uma mesma central de inteligência onde se tenham diferentes visões do processo de todo do desenvolvimento científico, contemplando o financiamento do mesmo e a produção decorrente.

O processo de identificação e análise dos dados pertinentes ao financiamento para a pesquisa, assim como a quantificação/qualificação da produção científica e tecnológica torna-

se, assim, um instrumento importante para diversas finalidades: orientar políticas; direcionar financiamento; desenvolver indicadores de produtividade científica, e orientar atividades de avaliação de desempenho.

A meta de construção de ambientes informacionais inteligentes que permitam o cruzamento dos dados extraídos de diferentes bases, conforme visto, é perseguida em todo o mundo. Os ganhos estão relacionados à obtenção de variáveis capazes de traduzir as diversas situações possibilitando a avaliação e monitoramento destes processos, o que traz continuidade à ação.

Entende-se que esta ideia ainda está distante da ideal, onde se deveria verificar o resultado social (impacto social), ou seja, verificar se aquele processo iniciado pelo financiamento, tendo a pesquisa como motor do desenvolvimento de uma determinada tecnologia e seus resultados fora absorvido pela sociedade e teve de fato a mudança prevista. Entretanto, mesmo a construção de ambientes informacionais descrita no parágrafo acima se torna uma peça fundamental na avaliação da ciência.

Ademais, conforme descrito no início deste, acrescenta-se ao aumento no volume de dados e velocidade em que eles aparecem, a variedade de fontes de informação a serem cruzadas. Neste sentido torna-se cada vez mais difícil a captura, tratamento e análise dos dados com vista a apresentá-los aos tomadores de decisão.

O movimento de mapeamento das fontes e o esforço de trazê-las a um ambiente comum onde possam ser analisadas e subsidiarem os tomadores de decisão, no sentido de avaliarem o fluxo de financiamento para a pesquisa é um objetivo almejado em todo o mundo. O próximo capítulo aborda o esforço internacional neste sentido.

Embora a preocupação seja crescente nos países centrais, eles ainda não têm essa contabilidade de forma organizada, clara e óbvia. Isso se deve, principalmente, pelo modelo e estrutura do fluxo de financiamento do Estado, onde os recursos financeiros fluem para diferentes agências de fomento, nacionais e locais, e cada uma delas tem uma lógica própria de operá-los. No modelo clássico de Política Científica, o Estado confia às agências a decisão de distribuir os recursos de forma a maximizá-los em termos de eficiência, eficácia e efetividade, sendo que o resultado desse investimento, por excelência, deve ser atestado por *rankings* de produtividade científica. O primeiro movimento que tem sido feito, em perspectiva internacional, é o desenvolvimento de estratégias que permitam a identificação, organização e visualização do conjunto de dados relacionados aos recursos financiados à ciência pelo Estado.

Este cenário não difere entre países centrais e periféricos. A mesma preocupação de mapear os dados sobre financiamento à pesquisa está presente no Brasil. É latente a necessidade de verificar se os recursos investidos estão sendo convertidos em benefício para sociedade. Esta é uma preocupação cada vez mais central no campo da pesquisa em saúde, por razões óbvias: as iniquidades em saúde são determinantes da capacidade de desenvolvimento dos países.

No Brasil, assim como nos países centrais, o fomento realizado pelo Governo se dá por meio das agências de fomento. Os recursos do Tesouro são distribuídos diretamente e principalmente para o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), para a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e para Financiadora de Estudos e Projetos (FINEP), especialmente por meio dos Fundos Setoriais. O Brasil também conta com agências de financiamento que atuam em nível estadual, as Fundações de Amparo à Pesquisa, as quais possuem um reconhecido papel neste cenário. Este conjunto de agências e fundos serão detalhados no Capítulo 3, porém ressalta-se aqui que as FAPs não serão consideradas no presente trabalho.

Estas inúmeras agências de fomento trabalham de forma independente, o que acarreta grande dispersão dos dados sobre o financiamento. Do melhor do conhecimento disponível, não está disponível qualquer dispositivo que permita a sistematização dos mesmos. Por outro lado, é também reconhecido, em perspectiva internacional, uma plataforma que congrega todo o conjunto de dados sobre os pesquisadores brasileiros (e aqueles que atuam no país) e, por decorrência, um potencial registro sobre as pesquisas desenvolvidas no Brasil. O Currículo Lattes é parte da Plataforma Lattes, a qual é um conjunto de sistemas de informação, bases de dados e portais da internet, concebido para integrar os sistemas de informação das agências federais, racionalizando o processo de gestão de Ciência e Tecnologia.

Compõem a Plataforma: o Currículo Lattes, Diretório de Instituições, Diretório dos Grupos de Pesquisa. O Currículo Lattes é um aplicativo que surgiu para apoiar a identificação e troca de informação entre pesquisadores brasileiros. Segundo o CNPq, que administra atualmente a Plataforma Lattes, desde os anos oitenta, identificou-se a necessidade de registro de currículos dos pesquisadores brasileiros por meio de um formulário padrão que proporcionasse informação sobre a distribuição da pesquisa científica no Brasil. Atualmente, o Currículo Lattes possui cerca de 1,9 milhões de currículos de pesquisadores e totaliza mais que três milhões de currículos, aqui considerando também os discentes registrados no sistema.

Esta demanda não parece estar localizada apenas no Brasil, ainda que este detenha o pioneirismo. Recentemente, Portugal e Canadá lançaram sistemas semelhantes ao Currículo Lattes, chamadas respectivamente de DeGóis e Common CV. No Canadá, o Common CV foi incorporado em 2015 e se tornou obrigatório para os professores universitários que desejam concorrer a financiamento. Entretanto, diferentemente dos outros dois países, esta base é fechada para consulta pública, pois o mesmo está sob a política de privacidade determinada pelo “Privacy Act” e outras políticas do governo Canadense. (Vide Anexo 1)

Segundo Mena-Chalco e Cesar Junior (2009), as instituições de ensino e pesquisa no Brasil se utilizam constantemente do Currículo Lattes para extrair dados a fim de gerar relatórios de sobre produção científica, orientações e avaliação da produção de grupos de pesquisa. Os autores acrescentam que se trata de uma tarefa árdua, pois os dados não estão estruturados. Ademais, a tarefa se torna suscetível a erros acarretando no tratamento manual dos resultados.

A Plataforma Lattes já tem um dispositivo que permite alguns cruzamentos, chamado de Painel Lattes. Entretanto, este dispositivo não proporciona, por exemplo, o cruzamento de dados com bases externas à Plataforma Lattes nem tampouco a visualização específica de dados importantes para as instituições, como o monitoramento da tipologia da produção de seus profissionais. Um dispositivo existente para extração e compilação automática dos dados do Currículo Lattes é o *scriptLattes*. Esta ferramenta é customizada para extrair dados relacionados à produção científica. Neste sentido pode se dizer que o *scriptLattes* avança na direção da extração guiada, mas também não realiza cruzamento com outras bases de dados de informação.

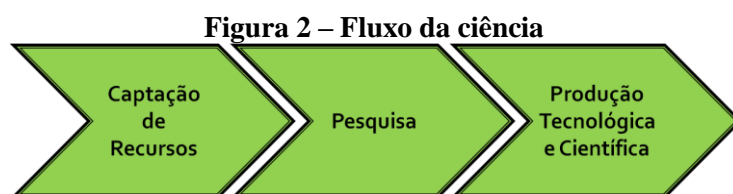
Esta base de currículos é a principal fonte para mapear produção científica dos pesquisadores nacionais. Porém, encontram-se dois grandes desafios na utilização destes dados, a saber: (i) o currículo é autodeclaratório e (ii) nem todos os dados digitados pelos pesquisadores se tornam públicos. O fato de o currículo ser autodeclaratório suscita que alguns dados sobre a produção do pesquisador possam não ser inseridos pelo mesmo, quando este julgar não ser importante ou mesmo se esquecer de colocar. O segundo item apontado é o que mais afeta na busca pela sistematização dos dados sobre financiamento. O Lattes possui também um campo para registro do financiador da pesquisa assim como valor do financiamento. Porém, o que poderia ser tomado como uma fonte potencial para coletar dados sobre o financiamento da pesquisa no Brasil não se torna possível, pois além do montante registrado não ser apresentado na visualização pública do CV, a ferramenta não possibilita a

busca por estes campos. Alguns pesquisadores se valem dos campos abertos, como “descrição do projeto”, declarando qual agência de financiamento está suportando o projeto. O campo descrição do projeto não é pesquisável implicando na dificuldade de recuperar tal informação, além de tais dados se apresentarem em linguagem natural, o que torna a tarefa complexa.

Outra fonte de dados valiosa, porém pouca – ou nada – utilizada no estudo sobre o financiamento do setor público federal para a pesquisa no Brasil é o Diário Oficial da União (DOU). Sob a perspectiva do financiamento público da pesquisa, o DOU poderia ser tomado como a fonte de maior confiabilidade, tomando-a como a voz oficial do Estado, conferindo transparência a todas as ações e iniciativas no seu escopo de atuação.

O Diário Oficial da União (DOU) foi criado em setembro de 1808 sob o nome de Gazeta do Rio de Janeiro, em decorrência da vinda da Corte Portuguesa para o Brasil. O objetivo desta publicação era registrar os atos normativos e administrativos oficiais do governo, dando transparência às ações para o povo. Em se tratando de financiamento público federal à pesquisa, e tomando como primeira etapa a captação de recursos, ou seja, que um determinado pesquisador foi agraciado com fomento para sua pesquisa, destaca-se que todos os financiamentos federais à pesquisa (em saúde, inclusive) são legitimados com publicação no DOU.

A captação de recursos pelos pesquisadores, considerada como a primeira etapa do fluxo linear da ciência conforme descrito no início deste capítulo e demonstrado na Figura 2, se expressa com a publicação de uma chamada para financiamento no DOU. Seguida ao Edital Público com a chamada para a pesquisa em determinado tema, e decorrido um período de tempo variável estimado entre 3 e 6 meses, o DOU registra o resultado do certame. A distribuição da verba propriamente dita é publicada posteriormente, identificando o nome do pesquisador e o volume de recursos financeiros que lhe foi concedido.



Fonte: Elaboração própria

O DOU é uma fonte de dados pública e está disponibilizado eletronicamente no sítio da Imprensa Nacional. É possível encontrar no portal da Imprensa Nacional documentos datados desde o ano de 1990 até os dias de hoje. O DOU é dividido em três seções tendo, cada uma, foco específico: a Seção 1 traz leis, decretos, resoluções, instruções normativas, portarias e outros atos normativos de interesse geral; na Seção 2 constam atos de interesse dos

servidores da Administração Pública Federal, enquanto a Seção 3 possui contratos, editais, avisos e ineditoriais². As chamadas para financiamento assim como a assinatura da concessão de verba podem ser encontradas na Seção 3.

Se Diário Oficial da União (DOU) é uma fonte de dados potencial para coletar e analisar dados sobre financiamento público à pesquisa, especialmente no campo da saúde, segue-se uma pergunta central: qual o padrão de financiamento da pesquisa em saúde, em perspectiva nacional, a partir do DOU? Entender esse padrão é fundamental para analisar, de forma exploratória, se o DOU está estruturado de forma tal que permita uma extração de dados que expresse, pelo menos, parte do investimento nacional em pesquisa.

Mensurar o esforço de pesquisa em saúde, em perspectiva mundial, faz parte hoje da agenda de agências supranacionais, de organizações não governamentais, e mesmo dos países centrais. Isso porque, na perspectiva da saúde global, as doenças já não têm fronteiras, as estratégias para seus enfrentamentos devem ser maximizadas, especialmente pelo crescente caráter interdisciplinar do conhecimento em saúde. Nesse sentido, o próximo capítulo, então, descreve o esforço de consolidar e analisar os dados sobre financiamento para pesquisa em perspectiva internacional.

² Ineditoriais: diz-se da parte do jornal vendida para publicação de informação de terceiros

3. Pesquisa em saúde: a perspectiva global

O advento da “saúde global” trouxe também para a pesquisa em saúde a preocupação de um olhar amplo e abrangente para mapear e identificar a produção de conhecimento no tema, especialmente no que diz respeito às doenças negligenciadas. De forma óbvia, a saúde global alerta que não há mais limites seguros para separar ricos e pobres, desenvolvidos e em desenvolvimento, e que uma das estratégias mais valiosas de prevenção orienta para o fomento de pesquisa em perspectiva internacional.

O desconhecimento sobre que pesquisas vêm sendo desenvolvidas, por quem, a que custo, e em que temáticas é responsável, segundo especialistas, por um “buraco negro” nas políticas de saúde. (TERRY et al., 2014)

Nos países desenvolvidos, cerca de 40% das pesquisas em saúde são financiadas pelo Estado e por filantrópicas, percentual esse que é bastante mais expressivo nos países em desenvolvimento. (POLICY CURES, 2015)

Nesse sentido, um mapa das atividades de pesquisa e desenvolvimento (P&D) atuaria no aprimoramento da coordenação das atividades, minimizando duplicações e auxiliando na gestão de recursos escassos para atender às prioridades de saúde pública dos países. Assim, os dados sobre gastos com pesquisa e desenvolvimento compõem uma importante fonte de informação para o desenvolvimento das políticas de pesquisa em saúde, tanto para os países desenvolvidos como para aqueles em desenvolvimento. (YOUNG et al., 2015)

Burke e Francisco (2006) resgatam a importância de se monitorar os gastos com pesquisa em saúde e apontam cinco importantes motivos para tal: (i) monitorar nível de esforço e descrever tendências, (ii) identificar lacunas onde é necessária atuação, (iii) avaliar o impacto das políticas públicas frente ao aumento do investimento em pesquisa e desenvolvimento para saúde, (iv) usar esta informação para propósitos de *advocacy* e (v) promover o debate público sobre os caminhos da pesquisa.

Os autores apresentam o estudo realizado pelo *Council on Health Research for Development* (COHRED). Este estudo objetivou medir e monitorar os investimentos para pesquisa em saúde nos países de baixa e média renda. Burke e Francisco (2006) demonstraram que muito embora os resultados tenham sido úteis para os tomadores de decisão dos países estudados para aprimorar as respectivas políticas públicas, o estudo não atingiu a sustentabilidade necessária de medir e monitorar o financiamento para pesquisa em saúde de forma continuada. Os autores acrescentam que para que este estudo seja retomado,

há necessidade de reiniciar o processo, praticamente do zero, pois o conhecimento referente ao estudo não foi internalizado pelos países.

Segundo os mesmos autores, os dados até aqui disponíveis apontam claramente para a necessidade de sistemas que monitorem os gastos com pesquisa em saúde. Eles acrescentam que esta necessidade não se limita apenas aos países de baixa e média renda; trata-se de uma demanda global. Outro ponto importante levantado pelos autores é da obrigatoriedade de se ter um sistema sustentável, ou seja, que permita a coleta e apresentação de resultados sobre os gastos com pesquisa em saúde de forma contínua. Os especialistas reforçam ainda que a coleta de dados não se traduz em uma tarefa fácil e que a existência de dados confiáveis, de boa qualidade e autênticos sobre pesquisa em saúde e respectivos gastos é essencial para democracia, direito de cidadania e fundamental para o desenvolvimento, implantação, monitoramento e avaliação de políticas públicas. (TERRY et al., 2014)

Olson e colaboradores (2011) complementam a visão acima incorporando a importância de se ter em mãos os dados do financiamento para a pesquisa. Isso seria tão mais oportuno e apropriado em tempos da Internet, onde há uma grande facilidade para publicizar os dados, imprimindo transparência na gestão pública. A partir destes dados, afirmam os autores, é possível verificar se uma agência de financiamento, um determinado laboratório ou ainda um pesquisador, está fazendo bom uso dos recursos.

Entretanto, o foco no monitoramento dos recursos financeiros direcionados à pesquisa e seus produtos mais tangíveis deve ser tomado com cautela. Nos padrões atuais de avaliação, considera-se a publicação científica um dos resultados mais importantes de uma pesquisa, o que se torna também um indicador de qualidade para o resultado de um financiamento. Ora, essa é uma visão restrita do processo da pesquisa e do seu alcance e de seus resultados, para além de seus produtos. Outro aspecto importante nesta avaliação para o sistema de Ciência e Tecnologia é a inovação, traduzida por novos processos e principalmente pelas patentes.

Olson e colaboradores (2011) continuam descrevendo os limites da medição de *performance* do financiamento para a pesquisa. Estes limites se baseiam nos resultados da pesquisa e que estes, muitas vezes são incertos ou só aparecem em longo prazo. Outro limite descrito pelos autores é que muitas vezes, o resultado de uma pesquisa pode depender de ações políticas. Neste sentido, acrescentam ainda que há poucas evidências que demonstrem que os resultados de uma pesquisa possam ser tomados como guia para os processos decisórios, por meio dos tomadores de decisão.

Os mesmos autores defendem que a ciência política deveria ser capaz de responder o quanto em recursos financeiros deveria ser alocado para a pesquisa em nível federal, e em que temáticas e/ou áreas do conhecimento, tendo em vista as demandas sociais. É nessa perspectiva que a política científica fica à espera de uma ciência que possa contribuir para a melhor gestão e orientação da pesquisa. Entretanto, é preciso lembrar que não há como verificar qual a razão entre o recurso financeiro aplicado na pesquisa e o desenvolvimento proveniente deste.

Neste cenário, Young e colaboradores (2015) apontam para a dificuldade de se conseguir dados sobre gastos com a pesquisa e desenvolvimento. Os autores ilustram esta dificuldade na comparação de dois artigos que possuem por objetivo confrontar os gastos dos Estados Unidos da América com os gastos mundiais em pesquisa e desenvolvimento em saúde. Dada a fragilidade dos resultados e a limitação de seu alcance, os autores apontam a importância de encontrar fontes que deem conta de apresentar dados necessários e significativos relativos à saúde. Tais dados necessitam ainda serem garimpados em diferentes fontes, de forma que se possa extrair aqueles referentes ao financiamento à pesquisa e desenvolvimento em saúde.

Burke e Francisco (2006) também apresentam a dificuldade de capturar os dados sobre financiamento de forma correta, dentre os gastos em Ciência e Tecnologia. Os autores ressaltam que tal dificuldade é, provavelmente, fruto dos governos investirem em pesquisa e desenvolvimento em saúde como parte de atividades de Ciência e Tecnologia, ou seja, os dados ficam incluídos em uma grande rubrica, sem um detalhamento necessário.

Viergever e Hendriks (2016) corroboram com a visão descrita e acrescentam que grande parte do financiamento à pesquisa e desenvolvimento em saúde é realizada pelo setor público e por organizações filantrópicas. Adicionam ainda a esta visão o papel central no desenvolvimento do conhecimento e de novos produtos para saúde, em especial no combate de doenças negligenciadas. Os autores apresentam que se for possível organizar os dados (i) o que é gasto, (ii) com o que tipo de pesquisa é gasto e (iii) como é decidido para onde os recursos financeiros serão alocados, seria também possível às agências de financiamento (i) sincronizar seus esforços, (ii) mitigar as possíveis duplicações de financiamento e (iii) aumentar a possibilidade de colaboração nas pesquisas. A transparência destes dados também facilitaria o processo de tomada de decisão, sendo este importante particularmente ao setor público.

Algumas iniciativas internacionais vêm tentando mapear os recursos destinados à pesquisa em saúde, em especial G-FINDER³ e o DIA2⁴.

O G-FINDER, mantido por um conjunto de instituições filantrópicas e supranacionais, busca coletar todo o investimento realizado por organizações públicas, privadas e filantrópicas em pesquisa e desenvolvimento direcionadas para um conjunto de 36 doenças negligenciadas. Estas doenças são caracterizadas por afetarem países em desenvolvimento, e por não terem recursos financeiros ou a devida atenção pelas políticas públicas.

O G-FINDER foi desenvolvido a partir da aplicação de um questionário aos governos dos países, desde 2008, e a cada ano é produzido um relatório com foco distinto. Em 2008 o foco era tentar determinar o quanto realmente estava sendo investido e neste sentido 143 agentes financiadores em 43 diferentes países responderam à coleta de dados. Este estudo categorizou os investimentos por instituição, por faixa de valor investido, pela doença estudada suportada por aquele investimento. Naquela ocasião constava do levantamento 30 doenças negligenciadas. O estudo encontrou que 20% do investimento global foram realizados por instituições públicas e empresas privadas em programas internos, enquanto os outros 80% por financiadores externos à instituição financiada.

Os anos que se seguiram focaram em tendências de investimento (2009) onde foi encontrado, por exemplo, que os países Brasil e Índia estavam, na ocasião, ocupando o grupo de “Top 5” de governos investidores. Identificou-se também que o aumento de financiamento pela Fundação Bill e Melinda Gates (<http://www.gatesfoundation.org/>) mascarou a diminuição de financiamento em vários setores, principalmente o público. Já em 2010 a preocupação foi com a possibilidade de mudança na pesquisa e desenvolvimento em decorrência da crise mundial de financiamento. O relatório trouxe que, para aquele ano, houve um aumento significativo no financiamento pelo setor público, porém concentrado em apenas dois países (United States of America – USA e United Kingdom – UK). O ano de 2011 foi consequência do ano anterior em relação ao foco, pois o relatório refletiu em seu título a preocupação com a inovação. Esta preocupação foi confirmada pela diminuição no montante financiado em diversos setores.

O relatório de 2012 buscou identificar tendências e padrões que poderiam ser inferidas a partir do primeiro relatório. Naquele ano verificou-se que os 2/3 do financiamento global para as doenças negligenciadas são realizadas pelo setor público e que destes, os países em desenvolvimento são os que têm a maior fatia, 95.9%. Em 2013, ao mesmo tempo em que foi

³ <https://gfinder.policycures.org/>

⁴ <http://www.dia2.org/>

presenciado o maior corte do financiamento destinado às parcerias para o desenvolvimento de produtos (do inglês PDP, *Partnership for Development Product*) confirmou-se, mais uma vez, que o setor público participa com 2/3 (dois terços) do financiamento para estas doenças. Para 2014, o relatório apresentou uma grata surpresa, pois pela primeira vez desde 2008 houve um aumento no financiamento para o desenvolvimento de produtos. O último relatório (2015) mostrou que o investimento com doenças negligenciadas classificadas como *second tier* manteve-se estável. Doenças classificadas como *second tier*, onde, dentre outras, se encontra a dengue, são aquelas que recebem entre 1.0% e 10% do total de financiamento. Neste ano, porém, foi incluído no relatório o Ebola o qual fez com que o investimento total neste grupo subisse 23%.

Uma síntese da tabela de investimento em dengue mostra que o investimento para pesquisa e desenvolvimento para esta doença subiu de 52,4 para 87,4 milhões de dólares em termos mundiais, nos anos entre 2007 e 2014. Entretanto, o número de financiamentos subiu de 48 para 217. De forma clara, embora o número de financiamentos tenha subido mais que quatro vezes, o valor investido não chegou a dobrar. No Brasil, o total de investimento naquele período – 2007 a 2014 – para pesquisa e desenvolvimento em dengue foi da ordem de 22 milhões de dólares. Embora os valores investidos pelo Brasil nesta doença não variem muito de ano a ano, dois anos chamam a atenção: o ano de 2009 com 13,4 milhões, sendo este responsável por mais da metade do investimento dos oito anos e o ano de 2014 com cerca de apenas 350 mil dólares. Ressalta-se aqui que todo o investimento realizado pelo Brasil naquele período para P&D em dengue registrado no G-FINDER foi realizado pelo setor público.

Os dados apresentados pelo G-FINDER são declaratórios, o que significa que não necessariamente refletem a realidade, porém são únicos nessa esfera, e deverão ser tomados em conta quando do desenvolvimento da metodologia e da análise dos dados resultantes da presente tese.

Este tipo de movimento para se mapear o fluxo de investimento se mantém em constante mudança. Para o ano de 2016 foi planejado um suplemento⁵ especial do G-FINDER contendo apenas os investimentos com as pesquisas sobre o vírus da zika.

Outro importante movimento no sentido de mapear, coletar e cruzar os dados sobre financiamento para a pesquisa com a finalidade de apresentá-los para que sejam analisados e deem suporte aos tomadores de decisão é o Deep Insights Anytime, Anywhere (DIA2). A

⁵ Este suplemento não fora lançado até a finalização desta tese.

ferramenta foi lançada em 2011, a partir da colaboração entre a Universidade de Purdue, Virginia Tech, Universidade do Estado do Arizona e a Universidade de Stanford, ambas nos EUA em resposta às demandas da Fundação Nacional de Ciência (National Science Foundation – NSF) que além de ter financiado o projeto fornece os dados que alimentam a plataforma. Estes dados incluem, entre outros, pesquisadores, valores, rede de colaboração e áreas de financiamento. Esta ferramenta tem por objetivo apresentar os dados de uma forma que possibilite uma rápida visualização dos mesmos não sendo necessário nenhum trabalho extra de limpeza ou desambiguação. O estudo trouxe um olhar diferente para os dados, agregando grafos de rede e *tree maps* aos tradicionais gráficos de barra, setorial e tabelas.

O DIA2 não é uma ferramenta específica para área da saúde, já que contempla o investimento em pesquisa em todas as áreas de conhecimento. Baseada no conceito de mineração de conhecimento e visualização interativa, a plataforma foi desenvolvida como um recurso central para pesquisadores e tomadores de decisão.

Molnar e colaboradores (2015) apostam que prover múltiplas formas de visualização dos dados permite que os usuários tenham *insights* a partir da visualização das relações, lacunas e outras conexões entre os atores que compõem o cenário do financiamento. Os autores apontam ainda que a lacuna entre a visualização de dados e os *insights* pode ser preenchida por meio de uma ferramenta computacional.

A busca pelo mapeamento e avaliação do fluxo de financiamento para a pesquisa no mundo está aquecida. Porém, no Brasil, não acontece o mesmo e, basicamente, o conhecimento gerado até hoje sobre o tema é fruto do estudo e vivência de um único pesquisador. O próximo capítulo terá como foco a pesquisa em saúde no Brasil, trazendo o histórico do modelo de financiamento para a pesquisa no país, apresentando o desafio de se reunir os dados sobre financiamento por conta da fragmentação dos mesmos e, por fim, ressalta a necessidade de se encontrar soluções que permitam a integração dos dados para que possam ser analisados de uma forma conjunta.

4. Pesquisa em Saúde no Brasil

O atual cenário do financiamento da pesquisa em saúde no Brasil começou a desenhar os seus contornos nos anos setenta do século passado após a criação das primeiras agências financiadoras – CNPq e FINEP. Este capítulo apresenta o desenvolvimento e a organização do sistema nacional de ciência e tecnologia no Brasil, tomando como base os modelos instituídos que suportaram em um primeiro momento o arcabouço institucional das referidas agências de financiamento. O texto apresenta também a diferença no nível de detalhamento dos dados apresentados pelas agências de fomento sobre o financiamento à pesquisa, o que gera dificuldade na consolidação e análise dos mesmos.

A criação do CNPq em 1951 surge como uma resposta a um modelo de inovação *science push*, ou seja, nasce da ciência o motor da inovação tecnológica. Este pensamento teve origem no período pós-guerra (1945) quando se iniciou o entendimento de ciência como “motor do progresso” (RUIVO, 1994). De forma clara, o CNPq foi criado naquele modelo de linearidade anteriormente descrito, onde:

uma sequência de estágios, em que novos conhecimentos advindos da pesquisa científica levariam a processos de invenção que seriam seguidos por atividades de pesquisa aplicada e desenvolvimento tecnológico resultando, ao final da cadeia, em introdução de produtos e processos comercializáveis. (CONDE; ARAÚJO-JORGE, 2003, p.729)

Todavia, especialistas que discutem as políticas de ciência, tecnologia e inovação em saúde, registram que a criação do CNPq serviu como um estopim para o afastamento entre “núcleo hegemônico da pesquisa em saúde e as políticas de saúde, que se traduziu em um afastamento crescente entre a temática da pesquisa e as necessidades de saúde da população.” (GUIMARÃES, 2004, p.377). Este fato se deve ao fato do CNPq adotar o modelo da lógica de financiamento individual, para o pesquisador/cientista, o qual decide o que quer pesquisar, o que nem sempre pode ser consonante alinhado com as demandas sociais, esse o coração da pesquisa em saúde coletiva.

No Brasil, esta lógica começa a mudar em 1965, com a criação do Fundo de Financiamento de Estudos de Projetos e Programas (FINEP), no governo Castelo Branco, por meio da assinatura do Decreto Nº 55.820. A ciência passa a ser vista como “solucionadora de problemas” (RUIVO, 1994). Embora o modelo de inovação tecnológica continue sendo linear, este é regido pelo mercado (demanda).

A FINEP nasce como um banco de investimento em desenvolvimento tecnológico, e financia tanto universidades que fazem P&D como empresas que procuram por estratégias para traduzir para prática suas ideias e necessidade de desenvolvimento de tecnologia.

Estes dois modelos são descritos por Tigre (2006) sob a vertente da gestão da inovação. Ele aponta que é importante entender os fatores indutores da inovação, (i) a oferta (*technology push*), (ii) a demanda (*demand pull*) e ainda (iii) os custos dos fatores de produção.

O fator *technology push* é derivado dos avanços da ciência, como atividades de pesquisa e desenvolvimento, capacitação tecnológica em empresas e universidades, difusão de conhecimento técnico-científico, gestão do conhecimento ou também para oferta de novos insumos produtivos. Ao fator *demand pull* atribuem-se as inovações desenvolvidas em resposta a demandas da sociedade (mercado) por melhor qualidade, aderência a padrões técnicos e ambientais, necessidades de segurança, customização, conveniência do usuário, eficiência econômica, ou ainda, novo design (TIGRE, 2006). O financiamento da pesquisa em saúde, ainda que se coloque confortavelmente em ambos os modelos, quanto fomentada por recursos públicos, deveria, claramente, estar melhor situada como resposta às demandas sociais.

É no cenário da ciência como “solucionadora de problema” que surge, em 1969, o Fundo Nacional de Desenvolvimento de Científico e Tecnológico (FNDCT). O FNDCT que busca fortalecer o sistema nacional de Ciência, Tecnologia e Inovação (C,T&I) foi

[...]instituído pelo Decreto-Lei n 719, de 31 de julho de 1969, e restabelecido pela Lei n 8.172, de 18 de janeiro de 1991, é de natureza contábil e tem o objetivo de financiar a inovação e o desenvolvimento científico e tecnológico com vistas em promover o desenvolvimento econômico e social do País. (BRASIL, 2007)

Desde a década de 1990, um novo modelo de financiamento em C&T, decorrente criação das agencias de regulação, e que trouxe outra fonte de financiamento, o Fundo Setorial em Saúde. Esses fundos trazem em suas origens o Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT), marco de mudanças nas políticas públicas. Porém, até 1998, os gastos em C&T eram custeados pelos recursos ordinários do Tesouro apesar da existência formal do FNDCT. No governo Fernando Henrique Cardoso, em 2001, a Lei nº 10.197 modificou o texto original acrescentando os artigos 3-A e 3-B que dispõem sobre o financiamento a projetos de implantação e recuperação de infraestrutura de pesquisa nas instituições públicas de ensino superior e de pesquisa. Já em 2007, no governo do então presidente Luis Inácio Lula da Silva, o FNDCT foi regulamentado pela Lei nº 11.540. Esta regulamentação ratifica, a partir de mudança no texto, a distribuição orçamentária para o desenvolvimento das regiões Norte, Nordeste e Centro-Oeste. As

instituições de ensino e pesquisa sediadas nestas regiões devem receber 30% (trinta por cento) dos recursos do fundo.

O FNDCT é composto por 16 Fundos Setoriais, sendo 14 referentes a setores específicos e dois transversais. Destes últimos, o Fundo Verde-Amarelo (FVA) voltado à interação universidade-empresa e o Infraestrutura com a finalidade de apoiar a melhoria da infraestrutura de Instituições de Ciência e Tecnologia (ICTs). A criação dos fundos setoriais foi importante redirecionador nas definições de prioridades para a aplicação dos seus recursos através de sua Secretaria Executiva, a Financiamento de Estudos de Projetos e Programas (FINEP).

De acordo com o sítio da FINEP (<http://www.finep.gov.br/>), “desde sua implementação nos anos recentes, os Fundos Setoriais têm se constituído no principal instrumento do Governo Federal para alavancar o sistema de C,T&I do País.” (FINEP, 2012)

A receita do FNDCT é proveniente de royalties sobre a produção do petróleo ou gás natural; percentual da receita operacional líquida de empresas de energia elétrica, percentual dos recursos decorrentes de contratos de cessão de direitos de uso da infraestrutura rodoviária para fins de exploração de sistemas de comunicação e telecomunicações, entre outros. Em outras palavras, os recursos que formam o FNDCT são provenientes de contribuições incidentes sobre o faturamento de empresas e sobre o resultado da exploração de recursos naturais pertencentes à União (BRASIL, 2007).

Os fundos são administrados por um Conselho Diretor composto pelo ministro do Ministério C,T&I, por representantes de outros ministérios e pelos presidentes de instituições de fomento (BNDES, FINEP, CNPq), por representantes das empresas, representantes da comunidade C&T e pelo presidente da EMBRAPA. Os recursos destes fundos são destinados, entre outros, para “projetos de instituições científicas e tecnológicas - ICTs e de cooperação entre ICTs e empresas” (BRASIL, 2007).

A gestão dos recursos do FNDCT está a cargo da Agência Brasileira de Inovação (FINEP), que atua como Secretaria Executiva do fundo. Porém, a FINEP acumula a função de agente executor. O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) também atua como agente executor. De forma reduzida, é possível descrever a FINEP como uma agência de fomento de projetos propostos por instituições públicas e privadas, enquanto o CNPq fomenta projetos de pesquisa individuais e concede bolsas de estudo. Verifica-se, portanto, uma complementaridade na atuação da FINEP e do CNPq.

O Fundo Setorial de Saúde (CT-SAÚDE) teve seus mecanismos de financiamento instituídos pela Lei nº 10.332, de 19 de dezembro de 2001, e regulamentados pelo Decreto Lei nº 4.143 de 25 de fevereiro de 2002.

O CT-SAÚDE tem como recursos o montante de 17,5% da Contribuição de Intervenção no Domínio Econômico (CIDE), o imposto que incide sobre a comercialização de gasolina. Os recursos passaram a ser destinados ao CT-SAÚDE, a partir de 2001 e por este motivo este fundo passou a constar nos relatórios de arrecadação, dotação orçamentária e execução financeira do Ministério da Ciência, Tecnologia e Inovação a partir do ano de 2002. Estes relatórios apresentam, dentre outros, o total de arrecadação, o total empenhado, o total executado e o total pago por fundo setorial.

Esse Fundo Setorial tem como foco: (i) a “capacitação tecnológica e inovação tecnológica nas áreas de interesse do Sistema Único de Saúde – SUS” e (ii) a “difusão e incorporação de novas tecnologias visando ampliação do acesso aos bens e serviços em saúde”. (BRASIL, 2002, p.3). Como objetivos específicos são citados: (i) contribuir para a melhoria dos serviços de saúde, (ii) ampliar o acesso aos bens e serviços de saúde, (iii) estimular o aumento do financiamento em C,T&I aplicados à saúde, (iv) estimular a equiparação tecnológica da indústria de bens e serviços aos países ditos de primeiro mundo, (v) “estimular a formação e a capacitação de recursos humanos para Pesquisa em Saúde” e (vi) “difundir o conhecimento científico e tecnológico”. (BRASIL, 2002, p.3).

A Finep não disponibiliza diretamente os dados sobre seus investimentos em pesquisa. Esta agência de fomento redireciona o acesso ao Portal da Transparência do Governo Federal. Não existe, portanto, um relatório consolidado com informações sobre o quanto ela distribui por ano, para que instituições ou para quais temáticas.

Além dos dois órgãos supracitados, CNPq e FINEP, o Brasil ainda conta no grupo das agências públicas de fomento em nível nacional, como o Banco Nacional de Desenvolvimento Econômico e Social (BNDES) e com as Fundações de Amparo à Pesquisa no âmbito estadual, que seguem a orientação do sistema nacional de ciência e tecnologia e visam estimular o desenvolvimento da ciência em caráter local.

Todas estas diretrizes citadas acima compõem uma política de fomento à pesquisa que busca catalisar o Sistema Nacional de Inovação com interseção clara no Sistema Nacional de Saúde. Sendo assim, é apropriado dizer que esta política atende às necessidades de um Sistema Nacional de Inovação em Saúde (SNIS).

O SNIS é resultado da Política Nacional de Ciência, Tecnologia e Inovação em Saúde (PNCTIS), a qual apresenta como um de seus eixos condutores a extensividade, expressa pela “capacidade de intervir nos vários pontos da cadeia do conhecimento” (BRASIL, 2006, p.19).

A extensividade

inclui toda pesquisa que visa ao avanço do conhecimento, seja aquele de aplicação imediata ou não. Inclui, portanto, além da produção de conhecimentos, as pesquisas voltadas para o desenvolvimento tecnológico e a inovação; a avaliação tecnológica, pesquisa clínica, pesquisas sobre padrões de uso e relação custo/benefício para diversos tipos de tecnologia em saúde, dentre outras (BRASIL, 2006, p.19).

Ademais, em suas estratégias, a PNCTIS aponta para a sustentação e o fortalecimento do esforço nacional em C,T&I em Saúde tendo como um norteador “a criação, ampliação, diversificação e garantia de continuidade das fontes de financiamento para ações de P&D em saúde” e como um de seus principais instrumentos o “fortalecimento dos mecanismos de fomento dos fundos setoriais à P&D” (BRASIL, 2006, p.24).

A despeito dessas políticas explícitas, muito pouco se sabe sobre a consequência direta dos investimentos em pesquisa no Brasil, de forma geral. Para o Brasil, dados provenientes do SCImago Journal and Country Rank (www.scimagojr.com) e do SCImago Institution Rank (www.scimagoir.com) apresentam um crescimento na visibilidade da produção científica do Brasil de mais de 30 vezes entre 1996 e 2015 na área da saúde, tendo sido as Ciências Biológicas com crescimento de cerca de 12 vezes seguido da Medicina com 6,5 vezes.

Esta realidade teve como um dos catalisadores o processo de ajuste da economia frente à realidade mundial a partir do início do governo do presidente Fernando Collor de Mello (1990-1992) com a abertura do mercado. O setor produtivo brasileiro sofreu – e ainda sofre – a competição acirrada das indústrias de países desenvolvidos e em desenvolvimento. Neste contexto o Brasil possui um caminho para se equiparar aos países de ponta: investir de forma sólida, consistente e coerente na educação e no desenvolvimento científico e tecnológico. (BURLE, 1993; OLIVEIRA; BIANCHETTI, 2006).

Entretanto, o Brasil carece de dados que permitam essa análise comparativa de seus investimentos e produtividade científica com outros países, desenvolvimentos e/ou em desenvolvimento. Mais, a pesquisa em saúde no Brasil é pouco discutida por autores daquele país. Poucos são os artigos publicados neste campo e a discussão por vir se baseia quase que exclusivamente em trabalhos publicados por Reinaldo Felipe Nery Guimarães que desde 1985 trabalha no campo do Planejamento, Gestão e Políticas de Ciência e Tecnologia e de Saúde. O autor afirma que no Brasil, o financiamento em pesquisa é majoritariamente

realizado pelo governo federal (GUIMARÃES, 2006). A estruturação do fomento está organizada na esfera nacional e estadual. Como já citado, relacionam-se aqui CNPq, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a FINEP e as Fundações de Amparo à Pesquisa (FAPs) de nível estadual – por exemplo, FAPESP e FAPERJ.

As iniciativas de fomento, portanto, sempre seguiram em cada agência de fomento, sua própria lógica, especialmente no que diz respeito à gestão dos processos. A realização da I Conferência Nacional de Ciência e Tecnologia em Saúde, em 1994, foi uma iniciativa que buscou criar mecanismos para coordenação da gestão da pesquisa em saúde, senão de seus resultados, pelo menos de suas prioridades. Muito embora várias boas ideias tenham saído desta reunião, pouco foi tirado do papel; muito porque o mandato presidencial desta época findou em 1995. (GUIMARÃES, 2004)

Este movimento recai sobre as ações de fomento no âmbito da pesquisa em saúde que sofre com o peso de um modelo antiquado com cerca de sessenta anos. Guimarães (2004) aponta uma série de características deste modelo de fomento. Embora o autor aponte que a “qualidade e a transparência nas ações de fomento, em particular as realizadas pelas agências do MCT⁶, pela Capes, bem como pela maioria das agências estaduais” (GUIMARÃES, 2004, p.379) se apoiam na “experiência brasileira com práticas de fomento em bases relativamente competitivas” a transparência se dá de forma fragmentada e não articulada. Nesta mesma linha, o autor cita que o modelo de fomento é descentralizado e plural; registrando que se trata de uma das qualidades do modelo uma vez que o fato de contar com diversas agências de fomento se torna “um instrumento de proteção dos executores de pesquisa contra eventuais obscurecimentos no que se refere à qualidade e à transparência nas ações de fomento.” (GUIMARÃES, 2004, p.380). Entretanto, esta característica torna-se um calcanhar de Aquiles quando da “ausência de mecanismos de coordenação adequados entre as múltiplas instâncias de fomento, em especial entre os dois atores principais, o MCT e o MS⁷.” (GUIMARÃES, 2004, p.380).

Outra característica decorrente desse modelo é a “baixa capacidade de articulação entre as ações de fomento científico-tecnológico e a política de saúde” levando a um corte de comunicação entre a geração do conhecimento e as indústrias, serviços de saúde e a sociedade em geral. Pois mesmo com “a existência de uma tradição importante em termos de institutos de pesquisa federais e estaduais”, esta falta de comunicação dificulta o processo de inovação

⁶ Ministério da Ciência e Tecnologia

⁷ Ministério da Saúde

caracterizado pela obtenção de patentes sobre um produto; ainda mais com “uma extensa e generalizada carência de atividades de pesquisa e desenvolvimento realizada nas empresas.” (GUIMARÃES, 2004, p.380)

Guimarães (2004, p.379) aponta ainda que o modelo de fomento também sofre com “a baixa capacidade de indução, especialmente nas mesmas agências do MCT, na Capes e em algumas agências estaduais” e, principalmente, com “uma quantidade de recursos para o fomento bastante aquém das necessidades” e desvinculado de um critério racional de prioridades refletido na Agenda Nacional de Prioridade de Pesquisa em Saúde, documento que deveria orientar as ações de fomento à pesquisa de acordo com as necessidades de saúde da população brasileira. Porém este documento possui, dentro das vinte e quatro subagendas, uma extensa lista – mais de 300 itens – contendo as “prioridades”. Nesta linha, a lista deixa de ser uma lista de prioridades e passa a ser uma lista de problemas de saúde. Na visão deste autor,

“o atendimento às necessidades de saúde nem sempre é uma variável dependente da pesquisa em saúde e, por outro, nem sempre há, no campo do saber e das práticas científicas e tecnológicos, conceitos, metodologia ou ferramentas adequados para que se possa produzir soluções através da pesquisa”. (GUIMARÃES, 2004, p.381)

O crescimento e envelhecimento da população, assim como o (re)surgimento de doenças, leva o setor de saúde de quase todos os países, com “a exceção dos países com pretensões hegemônicas globais ou regionais”, ao topo da lista dos setores que mais mobilizam recursos para a pesquisa, desenvolvimento e inovação. (GUIMARÃES; SERRUYA; DIAFÉRIA, 2008, p.12). Os autores acrescentam que os recursos desembolsados têm crescido em média 10% ao ano desde o ano de 1998, e destacam que o desembolso em países de renda média ou baixa tem diminuído em relação aos países desenvolvidos. Em outras palavras, países ricos têm investido proporcionalmente mais que os países com uma economia mais fraca, a cada ano que passa. Mesmo assim, os autores destacam a *performance* do Brasil, pois ele detém uma grande fatia do investimentos em saúde dentre os 5% destinados aos países de média renda, os quais desenvolveram “uma importante capacidade instalada de pesquisa, existindo para eles um papel identificável no cenário mundial de pesquisa e inovação em saúde”. (GUIMARÃES; SERRUYA; DIAFÉRIA, 2008, p.13)

Este quadro impulsionou o Brasil a alcançar a 15º posição dentre os países do mundo no que se refere a publicações científicas em periódicos indexados (BRASIL, 2008, p.9), sendo este um dado importante uma vez que este tipo de indicador se relaciona com o processo de inovação de um país. Entretanto, recupera-se aqui uma das características do

modelo de fomento apontada por Guimarães (2004) da baixa comunicação entre a geração de conhecimento e as indústrias. Neste sentido, não é possível determinar se realmente o aumento no número de publicações gera um resultado positivo dentro do processo de inovação.

Esse indicador no sistema de inovação é fruto do conceito de Sistemas Nacionais de Inovação desenvolvido na década de 80 na Europa e Estados Unidos. Nesta época os países buscavam “compreender os processos de articulação entre os inúmeros atores envolvidos no aparecimento de novos produtos e processos no mercado, em particular aqueles envolvendo conhecimento científico e tecnológico avançado”. Foram identificadas e estudadas as “relações entre atores públicos e privados com vistas ao desenvolvimento econômico”. Dentre os atores públicos destacam-se as instituições de pesquisa, as quais, a partir da geração de conhecimento e do desenvolvimento de medicamentos, vacinas e dispositivos diagnósticos, contribuem muito para o avanço do complexo industrial da saúde. Entretanto, “é necessário desobstruir ainda mais os canais de apoio a projetos de P&D diretamente” às empresas as quais são os principais locais de inovação. (GUIMARÃES; SERRUYA; DIAFÉRIA, 2008, p.18)

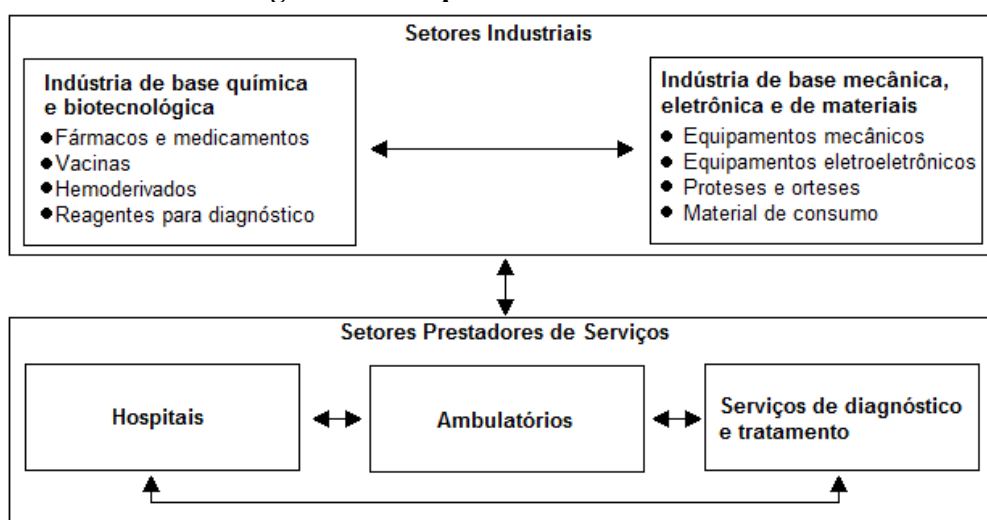
Porém, o planejamento se torna difícil diante do fato de serem “muito precárias as estimativas sobre gastos com pesquisa em saúde no Brasil” (GUIMARÃES, 2004, p.382) e da “escassez dos recursos financeiros” (GUIMARÃES; SERRUYA; DIAFÉRIA, 2008, p.16). Guimarães (2004, p.383) questiona se o recurso financeiro disponibilizado é muito ou pouco para o número de pesquisadores da área da saúde e conclui ser aquém das necessidades. Este autor acrescenta ainda que uma nova Política Nacional de Ciência, Tecnologia e Inovação em Saúde (PNCTIS) deverá ter como prioridade a busca de novas formas de obtenção de recursos para pesquisa em saúde.

Guimarães (2004, p.383) aponta uma possível solução para esta dificuldade sugerindo que se pense em “uma estrutura vinculada ao Ministério da Saúde especializada em captar, fomentar, acompanhar e avaliar a atividade de pesquisa”. Quanto à escassez de recursos financeiros, o autor sugere algumas soluções, a saber: (i) a aprovação do Fundo Verde-Amarelo o qual tem parte dos recursos alocados para saúde, (ii) o aumento do financiamento pelo Ministério da Saúde a partir da unificação das prioridades, tendo em vista também o financiamento proveniente das agências reguladoras e, de acordo como o autor a mais importante, (iii) a taxação das indústrias do álcool e do tabaco “com vistas à criação de um outro Fundo para financiar pesquisa em saúde”. (GUIMARÃES, 2004, p.384).

Ainda se tratando da escassez de recursos financeiros, o autor compara a pesquisa em saúde com a pesquisa agropecuária. O autor verificou que a pesquisa agropecuária recebeu mais recursos por pesquisador que a pesquisa em saúde. Ademais foi verificado que o Ministério da Agricultura, por meio da EMBRAPA, viabiliza quase que o dobro em termos percentuais que o Ministério da Saúde – 20% para o MS e 38,6% para o MAPA. (GUIMARÃES, 2004)

Guimarães (2004) passeia pelo complexo industrial da saúde, Figura 3, descrevendo o papel dos setores industriais e seus desafios em relação ao financiamento para saúde. O autor aponta que o maior desafio está para o setor da indústria de base química e biotecnológica, mais especificamente para a indústria de fármacos e medicamentos.

Figura 3 – Complexo Industrial da Saúde



Fonte: Modificado de Gadelha (2003)

O maior problema apontado pelo autor é a concentração do domínio patentário deste setor pelos países mais desenvolvidos. Este problema teria sido potencializado com a abertura comercial indiscriminada observada no Brasil durante a década de 1990, sem que fossem tomadas as devidas precauções frente à capacidade produtiva nacional. O autor aponta que tal medida “nos fez recuar em relação ao que já havíamos conquistado” e que aumentou a “fragilidade da estrutura industrial brasileira em saúde”, fazendo com que os setores industriais perdessem “competitividade no mercado brasileiro e no exterior”. (GUIMARÃES; SERRUYA; DIAFÉRIA, 2008, p.19) e (GUIMARÃES, 2004, p.386)

Guimarães, Serruya e Diaféria (2008) ressaltam a dificuldade de se reverter o quadro. Entretanto os autores apontam seis ações que ajudariam a trilhar este árduo caminho. Destas, (i) três relacionadas diretamente com pesquisa e desenvolvimento, (ii) duas referentes à

definição de uma política, seja na construção de bases sólidas entre as empresas, instituições de pesquisa e sistema de saúde; seja na regulação do mercado para tornar os produtores internos mais competitivos e (iii) uma diretamente associada ao financiamento à saúde com a finalidade de fortalecer a capacidade produtiva de insumos à saúde.

Porém, mesmo algumas pesquisas sejam financiadas, o que corroboraria com o item (i) acima, com certeza não é suficiente para fortalecer o Sistema Nacional de Inovação (SNI). Guimarães (2006) aponta que, no Brasil, o SNI é imaturo, pois dentre outros, o “*volume expressivo de recursos financeiros destinados à P&D em saúde, correspondente a 1,5% dos gastos nacionais com saúde e a 3,3% dos gastos nacionais públicos com saúde.*”(GUIMARÃES, 2006, p.6). Este investimento é insuficiente. O próprio autor articula neste sentido e alerta para os limites deste pequeno volume de investimento com base no Global Forum for Health Research (1998). Os dados levantados por ele informam que foram utilizados recursos no montante de “US\$ 73,5 bilhões, [sendo] mais de 90% nos países ricos e visando resolver os problemas dos países ricos.” (GUIMARÃES, 2004, p.376).

Guimarães (2010, p.1913) apresenta um novo rumo para este cenário. O autor aponta que o Brasil começou a viver “um novo longo ciclo, iniciado no último ano do século passado, com a criação dos fundos setoriais.”. Ele ilustra esta afirmação com o orçamento do Fundo Nacional de Desenvolvimento Científico e Tecnológico que fora elevado em 2010 para um patamar de três bilhões de reais. Guimarães (2010) ressalta também que a pesquisa no Brasil tem um forte vetor nos programas de pós-graduação. O autor aponta para um futuro onde haverá um redirecionamento dos recursos financeiros para a pesquisa realizada por pesquisadores doutores, e não por aqueles que ainda estão neste percurso. Em outras palavras, o autor sugere que a verba é direcionada para as pesquisas realizadas por aqueles que ainda estão no processo de formação, ou seja, cursando o doutorado. Porém que há uma tendência futura para o direcionamento da verba para as pesquisas realizadas por aqueles já com o título de doutor.

Entretanto, Guimarães (2010, p.1914) aponta que ainda é necessário um forte investimento nos programas de pós-graduação profissionais em detrimento ao acadêmico. Segundo o autor os primeiros deverão formar bons profissionais para o “mercado de trabalho extra-acadêmico” e que isto não se restringe somente aos os campos da saúde e das ciências da vida em geral, mas também nas humanidades e ciências da natureza: “haverá situações nas quais um olhar extra-acadêmico (mas não anti-acadêmico) será importante na definição de prioridades, na seleção de projetos e em sua avaliação.”. (GUIMARÃES, 2010, p.1914)

Conforme descrito, o Brasil ainda está no caminho de uma PNCT&I/S consolidada. Neste caminho ainda há barreiras que envolvem um precipício entre a pesquisa e a indústria, assim como a falta de recursos financeiros para o desenvolvimento das pesquisas. Guimarães (2010) concorda com Pang e colaboradores (2003) quando se discute os pontos a serem observados na elaboração da política e na prática da pesquisa. Estas funções são: (i) o gerenciamento, (ii) o financiamento, (iii) criação e manutenção de recursos e (iv) produção e utilização da pesquisa. Para o gerenciamento se destaca a identificação das prioridades de pesquisa em saúde e, por conseguinte, a coordenação da adesão às mesmas. Já para o financiamento, deve-se objetivar a criação de fundos de pesquisa seguros e alocá-los de forma responsável. Na criação e manutenção de recursos se espera construir, fortalecer e sustentar a capacidade humana e física para conduzir, absorver, e utilizar a pesquisa em saúde. A última função do sistema de pesquisa em saúde, produção e utilização da pesquisa, almeja produzir resultados de pesquisa cientificamente válidos, traduzir e comunicar pesquisa para informar a política de saúde, estratégias, práticas e opinião pública; e, por fim, promover o uso de pesquisas para desenvolver novos medicamentos, vacinas, material de insumo para uma melhor a saúde.

Das quatro funções apontadas por Guimarães (2010), resgata-se aqui que o Currículo Lattes poderia dar conta de pelo menos parte da quarta função referente à produção, porém registre-se, novamente, que o mesmo é autodeclaratório, acarretando problemas na análise destes dados como inconsistência.

Referente à primeira função do financiamento ressalta-se que o Ministério de Ciência e Tecnologia, assim como todas as agências no âmbito federal, possui uma página na internet com dados a respeito da execução financeira. As agências de fomento, tanto em nível federal quanto em nível estadual, tornam público o financiamento direcionado às pesquisas, entretanto não com o detalhamento necessário para que seja realizada uma unificação destes dados a ponto de se ter uma noção melhor sobre o financiamento à pesquisa no Brasil.

O Ministério de Ciência, Tecnologia e Inovação disponibiliza em seu sítio os relatórios de Arrecadação, Dotação Orçamentária e Execução Financeira em duas versões anuais: (i) aberto por ações e (ii) consolidado. O primeiro apresenta os valores discriminados para cada ação dos referidos fundos, enquanto o segundo, autodescrito pelo próprio nome, consolida os valores das ações pelos fundos. Existem relatórios disponíveis a partir do ano de 1999. O fundo setorial CT-Saúde começa a aparecer nos relatórios de 2002 e teve a distribuição de verba de acordo com o representado a seguir, Quadro 1:

Quadro 1 – Verba arrecadada e empenhada no CT-SAÚDE de 2002 a 2013

ANO	Arrecadado (R\$)	LOA (R\$)	Empenhado (R\$)
2002	41.334.76	50.540.000	421.100
2003	67.060.421	82.137.599	24.174.945
2004	61.030.204	53.241.735	26.912.745
2005	70.917.517	70.960.089	33.918.793
2006	74.223.077	62.362.849	54.508.945
2007	90.144.755	77.160.229	67.346.604
2008	102.776.614	90.555.909	80.855.263
2009	128.517.324	88.152.623	81.341.835
2010	135.705.627	88.000.000	87.087.466
2011	168.828.100	69.029.703	47.369.939
2012	221.654.550	79.455.123	46.270.283
2013	254.316.406	89.455.124	59.358.777

Fonte: BRASIL, ([s.d.])

A Lei Orçamentária Anual (LOA) é uma previsão de quanto será arrecado e estipula um teto que pode ser utilizado. Verifica-se aqui que em todos os anos o valor empenhado foi aquém do estipulado pela LOA.

O CNPq por sua vez, disponibiliza em seu portal (<http://www.cnpq.br>), dados das pesquisas por ele financiadas, a saber: título do projeto de pesquisa, tempo de vigência da pesquisa, montante financeiro recebido, pesquisador contemplado com a verba e sua instituição estado e município da instituição de origem do pesquisador principal. Os dados presentes no site desta agência de fomento não podem ser extraídos por relatório, a coleta de dados tem que ser executada “um a um” para o conjunto de variáveis de cada projeto, não permitindo facilmente qualquer tipo de análise.

A CAPES, assim como a FINEP, redireciona o acesso a seu portal sobre suas respectivas despesas com bolsas ou projetos financiados para o Portal da Transparência (<http://www.portaldatransparencia.gov.br>). Este portal apresenta quem são os beneficiados e os quais valores recebidos por eles. A CAPES possui também uma ferramenta chamada Geo CAPES (<http://geocapes.capes.gov.br>) que apresenta o número total de bolsas concedidas por programas, porém, sem os valores. Estes dados são úteis, entretanto, para calcular o montante total investido na capacitação, utilizando para cálculo do valor vigente da bolsa no Brasil.

No que diz respeito ao âmbito estadual, os dados sobre financiamento de projetos que as FAPs disponibilizam – quando o fazem – também se diferencia entre elas. A FAPERJ possui uma ferramenta de busca chamada “Resumo de Projetos Contemplados” que permite filtrar, por exemplo, por programa, nome do pesquisador e palavra chave. Ressalta-se que o resultado não recupera o montante financiado.

Já a FAPESP apresenta seus dados no que ela chama de “Biblioteca Virtual” e possui alguns tipos de filtros como área do conhecimento e ano de início do projeto, porém também não traz os valores. A Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) redireciona para o portal de transparência daquele estado que, embora somente possua filtros por ano/mês, contém dados como nome do pesquisador e montante recebido.

Também foi realizada uma rápida busca no site da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) e não foi encontrada nenhuma forma de verificar dados sobre o apoio à pesquisa. Para compor pelo menos um representante de cada região do país, também foram verificados os sites da Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) e Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB). A FAPEG também redireciona o acesso aos dados financeiros para o portal de transparência de seu estado, porém se diferencia dos dados da FAPEAM, pois estão consolidados no montante anual. Na FAPESB, assim como na FAPERGS, não foram encontrados dados financeiros sobre fomento à pesquisa.

A dispersão dos dados sobre financiamento à pesquisa dificulta a utilização do conjunto produzido, especialmente quando há necessidade de uma medida de desfecho. As decisões sempre envolvem grau de riscos e de incertezas, porém quando embasadas em um conjunto de dados sistematizados que apresentem padrões e apontem tendências é possível mitigá-lo.

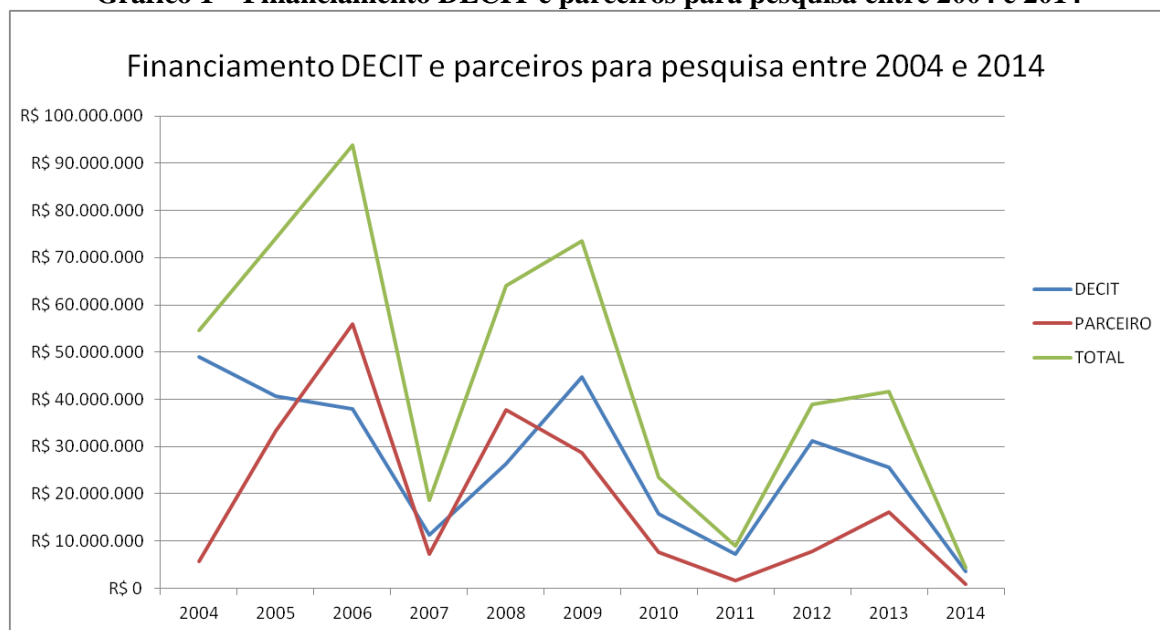
Cabe destacar que não são todas as bases citadas acima que possibilitam filtrar por temática muito embora seja desta forma que se organiza grande parte dos editais de financiamento lançados no Brasil. O filtro pela temática traria uma visão de quanto se tem investido em pesquisas naquele tema. No caso da dengue, por exemplo, embora de caráter estratégico no cenário da pesquisa em saúde e objeto de vários editais de financiamento, inexistente informação sistematizada a respeito do montante financeiro executado com vistas a subsidiar políticas e a alocação equitativa de recursos no país.

A seção “despesas” do site da FINEP divulga dados sobre a execução orçamentária e financeira. A descrição desta seção sugere que seja consultada a seção “acesso público” do Sistema Integrado de Planejamento e Orçamento (SIOP). O SIOP (<https://www.siop.planejamento.gov.br/siop/>), por sua vez, permite filtros pelas variáveis: ano, órgão e unidade. Porém, só apresenta o volume monetário total empenhado e liquidado por ano para o total de projetos financiados. Ademais, está apenas disponível filtro pelos últimos cinco anos (2010 a 2014).

Outro sítio eletrônico que disponibiliza dados sobre o financiamento em pesquisa no Brasil é o do Departamento de Ciência e Tecnologia (Decit), vinculado ao Ministério da Saúde, muito embora este não seja exatamente uma agência de fomento.

O Decit disponibiliza uma ferramenta de consulta chamada Pesquisa Saúde (<http://pesquisasaude.saude.gov.br/pesquisas.xhtml>) onde podem ser encontradas dados sobre projetos de pesquisas apoiados por ele desde 2002, com a colaboração do CNPq, Finep, Organização das Nações Unidas para a Educação, a Ciência e a Cultura (Unesco), Organização Pan-Americana da Saúde (OPAS), Fundações de Amparo a Pesquisa, Secretarias Estaduais de Saúde e de Ciência e Tecnologia. Esta ferramenta é um pouco mais completa que as demais e permite filtro pelas subáreas e busca por palavra chave. O Gráfico 1 a seguir apresenta o volume financiado pelo Decit e parceiro nos anos de 2004 a 2014.

Gráfico 1 – Financiamento DECIT e parceiros para pesquisa entre 2004 e 2014

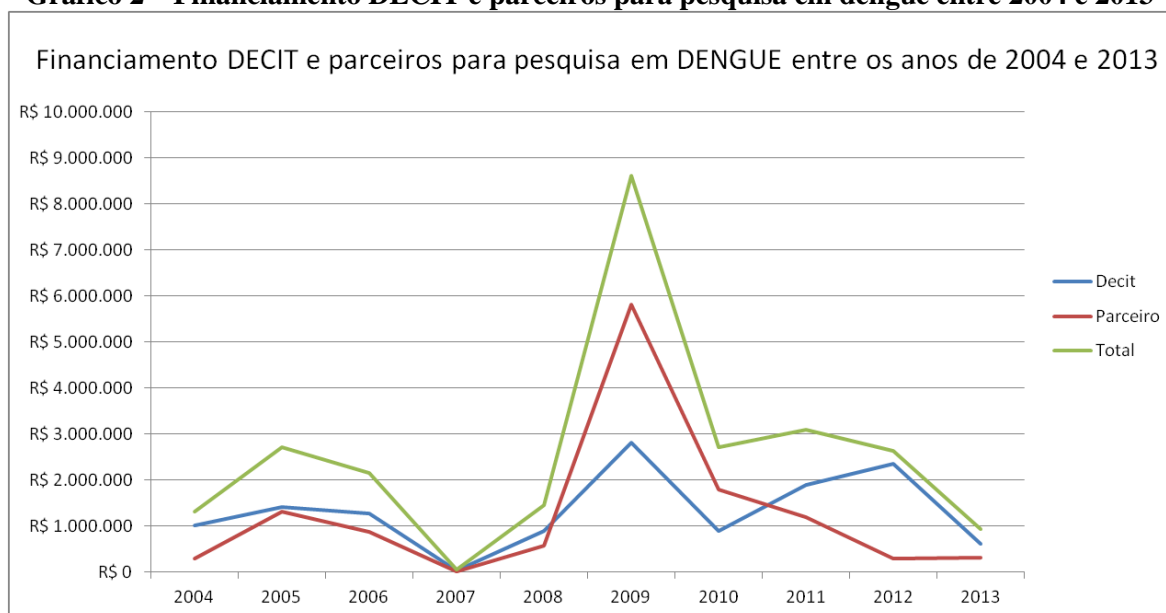


Fonte: BRASIL. MINISTÉRIO DA SAÚDE. SECRETARIA DE CIÊNCIA, (2007)

Dentre os campos apresentados na ferramenta do DECIT encontram-se o título do projeto, palavras-chaves e resumo. O termo “dengue” foi utilizado para filtrar o resultado de financiamento para apresentar a distribuição dos valores destinados para pesquisa em dengue ao longo dos anos.

O Gráfico 2 a seguir, apresenta o financiamento realizado pelo Decit e parceiros entre os anos de 2004 a 2013 para projetos com o termo dengue conforme descrito. Ressalta-se aqui que o ano de 2014 não apresentou nenhum projeto com aquele termo.

Gráfico 2 – Financiamento DECIT e parceiros para pesquisa em dengue entre 2004 e 2013



Fonte: BRASIL. MINISTÉRIO DA SAÚDE. SECRETARIA DE CIÊNCIA, (2007)

A fragmentação dos dados do financiamento em pesquisa no Brasil gera desafios para elaborar políticas públicas fundamentadas, isso para não falar sobre a falta de transparência com o uso de recursos públicos. Os dados de financiamento em pesquisa em saúde no Brasil estão dispersos em diferentes sites que não são sistemas interoperáveis, impossibilitando o cruzamento entre eles. A precariedade na interoperabilidade é caracterizada por diversas razões tecnológicas, entre elas a disponibilização de dados em formatos distintos; além do desalinhamento dos dados apresentados, ou seja, enquanto uma agência de fomento apresenta os dados pelo projeto financiado, outro apresenta por unidade federativa e uma terceira por instituição, impossibilitando assim o relacionamento entre os financiamentos e, por conseguinte, sua análise.

A dificuldade em se conhecer o cenário atual do fomento à pesquisa levou à conclusão, dentre outras, de que é difícil “definir as prioridades de pesquisa, em especial nas agências do MCT, na Capes [...]” (BRASIL, 2006, p.14). Esta dificuldade aponta para o que parece ser um dos principais empecilhos para a PNCT&I/S: “a dificuldade de coordenação, a pouca articulação e regulação governamental” (BRASIL, 2006, p.15). A Política Nacional de Ciência, Tecnologia e Inovação em Saúde (PNCTIS) aponta ainda que “não é fácil quantificar os esforços de CTI/S no País”, pois há dados “bastante precários, havendo pouca informação sobre o setor Saúde”. Acrescenta ainda que “no que se refere ao gasto em ações de CTI/S, não há informações consolidadas.” (BRASIL, 2006, p.11).

A descrição apresentada acima foi necessária para entender como se deu e vem se dando a lógica do financiamento de pesquisa em saúde no Brasil. O que se sabe sobre esse

tema se deve, particularmente, à produção do já citado Reinaldo Guimarães, que foi um ator central na estruturação do DECIT, o primeiro e mais importante movimento do Ministério da Saúde para incorporar a lógica da pesquisa no Sistema Único de Saúde. A discussão traz à tona o problema de se encontrar bases sólidas de financiamento à pesquisa, e aponta para a dispersão dos dados pelas diversas agências de fomento; dificultando a geração de conhecimento para suportar os formuladores de política a manter as políticas de ciência e tecnologia no Brasil.

Assim, no contexto desta pesquisa buscou-se no Diário Oficial da União (DOU) uma base sólida e inequívoca fonte de dados federais de financiamento à pesquisa que ajudasse a (re)pensar as políticas públicas no tema; e no Currículo Lattes uma base que, embora auto declaratória, pudesse ser integrada ao DOU para com isso contar a história do financiamento para pesquisa em dengue no Brasil.

Tal desafio é particularmente grandioso quando relacionado com as doenças negligenciadas, as quais possuem como algumas de suas características o baixo financiamento para pesquisa e a pouca atenção do Estado. Neste conjunto de doenças, encontra-se a dengue, transmitida pelo mosquito *Aedes Aegypt* o qual também é o vetor transmissor da febre chikungunya, febre amarela e zika vírus. O próximo item abordará o caráter estratégico de se estudar a dengue, apresentando rapidamente a história natural da doença, apontar para os trabalhos que se dedicam a discutir os custos da doença e ressaltará que embora alguns artigos apontem que essa temática tem sido agraciada com financiamento, não existem estudos que consigam traçar seu perfil de fomento.

4.1. O caráter estratégico da pesquisa em dengue

A dengue é uma doença viral que atinge humanos e preocupa pelas consequências clínicas dos indivíduos afetados e pelo desdobramento socioeconômico para o país (WHO, 2014). Ademais, temos observado um quadro de contínua elevação de sua incidência no Brasil e nos demais países de clima tropical e subtropical. Soma-se a isto, o surgimento de outras doenças vinculadas ao mesmo vetor da dengue (FIOCRUZ, [s.d.]). Neste sentido, avançar no estudo e no manejo da dengue contribuiu com avanços no manejo de outras doenças como a Zika e o Chikungunya.

Antes de tratar do cenário de pesquisa em dengue e das expectativas enquanto agenda de pesquisa nesta área, cabe explanar sobre a história natural da doença e sobre os seus desdobramentos clínicos para o paciente e socioeconômicos para o país.

O vírus da dengue tem no mosquito *Aedes Aegypti* o seu vetor de transmissão. O ciclo tem início quando a fêmea do mosquito que se alimenta de sangue humano tem contato com sangue de um humano contaminado. Uma vez portador do vírus e passado o tempo de incubação que pode variar de 4 a 10 dias o mosquito fica hábil para transmitir o vírus para outras pessoas ao longo de toda a sua vida. A alimentação do *Aedes Aegypti* ocorre mais intensamente pela manhã e à tarde, antes do escurecer, horário que a população ativa está a caminho do trabalho e de volta para casa respectivamente e, de certa forma, mais exposta aos mosquitos. A fêmea do *Aedes Aegypti* pode carrear diferentes cepas do vírus da dengue (WHO, 2014). Até o momento foram identificadas quatro diferentes cepas, a saber: DEN-1, DEN-2, DEN-3 e DEN-4. O contato do humano com uma destas cepas não o imuniza para os outros três tipos, ou seja, a infecção por uma das cepas só produz imunidade para ela mesma. O contato do humano com uma destas cepas do vírus, por outro lado, aumenta a sua chance de ter a forma mais grave da doença, chamada febre hemorrágica da dengue (FHD). A gravidade da febre hemorrágica da dengue está relacionada com a diminuição acentuada de plaquetas, o que pode levar ao óbito. (RAMIREZ-JIMENEZ et al., 2013; SELIGMAN, 2008; WHO, 2014).

A infecção por dengue provoca, em algumas pessoas que são infectadas, uma síndrome gripal forte que as afasta das suas atividades cotidianas por cerca de 15 dias. O quadro clínico é de febre alta com fortes dores musculares. Os sintomas podem incluir dor de cabeça, prostração, falta de apetite, fraqueza, dor retroorbital, náuseas, vômitos, erupções no corpo e coceira. A assistência médica se ocupa de minimizar estes sintomas e de acompanhar e hidratar o paciente para evitar o comprometimento das suas funções vitais. O protocolo de tratamento da dengue preconiza o uso de paracetamol enquanto analgésico e antitérmico. O acompanhamento do paciente pela equipe de saúde tem atenção especial na identificação dos casos de febre hemorrágica. (WHO, 2014; Ministério da Saúde, 2002; Ministério da Saúde 2013)

A atenção à saúde na dengue tem altos custos para um país. Os custos diretos com a dengue envolvem segundo Halasa, Shepard e Zeng (2012) o gasto com os medicamentos, com os funcionários de saúde e com o material hospitalar. Os custos indiretos, por sua vez, incluem os dias não trabalhados de pessoas infectadas e dos acompanhantes destes, bem como, a ausência da escola para pacientes em idade escolar. Segundo os autores os custos indiretos superam os custos diretos. Suaya e colaboradores (2009) avançam neste empreitada do cálculo dos custos da dengue e chegaram à proposta de que o custo da assistência médica

deve incluir os produtos utilizados na assistência ambulatorial e hospitalar e ser observado em separado por segmento da assistência – público e privado. No cálculo dos custos indiretos Suaya e colaboradores (2009) utilizaram um valor definido como salário mínimo diário, ou seja, o valor do salário mínimo dividido pelo número de dias do mês, e multiplicaram pelo número de dias médio de duração da doença. O acesso aos registros das empresas para identificação de faltas em decorrência da dengue, assim como, os consulta aos registros escolares para verificar ausências é deveras difícil. Halasa, Shepard e Zeng (2012) discutem este limite. Estudos que conseguiram identificar valores para custos da dengue apontam serem estes superiores a US\$ 40 por dia de doença. (AÑEZ et al., 2006; HALASA; SHEPARD; ZENG, 2012). Cabe destacar os achados de Añez e colaboradores (2006) de gastos superiores a US\$235.000 com a dengue no ano de 2001 pela Venezuela. Taliberti e Zucchi, (2010) observaram que o município brasileiro de São Paulo gastou US\$ 12.486.941,34 com o Programa Municipal de Controle da Dengue (PMCD) em 2005, o que inclui visitas domiciliares informativas e de inspeção, bloqueios de mosquitos com a utilização de larvicidas e inseticidas, busca ativa de casos, inspeção de pontos considerados estratégicos e levantamento de densidade larvária e pesquisa entomológica.

O resultado destas pesquisas é relevante para balizar a agenda de pesquisa em dengue e os valores de financiamento nesta área. Espera-se que as pesquisas em dengue se concentrem em (a) fortalecer o conhecimento sobre a dinâmica da infecção e o aprimoramento das ações antivetoriais; (b) estudos dos fatores de risco relacionados à ocorrência das formas graves e suas relações com a fisiopatogenia da doença; (c) aprimoramento de diagnóstico da dengue, manejo e tratamento das formas severas; (d) desenvolvimento de vacinas. Estas categorias foram propostas por Barreto e Teixeira em 2008 enquanto elementos para uma agenda nacional de pesquisa em dengue como produto do estudo da situação epidemiológica e das ações de controles executadas no Brasil naquele momento. Investimentos em pesquisas inseridas nestas categorias podem representar economia monetária para o país em decorrência de redução dos gastos diretos e indiretos relacionados com a dengue, assim como, colocarem em situação de destaque no mercado internacional de correlatos para diagnóstico e produção de vacinas.

Ações de enfrentamento à dengue no Brasil são de tamanha importância para o país que em julho de 2002 o Ministério da Saúde do Brasil publicou o Programa Nacional de Controle da Dengue em observância ao:

I - desenvolvimento de campanhas de informação e de mobilização das pessoas, de maneira a criar-se uma maior responsabilização de cada família na manutenção de seu ambiente doméstico livre de potenciais criadouros do vetor; II - fortalecimento da vigilância epidemiológica e entomológica para ampliar a capacidade de predição e de detecção precoce de surtos da doença; III - melhoria da qualidade do trabalho de campo de combate ao vetor; IV - integração das ações de controle da dengue na atenção básica, com a mobilização do Programa de Agentes Comunitários de Saúde e Programa de Saúde da Família; V - utilização de instrumentos legais que facilitem o trabalho do poder público na eliminação de criadouros em imóveis comerciais, casas abandonadas, dentre outras; VI - atuação multissetorial por meio do fomento à destinação adequada de resíduos sólidos e a utilização de recipientes seguros para armazenagem de água; VII - desenvolvimento de instrumentos mais eficazes de acompanhamento e supervisão das ações desenvolvidas pelo Ministério da Saúde, Estados e Municípios. (BRASIL, 2002a)

Muito embora este documento não explicita a agenda de saúde em dengue, ele define ações prioritárias e cria um grupo executivo ligado a Fundação Nacional de Saúde (FUNASA) para coordenar e implantar estas. Uma leitura atenta destas ações permite ver a agenda de pesquisa que está por detrás destas e que é necessária de ser construída em suporte a sua execução. Esta agenda inclui, dentre outros, a necessidade de avanço no debate ético do direito público versus o direito individual explícito aqui no conflito entre a liberdade individual e o controle sanitário (FUNASA, 2006).

Como exemplo de desdobramento deste texto legal e na direção de mobilização para a implantação das ações definidas por este, vemos a instituição da Rede Dengue Fiocruz em 2003 (FIOCRUZ, [s.d.]). A Rede Dengue Fiocruz foi instituída com recursos do Programa de Desenvolvimento e Inovação Tecnológica em Saúde Pública. Seus objetivos estão categorizados em (i) Educação e Promoção da Saúde, (ii) Gestão do Cuidado e Atenção às Urgências, (iii) Mapeamento de Vulnerabilidade Territorial, (iv) Avaliação e Monitoramento das Ações, (v) Pesquisa e Desenvolvimento Tecnológico, (vi) Comunicação e Informação e (vii) Articulação Intersetorial. A descoberta de duas outras doenças veiculadas pelo *Aedes Aegypti* ampliou, em 2015, o escopo da Rede Dengue Fiocruz para abranger também questões relativas ao chikungunya e a zika. Cabe colocar que a proposta reconhece dengue, zika e chikungunya como uma problema de saúde pública e se articula para atacá-lo com ações integradas e intersetoriais organizadas por projetos em atenção aos eixos preconizados pela diretriz do Programa Nacional de Combate à Dengue. Um dos produtos da Rede Dengue, Zika e Chikungunya, por exemplo, é o seu website rededengue.fiocruz.br. Nele é possível encontrar informação sistematizada sobre cada uma destas doenças, sua evolução e os seus desdobramentos tanto em linguagem popular quanto científica.

Segundo as últimas notícias publicadas no portal da Rede Dengue, Zika e Chikungunya até agosto de 2016 foram registrados 1.426.005 casos de dengue no Brasil no ano de 2016. A zika trouxe complicações em especial para gestantes e seus bebês – nove mulheres que adquiriram zika durante a gravidez foram a óbito; 42% das crianças que nasceram de mães infectadas por zika apresentaram pelo menos um destes quadros: microcefalia, lesões cerebrais, calcificações cerebrais, lesões na retina, surdez, dificuldades para se alimentar. Até outubro de 2016 foram notificados 9.862 casos de microcefalia no Brasil, incluídos neste número recém-nascido, natimorto e abortamento. Até setembro de 2016 foram notificados 236.287 casos de chikungunha no Brasil (FIOCRUZ, [s.d.]).

O caráter estratégico da pesquisa em dengue se dá não somente pelo necessário avanço do conhecimento na temática, mas também pelo fato da doença ocupar as manchetes de jornais brasileiros em períodos bem marcados do verão, em geral testemunhando o flagelo da convivência com um vetor que, a princípio, parece tão inofensivo e passa despercebido por muitos, mas que faz estragos enormes no cotidiano do país. Os investimentos e a produção científica brasileira nesta temática, por sua vez, merecem ser observadas reconhecendo a sua grandeza em termos de volume. Para tanto será resgatado o conceito e os recursos técnicos descritos para o manejo de *big data*. Sendo assim, o próximo capítulo trará uma abordagem do conceito de *big data* assim como de sua utilização e concluirá com o conceito de análise visual, do inglês *visual analytics*, um novo conceito que vem ganhando vulto por conta do primeiro.

5. Big data: para além de um grande conjunto de dados.

Big data enquanto conceito surge na década de 90 em designação ao manejo de um conjunto grande de dados com vistas à extração de informação para fundamentar atitudes. Este conceito em pareamento ao avanço tecnológico ocorrido nas últimas duas décadas sofre considerações tanto em relação ao tipo de dado que pode ser considerado como *big data* como quanto ao processo de extração e de apresentação das informações extraídas destes dados.

Com relação ao tipo de dado que cabe no conceito de *big data*, autores como Laney (2001) e Diebold (2012) colocam que importa não apenas o volume destes dados, mas a sua variedade e velocidade de surgimento. Laney (2001) iniciou esta discussão conceitual apontando como sendo três os desafios da tecnologia da informação para o manejo de *big data*, a saber: (i) o volume de crescimento dos dados, (ii) a velocidade em que os dados são criados e (iii) a extensão na variedade dos tipos de dados disponíveis. Volume, velocidade e variedade ficaram assim conhecidos como os 3Vs do *big data*. Fan e Bifet (2013) acrescentaram a esta discussão mais dois Vs: *variability* (variabilidade) que caracteriza mudanças da estrutura dos dados e como os mesmos são interpretados, e; valor dos dados, que agrega vantagem competitiva a partir das tomadas de decisão baseadas nos dados.

O processo de extração e de apresentação das informações em *big data*, por sua vez, demanda inovar para que seja possível interação com estes dados. Para o grupo Gartner INC. ([s.d.]) a inovação no processo é inerente a definição:

...grande volume de dados, uma alta velocidade e/ou uma alta variedade nos ativos de informação que demandam formas inovadoras de processamento de informações que permitem a visão melhorada, tomada de decisão, e automação de processos de baixo custo. (GARTNER, INC., [s.d.])

Filho e colaboradores (2015) também colocam da importância da estruturação de novos processos de análise dos dados estar presente na conceituação de *big data* ao definirem este como sendo uma "quantidade de dados suficientemente grande que leve a uma mudança nas formas tradicionais de análise de dados" (FILHO et al., 2015, p. 326). Estes autores tomam a posição de necessidade obrigatória da presença dos 3Vs em associações à presença do desafio metodológico no processamento e na análise dos dados para que possa ser atribuído a este a denominação de *big data*. Popowich, por sua vez, em sua fala no evento EDCON: Big Data & Analytics de 2014, apresentou o *visual analytics* como uma alternativa metodológica para processamento e análise de dados em *big data*. Para este pesquisador, é importante elevar a capacidade de interação com dados decorrentes de big data para qualificar

a produção decorrente desta interação. Neste sentido, uma evolução no campo do *visual analytics* significa avanço direto na utilização de *big data*.

Visual analytics, you know everyone here is talking about big data, but it only actually make sense of big data and do interact with your big data if you can visualize it, if you find interesting ways to interact with your data. You can get much more information from the other than traditional tables and bar charts. (EDCON, 2014)

A demanda de manejo de *big data* cresce entre as diferentes áreas de conhecimento. O uso de *big data* para *policy-makers* é de grande valia. Hay e colaboradores (2013) apontam que a adoção de *big data* catalisa a atenção dispensada pelos *policy-makers*. Esta utilização impulsiona um novo pensar na pesquisa a partir de um conjunto de dados que se afina com o conceito de *big data*.

Nos esportes, por exemplo, a possibilidade do uso de vídeos e telemetria para captura de dados proporciona um número sem precedentes de dados que carecem de análise para fundamentar qualificação. Esta afirmativa se aplica tanto para esportes coletivos como para os individuais (DING; ZHENG; RONG, 2014; SACRIPANTI, 2013; WANG; WENBO; SHEN, 2014). Na economia, algumas empresas como a AMAZON[©] processam um grande volume de dados gerados internamente para guiar suas estratégias de venda. Ademais, algumas destas empresas como a própria AMAZON[©] vendem infraestrutura em forma de serviço para quem possui um grande volume de dados e deseja analisá-los e/ou armazená-los. O Amazon Web Services oferece serviços de hospedagem de sites, backup e recuperação de dados, manejo de *big data*. Ainda a partir deste exemplo, cabe colocar que em 2012, a AMAZON[©] faturou uma quantia superior a 70 bilhões de dólares americanos, cerca de 15,5% do montante em vendas de *e-commerce* nos Estados Unidos. Este volume é refletido no volume de dados sobre as vendas, incluindo dados sobre os consumidores e suas compras (BOKHARI, 2013).

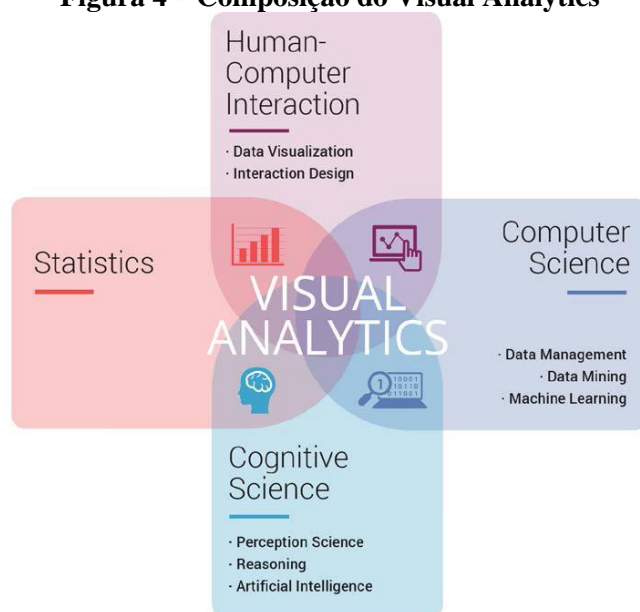
Na saúde, um exemplo de necessidade de manejo de *big data* vem do monitoramento de eventos adversos para com medicamentos, denominado de farmacovigilância. Neste caso, o volume de dados analisados é que permite predizer a segurança de um medicamento. Quando maior o volume de dados melhor o potencial de predição de eventos adversos. Abbott (2013) discute os limites no intercâmbio de dados de diferentes esferas governamentais na estruturação de um banco de dados com os contornos de *big data* para suportar a atividade reguladora do comércio farmacêutico. O autor coloca que a existência dos dados individuais dos pacientes não tem sido suficiente para viabilizar a análise e a ação sanitária. O texto usa do caso do medicamento Vioxx[©] para apresentar possibilidades de incentivo à participação dos entes governamentais e discutir os limites de articulação destes com o referencial da

tecnologia da informação em saúde. Para além deste exemplo, Filho e colaboradores (2015, p. 326) colocam que “apesar de a revolução do big data na saúde estar apenas começando, já é possível identificar três áreas auspiciosas para os próximos anos: a medicina de precisão, os prontuários eletrônicos do paciente e a internet das coisas.”

Para tanto, os dados devem ser manejados e apresentados de modo que seja possível a interação com eles. Esta é a proposta do *visual analytics*. O potencial do *visual analytics* é bem aplicado no enfrentamento do desafio de manejo de *big data* e aparece bem descrito no texto **Illuminating the path: the research and development agenda for visual analytics** publicado em 2005. Este texto, produzido pela National Visualization and Analytics Center e pelo United State Department of Homeland Security, surge em seguida ao atentado sofrido pelos Estados Unidos em setembro de 2001 e apresenta o *visual analytics* como uma forma de responder ao desafio de conseguir respostas rápidas para guiar uma ação efetiva em situações de emergência e/ou no enfrentamento de situações com múltiplos fatores a serem analisados e onde este se alteram com uma certa velocidade e sem regularidade. O texto toma o terrorismo nos Estados Unidos como caso de análise para explicar e posicionar as contribuições do *visual analytics* como ferramenta interativa de visualização e análise de *big data*.

O conceito de *visual analytics* é apresentado pelo texto como a ciência do raciocínio analítico facilitada por interfaces visuais interativas. A utilização deste conjunto de técnicas permite que dados com característica de *big data* fiquem visualmente explicativos e avança rompendo limites tecnológicos para oferecer análises oportunas e compreensíveis, comunicando de forma efetiva para sustentar a ação. O *visual analytics* é um campo multidisciplinar, Figura 4, que inclui: (i) técnicas de raciocínio analítico voltadas a estimular o surgimento de ideias capazes de apoiar a avaliação, o planejamento e a tomada de decisões, (ii) representações visuais e técnicas de interação voltadas a suportar a exploração e a compreensão de uma intensa quantidade de informação de uma só vez, (iii) manejo e representações de dados para conversão de dados conflitantes de modo dinâmico e capaz de suportar a visualização e a análise, (iv) técnicas de produção, apresentação e divulgação de uma informação em um determinado contexto e para uma variedade de públicos. Estas quatro categorias podem ser também descritas como: (i) agilidade analítica, (ii) facilidade de utilização, (iii) fonte de suporte para a reflexão, (iv) ferramenta de comunicação.

Figura 4 – Composição do Visual Analytics



Fonte: (VIVA, 2015, p.8)

A proposta de visualização de dados estatísticos com interatividade surgiu de modo mais estruturado na International Conference on Data Visualization que aconteceu em 1998. Em 2002, foi publicado o estudo Polaris que fundamentou o Software Tableau (STOLTE; TANG; HANRAHAN, 2002). No texto que apresenta o Polaris os autores frisam que a proposta vem em enfrentamento aos limites de explorar bases de dados grandes e multidimensionais. Os recursos apresentados incluem uma interface para construir diferentes explicações visuais e a capacidade de realizar consultas específicas nas diferentes explorações visuais. Os autores apontam que o desenvolvimento tecnológico permitindo carregar uma grande quantidade de dados na memória do computador possibilitou a prática de operações com maior desempenho, sem a latência da leitura e escrita dos dados em disco. Isto serviu para a melhoria das funcionalidades interativas dos sistemas de *Visual Analytics*. A partir de 2009 o uso de *visual analytics* se ramificou para as mais diferentes áreas tais como finanças e saúde pública e produtos como o Tableau que suportam esta exploração interativa ganharam destaque e vêm sendo tomados como base para novas versões e soluções tecnológicas.

Um exemplo de produto advindo da aplicação de técnicas do *visual analytics* em *big data* está publicado no site da empresa Flowingdata⁸. A partir de dados pessoais e de emprego de americanos é possível, por exemplo, refletir sobre o potencial de uma determinada região ou estado para a instalação de uma determinada indústria ou comércio. O produto

⁸ <http://flowingdata.com/2014/07/02/jobs-charted-by-state-and-salary>

correlaciona, dentre outros, o número de pessoas em uma determinada faixa salarial por tipo de emprego e estado americano. É possível explorar o país como um todo ou estado a estado. O autor comenta, por exemplo, as diferenças possíveis de serem identificadas entre a Califórnia e Washington:

For example, look at California. You see an increased prominence of farmworkers and laborers, whereas the farming, fishing, and forestry sector is nearly nonexistent in many other parts of the country. For a drastic change, switch to Washington, D.C., where people who work in the legal and business sectors are much more common. (YAU, 2014)

No Brasil, Carmo, Shimabukuro e Alcântara (2016) publicaram seu estudo que monitorou a coleta e a qualidade de dados ambientais por sensores por meio de técnicas de *visual analytics*. Os autores adotam o *visual analytics* em decorrência da coleta sistemática de dados ambientais realizadas por sensores produzir grande quantidade de dados temporais multivariados. O estudo utilizou do *visual analytics* para a extração de características do conjunto de dados tais como disponibilidade dos dados; funcionamento dos sensores; e evidências de padrões de falhas. Os gráficos interativos facilitaram a extração de informações de forma rápida e intuitiva, em contraste com o esforço que seria necessário para extrair as mesmas informações a partir de dados tabulares.

Outro grupo de pesquisa brasileiro que tem se dedicado a estudar o manejo de *big data* é o grupo coordenado pelos pesquisadores Christovam Barcellos e Marcel Pedroso do Laboratório de Informação em Saúde do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde da Fundação Oswaldo Cruz – LIS/Icict/Fiocruz. Este grupo foca nos desafios de *big data* em saúde e busca atender ao SUS. Os objetivos descritos na sua página eletrônica apontam para as diferentes temáticas e interfaces do *big data*, dentre elas, (i) a coleta e o processamento de informações; (ii) a interface de comunicação dos dados; (iii) a segurança das plataformas que sustentam a correlação dos dados e (iv) o desenvolvimento de aplicativos relacionados com a mineração e a visualização dos dados. São apontadas como linhas de pesquisa: (i) análise preditiva e algoritmos para mineração de dados e de textos, (ii) análise visual de dados para tomada de decisão em saúde, e (iii) infraestrutura, armazenamento e governança de dados em ecossistema Hadoop⁹. O grupo se propõe ainda a trabalhar com dados não estruturados como, por exemplo, na análise de mídias sociais. Os produtos, até o momento, incluíram o Seminário Ciência de Dados aplicada à Saúde e a formalização de curso de atualização com o mesmo nome.

⁹ Hadoop é uma plataforma de software em Java de computação distribuída voltada para clusters e processamento de grandes massas de dados - <https://pt.wikipedia.org/wiki/Hadoop>

6. Objetivos

Desenvolver uma metodologia com vistas a explicitar o padrão de financiamento público de pesquisa em dengue no Brasil a partir do Diário Oficial da União.

Mais especificamente, almeja-se:

- i. Identificar e modelar as dimensões representativas do padrão público de financiamento de pesquisa em saúde no Brasil;
- ii. Identificar e elaborar, a partir da modelagem do item anterior, um vocabulário de busca para financiamento de pesquisas em/para saúde para o DOU;
- iii. Modelar a extração de dados sobre pesquisa em dengue a partir do DOU;
- iv. Apresentar modelo visual que explicita o padrão de financiamento identificado.

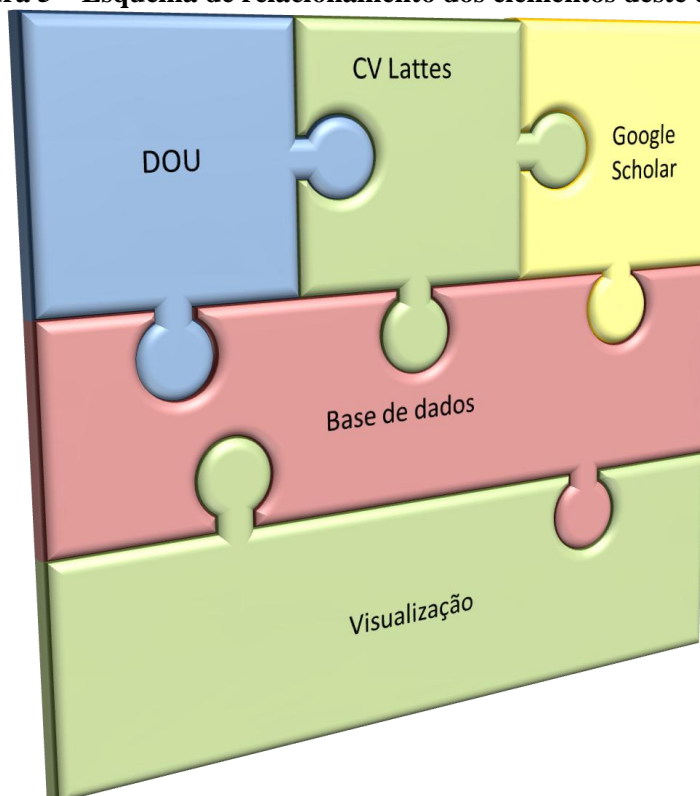
7. Metodologia

A proposta metodológica foi baseada nas possibilidades que a tecnologia proporciona de facilidades ao que se relaciona com a coleta e tabulação dos dados, assim como para com a visualização dos resultados.

Este capítulo foi estruturado em duas partes sendo na primeira apresentadas as fontes de dados, como elas se relacionam e os recursos tecnológicos envolvidos; e na segunda parte é apresentado o processo de coleta, armazenamento e extração dos dados, assim como a tabulação dos mesmos e procedimento de análise.

A figura a seguir, Figura 5, lista o conjunto de fontes que foi mobilizada para constituir a metodologia que vai ser apresentado ao longo deste capítulo.

Figura 5 – Esquema de relacionamento dos elementos deste estudo



Fonte: Elaboração própria

7.1. As fontes de dados e os instrumentais utilizados

A pesquisa se baseia em dados secundários disponíveis no Diário Oficial da União (DOU), no Currículo Lattes e no Google Scholar.

O DOU é uma fonte de dados pública e está disponibilizado eletronicamente no site da Imprensa Nacional (IN) conforme descrito anteriormente. O site da IN possui uma ferramenta de busca que permite recuperar páginas do DOU a partir de um termo dado. É possível buscar

pela grafia exata, bem como, buscar pela fonética. Um limite desta ferramenta de busca disponibilizada no site da IN diz respeito a variável tempo. A recuperação tem que se limitar em estar dentro de um mesmo ano. Ou seja, é possível realizar uma busca de 01/01/2010 a 31/12/2010 (um ano), porém não é possível buscar de 20/12/2010 a 20/01/2011 (um mês) por mudar o ano.

O DOU é um documento escrito em português (pt_BR¹⁰). Textos escritos correspondem a uma interpretação formal da linguagem natural (LN). Para que este tipo de texto seja entendido por máquinas faz-se necessário aplicar um conjunto de métodos e técnicas de um subcampo da Inteligência Artificial e da Linguística: o Processamento de Linguagem Natural (PLN). O PLN objetiva fazer com que os computadores entendam o que é escrito em linguagem humana. Para Chopra, Prashar e Sain (2013) linguagem humana é aquela falada ou escrita pelas pessoas para comunicação. A linguagem natural é definida por Lopes (2002) como a linguagem do discurso técnico-científico. Liddy (2001) acrescenta que o PLN é um conjunto de teorias e um conjunto de técnicas. A autora apresenta como definição:

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.¹¹
(LIDDY, 2001)

Uma das técnicas utilizadas no PLN é a mineração de textos. A mineração de textos consiste na obtenção de informações relevantes tendo como fonte textos descritos em linguagem natural. Aranha e Passos (2006, p.1) acrescentam que a partir da mineração é possível “extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural”.

A utilização da mineração de textos pode ser das mais diversas. Collier e colaboradores (2008), por exemplo, utilizaram a mineração de textos em páginas web para detectarem rumores sobre a saúde pública. Nesta mesma linha, Corley e colaboradores (2010) realizaram coleta de comentários (*posts*) da web e redes sociais com a finalidade de descobrir as tendências da incidência de casos de influenza. A mineração de textos também pode ser usada como técnica para padronização de termos. Morel e colaboradores (2009) realizaram análise em rede de coautoria de publicações da base de dados *Web of Knowledge do Institute for Scientific Information* (ISI) com a finalidade de realizar Planejamento Estratégico de

¹⁰ Para a computação o português utilizado no Brasil é representado como pt_BR

¹¹ Natural Language Processing é uma gama de técnicas computacionais para a análise e representação de textos que ocorrem naturalmente em um ou mais níveis de análise linguística com a finalidade de alcançar o processamento da linguagem parecida com a humana para uma série de tarefas ou aplicações.

Pesquisa, Desenvolvimento e na construção de Programas de Capacitação em doenças negligenciadas.

As técnicas de mineração de texto também podem ser utilizadas para identificar “palavras relevantes” em textos pela frequência com que elas aparecem, ou seja, quanto maior o número de vezes que uma determinada palavra aparece no texto, mais “forte” ela é em relação ao texto. Outra possibilidade é verificar a correlação de termos por sua proximidade, isto é, identificar se duas entidades possui algum tipo de vínculo, caso seus respectivos nomes estejam próximos um do outro no texto.

Entende-se que a Análise Visual pode agregar valor à mineração de textos, pois a primeira possibilita encontrar não somente o esperado, mas o inesperado (“About Visual Analytics”, [s.d.]). A associação do *Visual Analytics* ao *Data/Text Mining* leva ao conceito de *Knowledge Discovery and Data Mining* (KDD) o qual possui como principal objetivo a busca de padrões. O KDD pode ser aplicado em diversas áreas como médica, biológica, financeira e linguística (KEIM, 2010). Ademais a visualização é fundamental para o entendimento de fenômenos complexos, pois este campo tem o objetivo de descobrir padrões e tendências (LANE, 2010).

A estrutura tecnológica foi toda baseada em software livre e composta inicialmente por um servidor com o sistema operacional (SO) Linux e pequenos programas (*scripts*) escritos na linguagem Python.

O SO Linux foi desenvolvido inicialmente por Linus Torvalds em 1991. O estudante finlandês de computação estava disposto a criar um núcleo de sistema operacional compatível com os computadores de pequeno porte e semelhante ao UNIX que rodava em computadores de médio e grande porte. A linguagem de programação Python também teve seu lançamento em 1991. A linguagem Python possui um conjunto amplo de bibliotecas que a potencializam. O ambiente foi preparado para receber os programas desenvolvidos que realizam a coleta de dados e tabulação dos mesmos. A linguagem de programação Python é flexível, modular e prioriza a legibilidade do código tornando sua curva de aprendizado acentuada, o que caracteriza um rápido aprendizado em um curto tempo.

Esta linguagem de programação possui diversas bibliotecas para diferentes finalidades, como por exemplo, acesso à internet, funções matemáticas complexas, desenho de redes sociais, *parsing*¹² de arquivos no formato *eXtensible Markup Language* (XML) ou *HyperText Markup Language* (HTML) e, dentre outras, para desenvolvimento de jogos.

¹² Parsing: leitura de fonte de dados de forma estruturada.

Conforme mencionado, a linguagem de programação Python é modular. Isto significa que a ela podem ser incorporadas bibliotecas de códigos com funções utilizadas para fins específicos. Existem bibliotecas de códigos que, ao serem incorporadas no programa, permitem que o mesmo acesse um banco de dados, possibilitando incluir, alterar e apagar registros; dependendo, é claro, das permissões que o responsável pelo banco de dados configurou para aquele acesso. Outro exemplo são as bibliotecas gráficas que permitem a criação de imagens dos mais variados tipos, desde polígonos simples, até gráficos de barras, linhas ou setoriais. As possibilidades são inúmeras.

Este trabalho se valeu de uma série de bibliotecas de códigos que aumentou o potencial da linguagem Python em relação ao objetivo traçado. A seguir, foram descritas as principais bibliotecas que foram utilizadas com suas respectivas descrições, Quadro 2.

Quadro 2 – Bibliotecas de funções Python usadas para a coleta dos dados

Nome da biblioteca	Descrição
BeautifulSoup	Biblioteca de métodos para análise da estrutura de páginas/arquivos HTML/XML
Codecs	Biblioteca que contém métodos que trabalham com diferentes codificações de caracteres. (ex.: UTF-8, ISO-8859-1, etc).
cookielib	Biblioteca de métodos que gerenciam automaticamente os cookies ¹³ recebidos de páginas web.
datetime	Biblioteca de métodos capazes de manipular datas e horas (ex.: 10/2/2014 + 10 dias = 20/2/2014).
Getopt	Biblioteca de métodos que auxilia o programa a receber parâmetros passados por linha de comando pelo usuário.
Levenshtein	Algoritmo de cálculo de distância entre duas strings
Logging	Biblioteca de métodos voltados para a criação e manutenção de arquivos de log de eventos: informações, erros, avisos, etc
Mechanize	Biblioteca de métodos para navegação na internet. Voltado para manipular formulários na web, automatizando o preenchimento.
openCV	Biblioteca de métodos para trabalhos com imagens estáticas e em movimento.
Os	Biblioteca de métodos que provê acesso ao sistema operacional. Permite, entre outros, navegar pelos diretórios (pastas) do sistema.
Re	Biblioteca de métodos que provê operações de expressões regulares ¹⁴
Shlex	Biblioteca de métodos capazes de auxiliar na análise léxica
String	Biblioteca que contém constantes úteis como pontuação, letras, etc.
Sys	Biblioteca de métodos que provém funcionalidades específicas do sistema. Como, por exemplo, acesso à saída de vídeo (stdout) e entrada de dados (stdin).
Time	Biblioteca de métodos específicos para manipulação de dados de

¹³ “grupo de dados trocados entre o navegador e o servidor de páginas, colocado num arquivo (ficheiro) de texto criado no computador do utilizador. A sua função principal é a de manter a persistência de sessões HTTP “ (Wikipedia, 2014)

¹⁴ Expressão regular é uma seqüência de caracteres que forma um padrão de pesquisa. Muito utilizado para operações de busca e substituição. (Wikipedia, 2014)

Nome da biblioteca	Descrição
	hora.
urllib2	Biblioteca de métodos que utilizam protocolos de comunicação para internet (ex.: HTTP), permitindo acesso às páginas web com possibilidade de autenticação e redirecionamentos.

Fonte: Site The Python Package Index (<https://pypi.python.org/pypi>)

Algumas destas bibliotecas também auxiliaram a recuperação dos dados do Currículo Lattes o qual pertence à Plataforma Lattes. Aqui, as tecnologias de informação e comunicação (TICs) são as grandes aliadas para auxiliar nos processos de gestão. A Plataforma Lattes é um conjunto de sistemas de informação, bases de dados e portais da internet, concebido para integrar os sistemas de informação das agências federais, racionalizando o processo de gestão de C&T. Lançada em 16 de agosto de 1999, proporcionou um aumento significativo do número de CVs enviados ao CNPq, que chegou a mais de 100 por dia. Atualmente, o Currículo Lattes possui cerca de 3.098.215 currículos (BRASIL, [s.d.]

A Plataforma Lattes é uma base de dados composta pelo Currículo Lattes, Diretório de Instituições, Diretório dos Grupos de Pesquisa. O Currículo Lattes é um aplicativo que surgiu para apoiar a identificação e troca de informação entre pesquisadores brasileiros. Segundo o CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – Ministério da Ciência e Tecnologia), que administra atualmente a Plataforma Lattes, a necessidade de registro de currículos dos pesquisadores brasileiros por meio de um formulário padrão que proporcionasse informação sobre a distribuição da pesquisa científica no Brasil vem desde os anos oitenta do século passado. Esta demanda não parece estar localizada apenas no Brasil, ainda que este detenha o pioneirismo. Portugal lançou uma base de dados semelhante chamada de DeGóis (<http://www.degois.pt>) e, mais recentemente, o Canadá lançou o Canadian Common CV (<https://ccv-cvc.ca/>).

Segundo Mena-Chalco e Cesar Junior (2009) as instituições de ensino e pesquisa no Brasil se utilizam constantemente do Currículo Lattes para extrair dados a fim de gerar relatórios sobre produção científica, orientações e avaliação da produção de grupos de pesquisa. Os autores acrescentam que se trata de uma tarefa árdua, pois não se possui a informação estruturada. Ademais a tarefa se torna suscetível a erros acarretando no tratamento manual dos resultados.

A Plataforma Lattes já tem um dispositivo que permite alguns cruzamentos, chamado de Painel Lattes. Entretanto, este dispositivo não proporciona, por exemplo, o cruzamento de dados com bases externas à Plataforma Lattes nem tampouco a visualização específica de informações importantes para as instituições, como o monitoramento da tipologia da produção

de seus profissionais. Outro dispositivo existente para extração e compilação automática dos dados do Currículo Lattes é o scriptLattes. Esta ferramenta é customizada para extrair dados relacionados à produção científica e, embora seja possível utilizá-la para a extração da produção artística e técnica, pouco é utilizado neste sentido. Em outras palavras, o scriptLattes avança na direção da extração guiada, mas também não realiza cruzamento com outras bases de dados de informação.

Recentemente o CNPq implantou a utilização de *captcha* (Completely Automated Public Turing Test To Tell Computers and Humans Apart). *Captcha* é um tipo de proteção ao site para que programas automatizados não acessem seus dados. Com isso, o acesso aos dados dos currículos dos pesquisadores se torna moroso, dificultando a análise dos mesmos. Entretanto a utilização da ferramenta ScriptLattes desenvolvida por um grupo de pesquisa da Universidade de São Paulo (USP) em conjunto com a ferramenta Web ScriptLattes (WSL) que proporciona a interação visual com a primeira, facilita em muito a recuperação e análise dos currículos dos pesquisadores. Mesmo assim a recuperação da produção científica relacionada à pesquisas sobre Dengue ainda é um desafio.

É neste momento que o Google Scholar se torna uma oportunidade no acesso à produção científica do país. O Google Scholar, ou Google Acadêmico em português, é uma iniciativa da empresa Google para se buscar a literatura acadêmica. O Google Scholar proporciona uma forma simples de busca da literatura científica. Esta forma simples é baseada na ideia de a partir de um único lugar ser possível buscar em diferentes fontes e, por conseguinte, diferentes áreas de conhecimento, por artigos, livros, teses, resumos e outras tipologias documentais. Também é possível exportar suas referências para softwares de gerenciamento de referências como o EndNoteWeb.

O Google Scholar permite a busca em bases que indexam a literatura científica. Usualmente estas bases são pagas para que se tenha acesso, entretanto por acordos comerciais entre o Google Inc.[©] e os detentores destas bases, é possível ter acesso às citações. Todavia, para ter acesso ao texto completo, muitas vezes é necessário permissão de acesso por meio de assinatura com os editores científicos ou os detentores das bases científicas.

A ferramenta de busca do Google Scholar permite filtrar por ano, citações e patentes; assim como buscar por termos em campos específicos como título ou em outra parte da produção científica e, por fim, busca por autores.

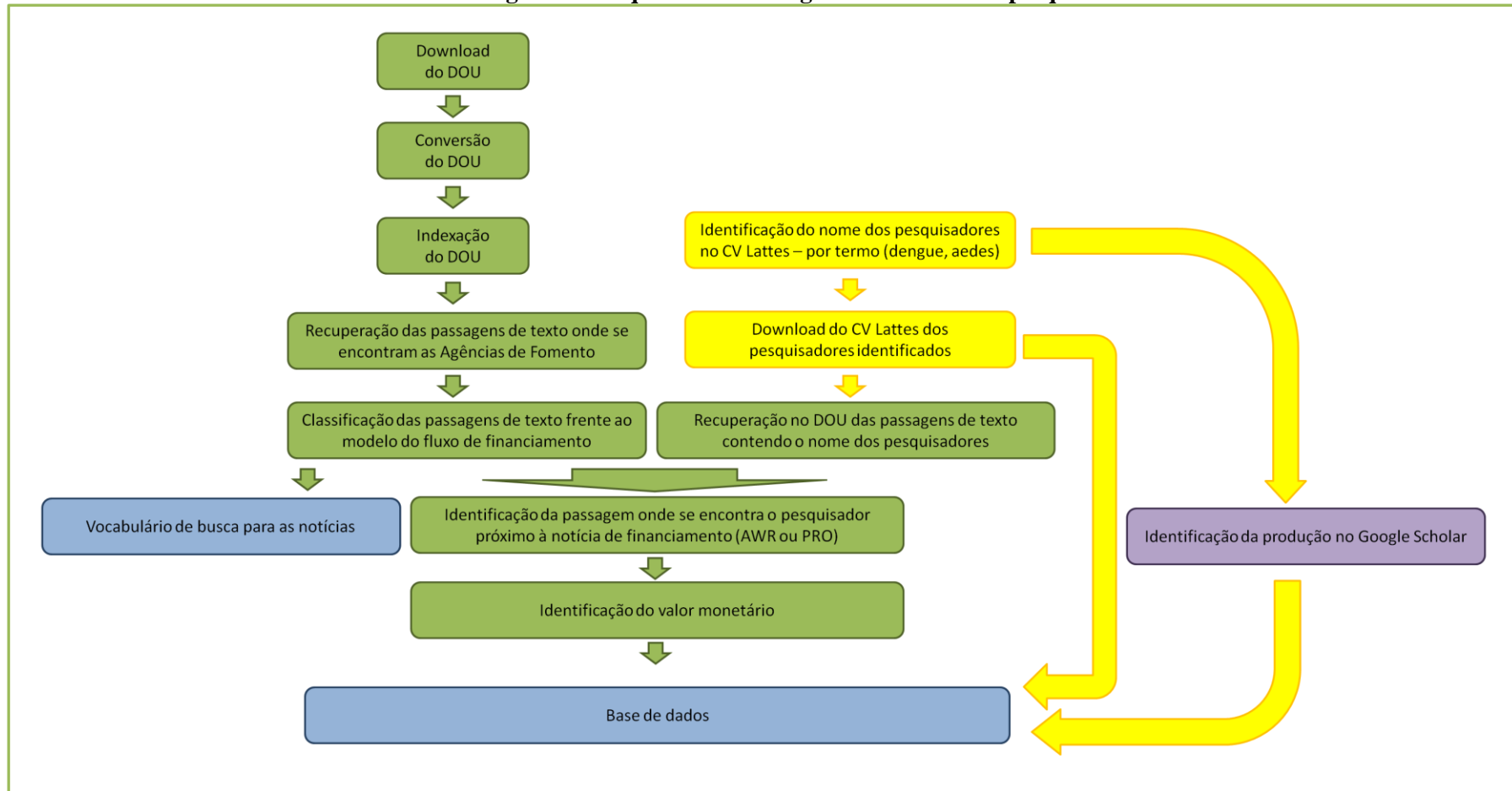
A partir destas três fontes de dados – Diário Oficial da União, Currículo Lattes e Google Scholar – foi possível constituir o nível superior da plataforma metodológica deste

estudo. Tanto o Currículo Lattes quanto o Google Scholar, que fizeram parte deste estudo podem ser substituídos por outros sistemas ou bases que convier ao estudo como uma base local à instituição para os pesquisadores e uma base como Medline ou Scielo respectivamente. Entretanto, ressalta-se que esta articulação de fontes é necessária para esta modelagem e que estas fontes embora não permitam precisão, são fontes de aproximação.

7.2. Passos da Pesquisa

O objetivo de desenvolver uma metodologia para explicitar o padrão de financiamento público de pesquisa em dengue no Brasil a partir do Diário Oficial da União se confunde com os próprios passos desta pesquisa. Sendo assim, optou-se por iniciar esta subseção apresentando o esquema representativo do modelo de extração de dados do DOU, da identificação dos pesquisadores de dengue na plataforma Lattes e da recuperação do quantitativo de artigos publicados indexados pelo Google Acadêmico. Este esquema, Figura 6, traz a sequência realizada para cada um destes processos e como elas se interligam, tendo como desfecho a criação de uma base de dados.

Figura 6 – Esquema metodológico utilizado nesta pesquisa



Fonte: Elaboração própria

O texto a seguir descreve o esquema apresentado na página anterior apresentando como foi realizada cada etapa com seus percalços e motivos da escolha de um caminho determinado.

Isto posto, de sorte a atender o objetivo desta pesquisa foi necessário preparar o campo, ou seja, recuperar o corpus do Diário Oficial da União (DOU) e entender suas características técnicas. Tal entendimento foi de fundamental importância para determinar as limitações desta fonte de dados para a pesquisa.

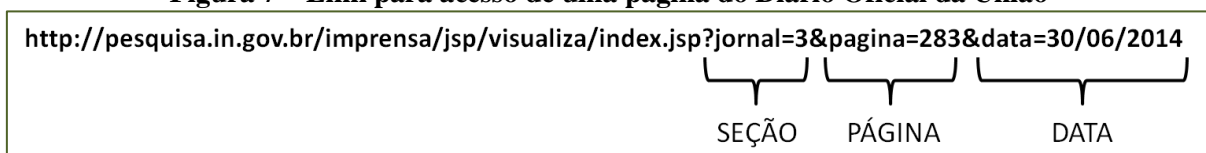
O Diário Oficial da União (DOU) é disponibilizado no formato Portable Document Format mais conhecido pelo seu acrônimo PDF. Cada página do mesmo equivale a um arquivo independente totalizando, em média, quatrocentos e cinquenta arquivos por dia de publicação.

A coleta e a tabulação dos dados foram realizadas automaticamente por *scripts*. *Scripts* são pequenos programas de computador com finalidade específica. Para fins desta pesquisa os *scripts* foram desenvolvidos em linguagem Python e utilizaram as bibliotecas de métodos descritas anteriormente.

A primeira etapa realizada foi a coleta de dados efetuando o *download* do DOU para um servidor Linux capaz de armazenar os arquivos coletados. Tecnicamente, o DOU é disponibilizado na rede mundial de computadores – *internet* – por página, ou seja, é possível recuperar cada página do DOU independentemente, porém não é possível visualizar ou baixar todo um número do DOU em um arquivo único. Estas páginas são disponibilizadas no formato Portable Document Format (PDF).

O link abaixo representa a recuperação de uma página do DOU, Figura 7:

Figura 7 – Link para acesso de uma página do Diário Oficial da União



Fonte: Elaboração própria

Como o processo manual de captura de cada uma das páginas do DOU seria muito moroso, foi necessário desenvolver um *script* que fizesse o *download* automático do Diário Oficial da União disponível no *site* da Imprensa Nacional. A Imprensa Nacional disponibiliza cada página de cada número do DOU datado de 2 de janeiro de 1990 até os dias de hoje, ou seja, conforme descrito, não há possibilidade de baixar o número do DOU por completo, apenas suas páginas uma a uma. Esta é uma característica importante da fonte de dados, pois direciona o formato de armazenamento de toda a fonte de dados.

As páginas foram armazenadas em uma estrutura de diretórios organizadas por ano e meses. Para facilitar o entendimento de que página pertence a que número do DOU, as páginas que foram baixadas, ao serem salvas como arquivo no disco do servidor, receberam seus nomes de acordo com uma regra que corresponde à data de publicação, seção e página baixada. A regra definida para o nome de cada arquivo é: DATA-SEÇÃO-PÁGINA.pdf. A seguir, cada uma das partes do nome do arquivo é apresentada no Quadro 3.

Quadro 3 – Formato do arquivo salvo

PARTE	FORMATO	EXPLICAÇÃO	EXEMPLO
DATA	AAAAMMDD	Ano com 4 algarismos Mês com 2 algarismos Dia com 2 algarismos	20140630
SEÇÃO	N	Seção com 1 algarismo	3
PÁGINA	NNN	Página com 3 algarismos	283

Fonte: Elaboração própria

Para fins de exemplificação, a página representada pelo link da Figura 7, quando salva em disco, recebera o nome 20140630-3-283.pdf.

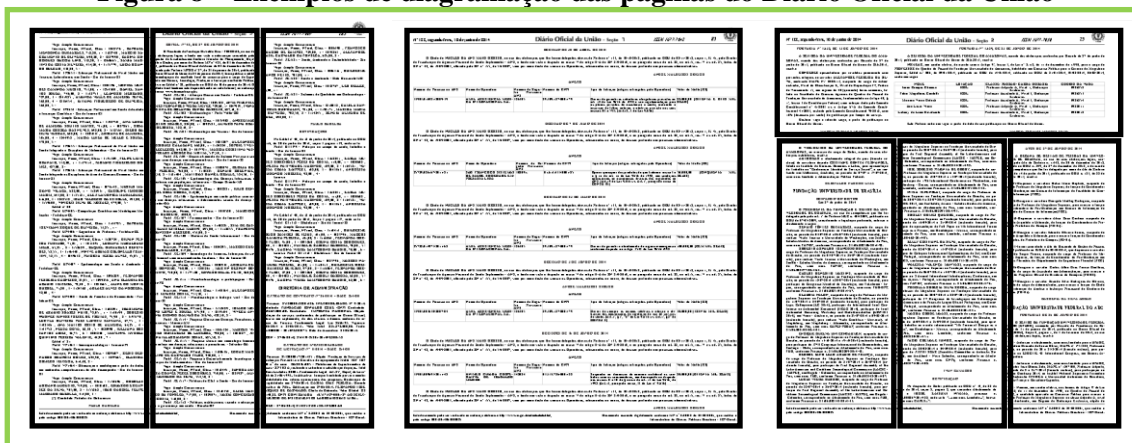
Para que fosse possível que o computador lesse os arquivos no sentido de indexar o texto e minerar os dados, foi necessário converter os arquivos do formato PDF para o formato texto, ou simplesmente TXT. O formato TXT foi o primeiro formato digital para arquivos texto e que perdura até hoje por facilitar a manipulação do texto nele contido. A automatização de uma série de operações quando se utiliza este formato é bem simples.

Após baixar um número completo, um *script* realiza a conversão de cada uma das páginas do formato *Portable Document Format* (PDF) para o formato texto (TXT). Este último legível por qualquer programa e de fácil indexação. É possível, em arquivos tipo “TXT”, buscar, incluir, retirar ou substituir palavras, assim como alterá-las utilizando comandos que se baseiam em regras.

Em um arquivo no formato DOC, DOCX ou PDF é possível buscar uma determinada data (ex.: 30/06/2014). Porém em um arquivo no formato TXT, utilizando os comandos pertinentes e sem nenhum outro subterfúgio, a recuperação de todas as datas é facilitada se a busca for baseada na regra “dd/dd\dddd”, onde a letra “d” representa qualquer algarismo numérico. Os arquivos no formato TXT são salvos em uma estrutura análoga à primeira descrita anteriormente e representada no Quadro 2, ou seja, os arquivos receberam o mesmo nome que o original exceto a extensão que, ao invés de ser “pdf” passou a ser “txt”. Além disso, os arquivos também são salvos em diretórios organizados pelos anos e meses.

O trabalho de conversão, embora inicialmente pareça simples, é de extrema complexidade. A complexidade é resultado de questões diversas como hifenização de palavras, mudança de parágrafo, existência de tabelas e principalmente pela diagramação do DOU. Uma página do DOU pode possuir uma, duas ou três colunas. Porém, em uma mesma página, é possível encontrar uma combinação de quantidade de colunas conforme exemplificado na Figura 8.

Figura 8 – Exemplos de diagramação das páginas do Diário Oficial da União



Fonte: Elaboração própria

É possível verificar acima diferentes diagramações do DOU referentes às páginas 285 da seção 3; 85 da seção 1 e 25 da seção 2; sendo todas de 30/06/2014. Ressalta-se aqui que, embora as diferentes diagramações tenham sido apresentadas em seções diferentes, o mesmo pode ocorrer dentro de uma mesma seção.

Sendo assim, buscou-se na internet uma ferramenta que fizesse a conversão do formato PDF para o formato TXT. Verificou-se que muitas são as ferramentas que fazem este tipo de tarefa, porém cada qual possuía uma limitação que dificultaria o trabalho.

A primeira ferramenta testada foi a **pdf2htmlEX**. Esta ferramenta é fruto da pesquisa de Wang e Liu (2013) e tem como objetivo a conversão do formato PDF para publicação na web. Seu resultado visual é muito bom. Porém, o código gerado em HTML, o que, inicialmente, não seria um problema para realizar o *parsing*¹⁵, é deveras complexo, o que torna sua compreensão por máquina de mesma ordem.

Outra ferramenta testada foi o **pdfreflow**. Esta necessita que haja uma conversão prévia utilizando o **pdftohtml** disponível no próprio sistema operacional Linux utilizado. A primeira conversão gera um arquivo formato *Extendable Markup Language* (XML) e a segunda converte de XML para TXT. Entretanto, o pdfreflow não suporta texto em colunas, o

¹⁵ Descrito anteriormente.

que é totalmente necessário para trabalhar com o DOU, uma vez que encontramos em sua diagramação páginas com uma, duas ou três colunas.

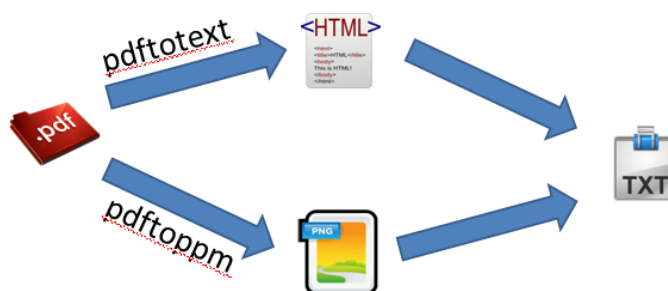
Uma ferramenta com um bom potencial para trabalhar as páginas do DOU é a **pdfbox-app-1.8.9.jar**. Desenvolvida na linguagem JAVA, esta trabalha bem as páginas, entretanto alguns problemas no tratamento de colunas e reconhecimento de tabelas dificultariam a conversão.

A ferramenta **pdf2txt** é um *script* Python que chama a biblioteca PDFMiner (biblioteca Python que abre e lê o PDF). Este *script* possui o parâmetro “-Y” que recupera as coordenadas x1-y1 e x2-y2 para cada letra contida no texto. As coordenadas x1-y1 correspondem ao canto superior esquerdo da letra e as coordenadas x2-y2 ao canto inferior direito da letra. A partir destas coordenadas, seria possível determinar a posição da letra em relação ao texto. Porém, seria difícil trabalhar os espaços entre as palavras.

Mais uma ferramenta avaliada foi a **PdfMasher**. É uma ferramenta que funciona em modo gráfico no Windows e, embora esta seja boa para recuperar a ordem do texto, o fato de trabalhar no modo gráfico e principalmente necessitando interação humana dificultaria em muito tornar o processo automático.

De um modo geral, cada ferramenta possuía uma ideia interessante que foi aproveitada para a definição do processo de conversão. Optou-se por fim na construção de uma ferramenta própria (*script*), que lançou mão do melhor de cada uma das ferramentas avaliadas anteriormente, apresentado no esquema a seguir, Figura 9:

Figura 9 – Processo de conversão de uma página do Diário Oficial da União para texto



Fonte: Elaboração própria

Optou-se pela utilização da ferramenta **pdftotext**, pertencente à *poppler tools*, que já acompanha o sistema operacional Linux e não depende de nenhum programa externo para seu funcionamento. Para este fim, o **pdftotext** converte a página do DOU para um arquivo no formato HTML contendo as coordenadas (x1-y1 e x2-y2) de cada **palavra** analogamente ao que o **pdf2txt** realiza com as letras conforme descrito anteriormente. O mesmo arquivo PDF é convertido para imagem utilizando a ferramenta **pdftoppm**. Foram mantidas as dimensões

altura e largura da página as quais podem ser encontradas no arquivo gerado pelo pdftotext. O pdftoppm também é nativo ao sistema operacional Linux e, assim como o pdftotext, compõe as ferramentas do *poppler tools*.

A conversão para imagem faz-se necessária para o processo de identificação das tabelas. Para tanto foi utilizada uma biblioteca Python chamada Open Computer Vision, ou simplesmente OpenCV. Esta biblioteca tem a finalidade de trabalhar imagens estáticas e em movimento. Uma vez que o objetivo deste trabalho não é o reconhecimento de tabelas ou conversão de PDF em texto, foi estipulado então que somente seriam trabalhadas as tabelas “perfeitas” ou “semiperfeitas”. Esta denotação se deve à característica de, no primeiro caso, cada célula da tabela estar contida dentre linhas ou como no segundo caso onde pode haver acoplamento de células (*merge*), porém em ambos os casos sempre será encontrado o quadro externo. A figura a seguir, Figura 10, representa os formatos de tabela que são detectados para a conversão para o formato TXT:

Figura 10 – Exemplo de tabelas trabalhadas na conversão de PDF para texto

			MERGE CELL			
MERGE CELL						

Fonte: Elaboração própria

Para a detecção das tabelas foi utilizado o método **Hough Transformation**. Na análise automatizada de imagens digitais, muitas vezes surge a etapa de detectar formas simples, como linhas retas, círculos ou elipses. Para tanto, foi utilizado um processo anterior de detecção de bordas (edge detection) para obtenção de pontos (pixels) da imagem que estão na curva/reta desejada. Este processo foi utilizado para detecção das linhas **verticais** contidas na imagem gerada.

Na figura a seguir, Figura 11, a exemplificação da detecção de linhas verticais e horizontais em uma página do Diário Oficial da União. Em azul encontram-se as linhas verticais e em vermelho as linhas horizontais detectadas na imagem.

Ressalta-se que para fins desta conversão só utilizaram-se as linhas verticais detectadas para a definição de áreas na imagem contendo possíveis tabelas.

Figura 11 – Exemplo de detecção de linhas verticais e horizontais por Hough Transform



Fonte: Elaboração própria

As subáreas detectadas foram recordadas foram recordadas do texto e analisadas utilizando a biblioteca Python Image Library (PIL) que permite a verificação de cada ponto (pixel) da imagem. Verificou-se, portanto, a existência de linhas horizontais que tocassem ou cruzassem as linhas verticais detectadas e por fim se este conjunto de linhas formava uma tabela.

A partir dos dados de coordenadas das células obtidas nesta etapa em conjunto com os dados das coordenadas das palavras obtidas na primeira etapa, é gerado o texto referente à página do DOU. Este processo de conversão dura em média sete segundos para cada página, ou seja, cerca de uma hora para um número inteiro do DOU.

Para a delimitação das células de uma linha em uma tabela foi utilizado o conjunto de caracteres *underline-pipe-pipe-underline*, equivalente a “||” (sem aspas).

A figura a seguir, Figura 12, representa por meio de um esquema gráfico as ações destas etapas descritas: captura das páginas do DOU, conversão para o formato TXT.

Figura 12 – Processo de download e conversão das páginas do DOU



Fonte: Elaboração própria

Todo este processo é executado automaticamente e diariamente para que se mantenha um conjunto de dados atualizados de acordo com o que é publicado pela Imprensa Nacional.

Ressalta-se aqui que todos os números do DOU do ano de 1990 até o ano de 2001 inclusive não foram gerados eletronicamente. Aqueles números passaram por um processo de digitalização reconhecimento ótico de caracteres, em inglês *optical character recognition* (OCR).

O processo de conversão dos documentos tem o intuito de tornar os documentos passíveis de busca por meio da utilização de termos. Entretanto é um processo demorado, custoso e não apresenta tantos benefícios, pois alguns caracteres são reconhecidos automaticamente de forma equivocada. É possível exemplificar com a passagem representada na figura abaixo, Figura 13, retirada da página 143, Seção 2 do DOU de 3 de maio de 1990.

Figura 13 – Exemplo de texto retirado do DOU de 3/5/1990, Seção 2, página 143
DAS EM NCZ\$40,00 (QUARENTA CRUZADOS NOVOS); 06)UM SECADOR DE CABELOS ' MARCA SIKASIWA, Nº 711128, EM ÓTIMO ESTADO DE CONSERVAÇÃO E FUNCIONA = MENTO, AVALIADO EM NCZ\$150,00 (CENTO E CINQUENTA CRUZADOS NOVOS); 7) ...

Fonte: Diário Oficial da União, 3/5/1990

O texto que aparece é “[...] 06) UM SECADOR DE CABELOS MARCA SIKASIWA, Nº 711128, EM ÓTIMO ESTADO DE CONSERVAÇÃO E FUNCIONA – MENTO, AVALIADO EM NCZ\$150,00 (CENTO E CINQUENTA CRUZADOS NOVOS);[...]”. Entretanto, ao verificar o que foi reconhecido após passar pelo processo de reconhecimento ótico de caracteres é “()CUM SECADOR DE CABEIAS MARCA SIMASIWA, U 2 711128, EM &mo ESTADO DE' CONSERVAÇXO E FUNCIONA =1 =20, AVALIADO EM NCZS150,00 (CENTO 2 CINQUENTA CRUZADOS[...]. Algumas letras como “ó” ou “Ó” têm grande possibilidade de serem reconhecidos como número “6”, assim como a letra “g” com o número “8” ou “9” dependendo da fonte utilizada.

Sugere-se aqui que trabalhos que venham a explorar o Diário Oficial da União de forma automática, busquem a partir de publicações do DOU posteriores a Janeiro de 2002, época que o DOU passou a ser gerado diretamente em PDF. Damasceno e colaboradores (2011) apontam que “fazer uso das informações do Diário Oficial da União é uma tarefa árdua, pois [...] os dados sensíveis [...] não possuem um padrão”.

Ressalta-se que ainda no ano de 2015 existem algumas páginas do DOU onde são encontrados textos em formato de imagem onde o recurso de OCR não foi utilizado, como por exemplo, no DOU de 30/04/2015, seção 1, página 063. Esta página do DOU não tem como ser recuperada a partir da ferramenta de busca por termos.

Contornar estes problemas é muito custoso no que se refere a tempo de desenvolvimento e o benefício não seria tão grande por entender que seriam poucas páginas. Sendo assim, deu-se continuidade ao processo.

Após *download* e a conversão para o formato texto também foi realizada a indexação dos textos utilizando duas ferramentas próprias para este fim: o Whoosh e o ElasticSearch. A escolha de dois processos distintos de indexação se deu por dois motivos: o primeiro motivado pela demora na indexação do Whoosh e para tanto se pensou em uma contingência de índice e o segundo por uma futura comparação entre os dois indexadores. A indexação de um texto permite uma rápida recuperação das páginas que contém o termo buscado. O Whoosh é totalmente integrado ao Python, necessitando apenas que o script que realiza a indexação acesse-o e solicite a indexação. Já o ElasticSearch, embora mais rápido que o primeiro, necessita que o serviço de indexação esteja rodando em algum servidor para que um programa faça a requisição de indexação ou recuperação. Sendo assim, este serviço fica responsável pela indexação e recuperação; e envia o retorno solicitado ao programa que realizou a requisição.

A indexação do texto proporcionou uma nova possibilidade de busca de termos no Diário Oficial da União sem restrição de período de tempo como na ferramenta de busca da Imprensa Nacional. Sendo assim, como primeiro passo para identificar o contexto desta pesquisa no Diário Oficial da União, foram realizadas três buscas no DOU contendo os termos “*aedes aegypt*”, “*dengue*” e mosquito. Verificou-se a distribuição destes termos ao longo dos anos (2005 a 2014) com a finalidade de entender a importância do assunto no decorrer daqueles anos e como estaria distribuída, no contexto do DOU, estes termos. Em outras palavras, verificou-se a relação dos termos com os assuntos publicados no DOU. Sendo que a Seção 1, conforme já descrito, traz leis, decretos, resoluções, instruções normativas, portarias; a Seção 2 publica atos de interesse dos servidores da Administração Pública Federal, enquanto a Seção 3 possui contratos, editais, avisos e ineditoriais. A partir deste levantamento tentou-se identificar o financiamento para pesquisa em dengue. Tendo em vista que o que fora levantado não trouxe dados sobre financiamento, buscou-se outro caminho.

O caminho escolhido com o objetivo de **identificar e modelar o fluxo de informação sobre o financiamento público federal a partir das notícias do Diário Oficial da União** foi o de realizar uma busca englobando as três seções do DOU na base de dados tanto com os nomes das agências de fomento de nível federal quanto por suas siglas, a saber: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de

Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Departamento de Ciência e Tecnologia (DECIT) e Financiadora de Estudos e Projetos (FINEP). Os resultados foram filtrados pelos anos de 2005 a 2014 inclusive. O resultado desta busca foi salvo em um arquivo no formato TXT nomeado **10_search_agencias.txt** onde os campos recuperados foram delimitados pelo caractere especial *pipe* “|” repetido três vezes. Os campos salvos foram data, seção, página, passagem de texto exemplificado abaixo, Quadro 4.

Quadro 4 – Dados recuperados do DOU referentes às agências de financiamento federal

03/01/2005 3 002 Espécie: Termo Aditivo n.º 01.02.0014.01; Data de Assinatura: 29/12/2004; Partes: Financiadora de Estudos e Projetos - FINEP; CNPJ n. o - 33.749.086/0001-09 e FUBRA; CNPJ n. o - 03.151.583/000140; Objeto: Prorrogação de prazos; Prazo de Utilização: 09/01/2006; Prazo de Prestação de Contas: 09/03/2004. Ministério da Ciência e Tecnologia .
--

Fonte: Diário Oficial da União

Entende-se aqui por passagem uma parte do texto onde o termo foi encontrado. Não é possível afirmar que tal passagem seja uma ou mais frases ou ainda um ou mais parágrafos, pois o processo de conversão de PDF em TXT não é preciso na identificação da quebra de linha por conta da dificuldade de identificação automática da diagramação do texto das páginas do DOU.

Neste arquivo cada palavra fora contabilizada e ordenada em ordem decrescente de quantidade para cada ano e seção. Ou seja, quanto maior a frequência em que uma palavra aparece nas passagens maior a possibilidade de sua importância no contexto. Esta contabilização desconsiderou as *stopwords* as quais são palavras com uma alta frequência, porém que não agregam conteúdo ao texto.

A lista de termos em ordem decrescente de quantidade de aparecimento apontou uma ideia do contexto publicado de acordo com a designação de cada uma das seções do DOU.

Tendo em vista que notícias sobre financiamento são publicadas na Seção 3 do DOU, foram verificados na lista de palavras por frequência desta seção, independente do ano, os termos que poderiam ser chave para identificação daquelas notícias.

Verificando que todo o processo utilizaria somente a seção 3, a partir do primeiro arquivo foram recuperados e salvos em um novo arquivo (**15_secao_3.txt**) apenas os registros que constavam na seção 3. Este arquivo possui os mesmos campos, ordem destes campos e delimitador que o arquivo inicial.

Com base neste arquivo foi acrescentado um campo em cada um dos registros contendo o código agência de fomento referente à passagem de texto. Inicialmente, cada agência de fomento recebeu um número de identificação específico conforme a seguir no Quadro 5:

Quadro 5 – Base para identificação das agências

VALOR	AGÊNCIA	SIGLA
1	Conselho Nacional de Desenvolvimento Científico e Tecnológico	CNPq
2	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior	CAPES
4	Financiadora de Estudos e Projetos	FINEP
8	Departamento de Ciência e Tecnologia	DECIT

Fonte: Elaboração própria

A escolha por códigos de identificação não sequencial se deu pelo fato de algumas passagens de texto conter mais de uma agência de fomento. Para estes casos o código de identificação atribuído foi o resultado da soma dos códigos de identificação base. Ou seja, a passagem de texto que contivesse o CNPq e a CAPES, receberia o valor 3 (três). De mesma lógica, caso todas as agências aparecessem em uma passagem de texto, o código, a partir deste modelo, seria 15 (quinze).

O quadro abaixo, Quadro 6, contendo grifos próprios exemplifica uma passagem de texto que recebera o código de identificação 5 (cinco) referente às agências de fomento CNPq e FINEP na mesma passagem de texto.

Quadro 6 – Passagem de texto com código de identificação “5” para agência de fomento

CONTRATADO: R.A S. SHOPPING CENTERS LTDA, GULFSHOPPING S/A, BRINCA BRASIL COMERCIO E REPRESENTAÇÕES LTDA, BERNARD J. KAPLAN SHOPING CENTERS PROMOÇÕES S/A, FLEURY ALLIEGRO IMÓVEIS S/C LTDA, LUIZ ALBERTO GAMA DE MENDONÇA, FUNDAÇÃO DE PREVIDENCIA COMPLEMENTAR DOS EMPREGADOS DA **FINEP**, DO **IPEA**, DO **CNPQ**, DO INPE E DO INPA, CAIXA DE PREVIDÊNCIA DOS FUNCIONÁRIOS DO BANCO DO BRASIL - PREVI, FUNDAÇÃO DE PREVIDENCIA DOS SERVIDORES DO IRB, IRB-BRASIL RESSEGUROS, AD SHOPPING AGENCIA DE DESENVOLVIMENTO DE SHOPPING CENTERS S/C LTDA. OBJETO: Cessão de Espaço no Imóvel sito à Avenida Profa. Izoraida Marques Peres, nº 401, Sorocaba/SP. PROCESSO: 7063.01.064/2004. GERENTE DE FILIAL: Angela Kiyoko Hiramatsu Kakazu. VALOR: R\$ 2.500,00 (dois mil e quinhentos reais) mensais. VIGÊNCIA: 24 meses MODALIDADE: Dispensa de licitação, conforme art.24, Inciso X, da lei nº 8.666/93. RUBRICA: 5704-1 - Aluguel de Imóveis para Uso. DATA ASSINATURA: 06/01/2005

Fonte: Diário Oficial da União, 15/02/2005, Seção 3, página 33

O resultado da codificação das agências gerou o arquivo nomeado **20_identifica_agencias.txt** contendo, além dos mesmos campos do citado anteriormente (15_secao_3.txt), o código de identificação da(s) agência(s) de fomento. Deste arquivo, foram lidas 100 passagens aleatórias com o intuito de classificá-las de acordo com um possível fluxo de notícias que acompanhem o fluxo de financiamento para pesquisa. A identificação do fluxo de notícias não permitiu verificar o encerramento do processo de financiamento. Frente a este quadro buscou-se na lista da frequência das palavras ordenadas decrescentemente alguma que pudesse representar a finalização do processo. Foram encontrados os termos esperados: “prestação” e “contas”. A partir destas, dentro do subconjunto de passagens referenciado anteriormente pertencentes à seção 3 foram recuperadas passagens contendo estes dois

termos. O número de passagens recuperadas foi pequeno. As passagens foram lidas no intuito de encontrar dados sobre a prestação de contas. Foram encontradas “notificações”, “editais” ou “convênios” e “termos aditivos”. Para estes casos, notificação – ou citação – para prestação de contas trata-se de um aviso público feito pela agência de fomento a um determinado pesquisador ou bolsista que não tenha realizado a prestação de contas dentro do prazo, ou seja, uma cobrança para que a prestação de contas seja realizada. As passagens “editais” ou “convênio” referem-se a publicação destes termos propriamente dito contendo os prazos de prestação de contas e referências às Instruções Normativas que regem esta atividade. “Termos aditivos” podem ser renovações de prazo, atualização de valor ou outro ato que indique a continuação do projeto.

A partir da recuperação das passagens descrito nos parágrafos acima, foi possível identificar a ordem de publicação das notícias e assim, **identificar e modelar o fluxo de informação sobre o financiamento público federal a partir das notícias do Diário Oficial da União**, vide o capítulo Resultados.

A frequência das palavras organizadas por ano e seção não aponta para um vocabulário específico para busca em cada uma das etapas identificadas. Sendo assim, para **identificar e elaborar, a partir da modelagem anterior, um vocabulário de busca para financiamento de pesquisas em/para saúde para o DOU** buscou-se classificar as passagens de texto do arquivo **20_identifica_agencias.txt**. Para tal, foram utilizados dois métodos distintos, a saber: (i) classificação utilizando o modelo Naive Bayes e (ii) classificação utilizando frequência do termo–inverso da frequência nos documentos, do inglês *term frequency–inverse document frequency (tf-idf)*.

A classificação Naive Bayes faz parte do grupo de classificadores probabilísticos e é baseado no Teorema de Bayes. Classificadores deste grupo possuem a capacidade de prever qual a classificação um determinado dado receberá baseado na distribuição probabilística sobre um conjunto de classificações. Em outras palavras, a partir de um conjunto de classificações previamente realizado é possível prever qual a classificação de um novo elemento baseando-se na probabilidade daquele frente ao que já fora determinado.

A frequência do termo–inverso da frequência nos documentos (*tf-idf*) por sua vez é baseada no modelo descrito por Salton e McGill (1983). Este método é bastante utilizado em sistemas que necessitam o agrupamento de textos por similaridade. O agrupamento é realizado tomando por base a frequência das palavras contidas no texto, entretanto palavras

que possuam uma alta frequência em mais de uma classificação, perdem “peso” classificatório.

Para esta tarefa, foram classificadas manualmente 1.200 passagens dentro das categorias: chamada para financiamento (CALL), publicação de resultado (RES), recebimento de financiamento (AWR), aditivo para o financiamento (PRO), prestação de contas (CON) e outro. Dentre o subgrupo “outro”, verificou-se a existência de passagens de texto que caracterizavam de forma bem clara “compras” e “concursos”. Passagens contendo estas foram classificadas com o respectivo nome destas duas categorias por entender que as mesmas podem iluminar futuras discussões. Por fim, classificou-se “outro”. Este arquivo recebeu o nome de GOLD.txt por se tratar do padrão ouro da classificação das passagens de texto.

A fim de identificar qual o melhor modelo para a classificação automática das demais passagens foi verificada a acurácia dos métodos Naive Bayes e TF-IDF (*term frequency-inverse document frequency*). Para isto, foi utilizado o método de validação cruzada (*ten-fold cross validation*) onde o conjunto ouro foi dividido em 10 subconjuntos diferentes de teste contendo a mesma quantidade de passagens de texto em cada um. Dois scripts foram desenvolvidos para a avaliação da classificação. O primeiro verificou a classificação realizada pelo método Naive Bayes enquanto a segunda a classificação TF-IDF. A mesma estratégia foi adotada em ambos os scripts, os quais seguiram os seguintes passos: (i) dez por cento do conjunto ouro foi reservado como conjunto de teste, enquanto os outros noventa por cento foram alocados como conjunto de treinamento; (ii) após o treinamento do modelo, o algoritmo classificou o conjunto de teste; (iii) a classificação do conjunto de teste foi comparado com a classificação manual e os resultados foram usados para compor a matriz de confusão. O processo de validação cruzada foi repetido 10 vezes onde cada conjunto de teste foi utilizado somente uma vez assim como o conjunto de treinamento. Cada vez que o script rodava, era usado, naquele caso, um novo conjunto contendo dez por cento das passagens como conjunto de teste e as outras noventa por cento das passagens como conjunto de treinamento.

Mesmo o método *tf-idf* tendo tido um melhor resultado sobre o Naive Bayes, apresentavam-se então três caminhos, a saber: (i) tentar minimizar as diferenças de classificação, (ii) adotar o resultado do método *tf-idf* por ter tido um melhor resultado e (iii) verificar o vocabulário pertinente a cada uma das etapas do financiamento gerada por cada uma das formas de classificação (Naive Bayes e *tf-idf*) e depois compará-los buscando discrepâncias nos mesmos. Considerando a primeira forma muito custosa em relação ao

tempo necessário para minimizar tais discrepâncias; e a segunda forma por estar podendo abrir mão de falsos negativos na classificação, optou-se por gerar o vocabulário calculando a frequência das palavras para cada uma das etapas de financiamento descritas no modelo gerado no objetivo para cada um dos métodos de classificação. De posse do cálculo de frequência, foram listados os termos e a quantidade das 100 palavras com maior frequência para cada etapa de cada método. A partir das listas, verificou-se a interseção entre as mesmas. Ou seja, verificou-se se os termos existentes em uma lista de uma etapa específica para um método de classificação, também aparecia na lista de mesma etapa de financiamento para o outro método. Encontrou-se uma grande quantidade de similaridade nas duas listas para as etapas de financiamento, com exceção para a etapa de resultado (RES).

Neste momento, os dois scripts de classificação automática, baseados respectivamente no Naive Bayes e no TF-IDF, foram rodados para todas as demais passagens de texto gerando como resultado dois arquivos distintos. Identificaram-se então todas as passagens de texto que continham a classificação RES, tanto para Naive Bayes quanto para TF-IDF. Destas, foram escolhidas aleatoriamente 100 e lidas para identificar o porquê da discrepância. Foi possível identificar que as sequências “torna público o resultado” ou “torna pública a divulgação do resultado” aparecia em todas. Os termos “torna”, “públic” (sem a letra “o” ou “a”) e “resultado” foram utilizados para, de forma automática verificar os pares de classificação que não coincidiam. Um total de pouco mais que 20% foi constatado como pertencendo à classificação de divulgação de resultado (RES). Este procedimento aumentou o grau de coincidência entre as listas de palavras para a etapa classificada como RES para os dois métodos de classificação. As cinco listas de termos estão apresentadas no capítulo Resultados desta tese.

A metodologia adotada até este momento foi de grande importância para se entender a estrutura do financiamento atrelado às notícias publicadas no Diário Oficial da União. Entretanto o mapeamento do fluxo de financiamento, assim como a classificação das notícias referente a este fluxo e o vocabulário referente a cada etapa do fluxo não são suficientes para **modelar a extração de dados sobre financiamento para pesquisa em uma determinada doença a partir do DOU**, no caso deste estudo a **dengue**. Sendo assim, inicialmente, foi realizada uma busca contendo o termo “dengue” e outra contendo o termo “aedes” nas notícias classificadas como AWR o qual se refere à concessão dos recursos financeiros. Poucas foram as notícias recuperadas, o que indicou que a estratégia adotada não seria a melhor.

Sendo assim, foi utilizada outra importante fonte de dados do cenário da pesquisa com o intuito de auxiliar na identificação do volume de recursos financeiros alocados para pesquisa nesta temática, especificamente dengue. Isto foi necessário para se ter uma base de nomes de pesquisadores na temática estudada. Esta fonte de dados foi o Currículo Lattes pertencente à plataforma Lattes. A estratégia adotada foi a utilização da busca avançada deste sistema nacional de currículos. Para esta busca foi utilizada a seguinte estratégia:

- Somente doutores foram recuperados;
- No campo “esta frase exata” foi digitada o termo dengue.

A lista de pesquisadores recuperados foi salva em um arquivo nomeado **50_pesquisadores_dengue.txt** contendo o nome do pesquisador e o número identificador de 10 (dez) dígitos (ID10). O ID10 é um número utilizado quase que exclusivamente internamente pelo sistema. Para os usuários do mesmo existe uma identificação de 16 (dezesesseis) dígitos (ID16) que compõe o link de acesso ao currículo do pesquisador. Os dois identificadores – ID10 e ID16 – são únicos no sistema e são vinculados entre si. Existe apenas um pesquisador associado a um ID10 e ID16.

Os dois links abaixo, respectivamente com o ID10 e com o ID16 pertencem ao mesmo pesquisador:

- <http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4265233P5>
- <http://lattes.cnpq.br/9393004066976712>

O arquivo descrito acima, contendo o nome dos pesquisadores, serviu de base para a busca dos pesquisadores na Seção 3 do Diário Oficial da União com o intuito de identificar as datas e páginas nas quais seus nomes apareciam. Como resultado deste procedimento obteve-se um arquivo (**55_pesquisadores_dengue_no_dou.txt**) contendo o nome do pesquisador seguido do conjunto de cada data e respectiva(s) página(s) onde consta seu nome (ex.: Pesquisador A ; 01/10/2001,102,203 ; 21/3/2004,67). Este arquivo foi utilizado para identificar os pesquisadores que apareciam até **três** páginas de distância de uma notícia classificada como AWR (concessão de verba) ou PRO (extensão) pelo método *tf-idf* (**30_tfidf.txt**). O motivo de verificar os nomes mesmo que distante das páginas originárias se deu pelo fato de em diversos casos a notícia de fomento apresentar uma relação de contemplados que pode vir a mudar de página. Já o motivo para incluir a classificação PRO se baseou no fato que a renovação do projeto de pesquisa pode incluir uma repactuação de verba independente da prorrogação do prazo.

Tendo em mão esta lista de pesquisadores relacionados com uma notícia classificada como AWR ou PRO (**60_verifica_pesquisadores_awr_results.txt**), foram identificados valores monetários que constassem na passagem de texto onde aparecia o nome do pesquisador sendo esta passagem de texto próxima de outra classificada como AWR ou PRO.

A escolha por um método que identificasse o valor monetário não foi simples. Para este fim, trabalhou-se, inicialmente, no reconhecimento de entidades nomeadas, do inglês *Named Entity Recognition* (NER). Gu (2008) descreve o reconhecimento de entidades nomeadas como uma sub tarefa da Extração de Informação (Information Extraction – IE) e da mineração de textos (Text Mining – TM). O autor acrescenta que dada uma sentença, o objetivo na detecção de entidades nomeadas é normalmente encontrar nomes de pessoas, organizações, localização ou outras que façam sentido ao objetivo da pesquisa.

Named Entity Recognition é a forma de extrair informação na qual se busca classificar cada conjunto de palavras do documento como sendo o nome de uma pessoa, uma organização, uma localização, data, hora, valores monetários, outros tipos que venham a fazer sentido para pesquisa ou ainda nenhum. Esta tarefa tem bastante importância para buscadores na internet, tradução automática por máquina, indexação automatizada de documentos ou como base para uma extração complexa de informação.

Portanto, entidade é o nome dado para um conjunto de palavras sequenciais que se referem a um dado importante do texto. Para este caso foram definidas as seguintes entidades: Pessoa, organização, localização, data, valores monetários, saúde, identificação, evento e financiamento.

A seguir, Quadro 7, são apresentados os tipos de entidade, suas respectivas siglas, assim como exemplos de entidades referentes.

Quadro 7 – Tipos de entidades nomeadas

Tipo	Sigla	Do inglês	Exemplo
Pessoa	PER	PERSON	Nomes de pessoas
Organização	ORG	ORGANIZATION	Empresas públicas, privadas...
Data	DATE	DATE	Datas nos mais diversos formatos
Lugar	LOC	LOCATION	Cidade, Estado, País, o nome de um rio...
Valor	CUR	CURRENCY	Valores (R\$ 1,00, um real)
Saúde	MED	MEDICINE	Medicamentos, doenças...
Identificação	ID	IDENTIFICATION	Leis e Decretos
Evento	EVT	EVENT	Congresso, conferência
Financiamento	FUN	FUNDING	Termos de apoio, financiamento

Fonte: Elaboração própria

Para se chegar à identificação de entidades é necessário realizar a classificação sintática chamada de *Part of Speech Tagging*, ou POS-TAG. O POS-TAG é o nome dado a conjuntos abertos de palavras, definidos a partir de propriedades, funções semânticas ou gramaticais (BASILIO, 2011).

A fim de realizar a classificação POS-TAG foi utilizada a ferramenta desenvolvida pelo Departamento de Informática da Universidade de Lisboa chamada de LX-TAGGER (BRANCO; SILVA, 2004). A automatização desta foi realizada por meio de um *script* Python que lê toda a estrutura de arquivos no formato TXT, classifica cada um e salva a classificação em uma nova estrutura hierarquizada por ano e mês. Analogamente aos arquivos PDF e TXT descritos anteriormente, estes recebem o nome contendo ano com quatro dígitos seguido pelo mês e dia ambos com dois dígitos cada, *hífen*, o número da seção com um dígito, novamente um *hífen* e por fim o número da página com três dígitos. A extensão dada a cada arquivo classificado foi *tag*. O arquivo classificado, por exemplo, da página 50, da seção 3 do dia 24/11/2014 foi nomeado 20141124-3-050.tag.

A classificação POS TAG auxilia no reconhecimento de entidades nomeadas (NER). Para tanto, foi utilizada a biblioteca *Natural Language Tool Kit* (NLTK) existente para a linguagem Python. Entretanto, após alguns testes com a mesma, verificou-se que esta não estava preparada para o reconhecimento de entidades em textos que não estivessem no idioma inglês. Foi verificada a necessidade de treinar o sistema para reconhecer entidades a partir de textos em português.

Uma forma de treinar o sistema é utilizando a técnica de *Maximum Entropy* (maxent). Verificou-se que o treinamento é realizado a partir de um *script* chamado “named_entity.py” que se encontra no diretório “chunk” da biblioteca NLTK. Este arquivo foi copiado com o nome de “named_entity_pt.py” indicando a linguagem português (pt) no mesmo. Neste foram alterados os seguintes métodos da classe NEChunkParserTagger:

- `_classifier_builder`: a qual chamava o método `MaxentClassifier.train` passando como parâmetro o algoritmo MEGA Model Optimization Package (MEGAM) e passou a utilizar o Generalized Iterative Scaling (GIS). O motivo desta mudança foi a decisão da não instalação de mais um pacote externo ao servidor. O GIS já é incorporado no pacote do NLTK enquanto o MEGAM não.
- `_portuguese_wordlist`: Este método substitui o `_english_wordlist`. O novo método “ensina” ao sistema as palavras em português contidas nos corpus

“mac_morpho” e “machado”. O primeiro corresponde aos dados obtidos do jornal Folha de São Paulo e o segundo corresponde à obra completa de Machado de Assis. Os dois corpus já incorporados a NLTK.

Além disso, foi gerado um arquivo no formato CONLL 2002. Este formato foi descrito na *Conference on Natural Language Learning* do ano de 2002 para os idiomas Espanhol e Holandês.

Este arquivo recebeu o nome de iob_portuguese.train. O formato CONLL 2002 também é conhecido por *Inside, Outside, Begin* (IOB). Este formato possui, basicamente, três colunas separadas cada uma delas por um espaço. A primeira coluna é a palavra do texto, a segunda coluna a TAG de *Part of Speech* (POS) e a terceira a TAG IOB. A TAG IOB define se uma palavra faz parte ou não de uma entidade.

O texto abaixo retirado da página 37, seção 2, do DOU de 27/11/2014:

No - 1.095 - Autorizar o afastamento do país do servidor PAULO EDUARDO POTYGUARA COUTINHO MARQUES, Tecnologista em Saúde Pública do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, da Fundação Oswaldo Cruz, com a finalidade de realizar parte do curso de doutorado, intitulado: "Padrão de financiamento à pesquisa em dengue a partir do Diário Oficial da União", na Escola de Ciência da Computação da Universidade Simon Fraser, em Vancouver - Canadá, no período de 4 de fevereiro de 2015 a 2 de fevereiro de 2016, inclusive trânsito, com ônus para CAPES (Processo nº 25380.001631/2014-53)

Possui as seguintes entidades:

Quadro 8 – Entidades nomeadas

Tipo	Entidade
PER	PAULO EDUARDO POTYGUARA COUTINHO MARQUES
ORG	Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
ORG	Fundação Oswaldo Cruz
ORG	Escola de Ciência da Computação da Universidade Simon Fraser
ORG	Universidade Simon Fraser
LOC	Vancouver
LOC	Canadá
DATE	4 de fevereiro de 2015
DATE	2 de fevereiro de 2016
ORG	CAPES

Fonte: Elaboração própria

Como exemplo de anotação no arquivo IOB, segue abaixo parte em duas colunas por questão de organização do texto e onde as reticências identificam parte do texto omitido por não pertencerem a nenhuma entidade e para simplificação do exemplo:

(...)	de_ V I-ORG
o DA O	a DA I-ORG
servidor CN O	Computação CN I-ORG
PAULO PNM B-PER	de_ PREP O
EDUARDO PNM I-PER	a DA O
POTYGUARA PNM I-PER	Universidade PNM B-ORG
COUTINHO PNM I-PER	Simon PNM I-ORG
MARQUES PNM I-PER	Fraser PNM I-ORG
, PNT O	, PNT O
(...)	em PREP O
em_ PREP O	Vancouver PNM B-LOC
a DA O	- PNT O
Escola PNM B-ORG	Canadá PNM B-LOC
de PREP I-ORG	(...)
Ciência PNM I-ORG	

Para que este tipo (IOB) de anotação fosse realizado fora necessário um arquivo Padrão Ouro (*Gold Standard*) de onde a máquina tomara como base para anotar outro texto.

Exemplo:

Texto	POS-TAG	IOB
Paulo chutou a bola	Paulo/PNM chutou/V a/DA bola/CN	Paulo PNM B-PER chutou V B-ACAO a DA O bola CN B-OBJ

De forma simples, ao analisar a frase “Ana pegou a laranja”, o reconhecimento de entidades nomeadas (NER) identificaria automaticamente, segundo o exemplo acima que “Ana” é uma pessoa (PER), “pegou” é a ação (ACAO) e que “laranja” é o objeto (OBJ).

Para montar o arquivo *Gold Standard* para esta pesquisa foram recuperados inicialmente os nomes de 330 (trezentos e trinta) empresas públicas disponíveis no Sistema de Informações Organizacionais do Governo Federal (SIORG) por meio de um arquivo no formato *JavaScript Object Notation* (JSON). O formato JSON é amplamente utilizado para interoperabilidade entre sistemas, isto significa que o padrão JSON possibilita facilmente a leitura e entendimento por outros sistemas.

A partir dos 330 nomes foram recuperados aleatoriamente 324 parágrafos do Diário Oficial da União que possuíam aqueles nomes. Os seis casos de perda podem ser caracterizados como erro de digitação no SIORG e/ou no DOU. Como exemplo é

possível destacar o “Tribunal **Reginal** Eleitoral da Bahia” donde na palavra Regional falta a letra “o”.

A partir dos parágrafos recuperados, foi realizada a classificação POS Tagging utilizando o LX-Tagger conforme descrito anteriormente. Tomando como base esta classificação, foi preparado um arquivo no formato CONLL 2002 (IOB), tendo cada palavra do novo arquivo recebido a classificação *Outside* (O). Este arquivo foi revisto manualmente de sorte a classificar as palavras conforme *Begin* e *Inside* entidades.

Após todo o arquivo classificado e revisto, utilizou-se um *script* em Python para ler este arquivo (**pt_portuguese.train**), treinar o sistema com estas classificações e gerar um novo arquivo que recebera o nome de **pt_chunker.pickle**. Ressalta-se que arquivos do tipo *pickle* são a representação serializada do objeto salvo. O processo de serialização (*pickling*) consiste na utilização de uma máquina virtual baseada em pilha simples que registra as instruções utilizadas para reconstruir o objeto. Em outras palavras, a partir de um arquivo no formato *pickle* não é necessário retreinar o sistema toda vez que for necessário reconhecer entidades, pois o sistema carrega o treinamento já realizado a partir deste arquivo no formato *pickle*.

Em posse do treinamento, se deu início ao reconhecimento dos valores monetários pertencentes à passagem de texto onde o pesquisador aparecia próximo a uma notícia AWR ou PRO. Este procedimento acabou sendo abandonado, pois o Reconhecimento de Entidades Nomeadas recuperava todos os valores da passagem de texto, o que muitas vezes, por conta de problemas decorrentes à conversão de PDF para TXT, acarretava na recuperação de valores que não condiziam com o financiamento para aquele pesquisador.

Tendo em vista o problema encontrado buscou-se então outra forma de identificar os valores monetários. Foi adotada a busca por expressão regular. A expressão regular é um padrão que especifica um conjunto de cadeias de caracteres. Ou seja, expressão regular auxilia a recuperação de todas as partes de texto que corresponda a uma regra pré-determinada.

Foi desenvolvido um script em Python que buscou o **primeiro** valor monetário **após** o nome do pesquisador para cada uma das passagens de texto. Porém, caso não fosse identificado nenhum valor, buscar-se-ia o **último** valor **antes** do nome. Para os casos em que não fosse detectado nenhum valor monetário, aquela passagem de texto era descartada.

Este procedimento gerou um arquivo contendo o código da agência de fomento, a data de publicação do fomento, o nome do pesquisador e o valor identificado pelo procedimento descrito acima (**75_new_confronta_awr_pesquisadores.txt**).

Este arquivo serviu de base para a descrição do padrão de financiamento para pesquisa em dengue no Brasil. A pergunta “o que é produzido a partir deste recurso?”, mesmo em vista do descrito de que ciência não é uma linha de montagem, é latente. De modo a **explorar a relação entre financiamento e produção científica** foi lançada mão de outra fonte de dados, o Google Acadêmico (Google Scholar). Foi realizado um levantamento da produção de todos os pesquisadores identificados no Currículo Lattes. Buscou-se então no Google Acadêmico (Google Scholar) a produção dos referidos pesquisadores utilizando as regras abaixo:

- Filtro de data: de 2005 e 2015 inclusive;
- Filtro “com a frase exata” utilizando o termo “dengue”
- Não foram recuperados patentes ou citações.

Este levantamento gerou um arquivo (**50_producao_scholar_2005_2015.txt**) contendo em cada linha o nome do pesquisador, o ano e a quantidade de produção científica para aquele ano.

Ainda nesta linha, buscando atividades científicas dos pesquisadores, foi utilizada a ferramenta ScriptLattes. Ao baixar os currículos dos pesquisadores, o ScriptLattes gera um arquivo contendo, de forma estruturada em um arquivo tipo XML, os dados existentes no currículo de cada pesquisador, além de dados processados pela própria ferramenta. Este arquivo foi acessado de modo a recuperar variáveis de análise para este estudo. Sendo assim, foram recuperados:

- a rede de coautoria de cada um dos pesquisadores para cada ano;
- a instituição à qual o pesquisador está vinculado;
- dados geográficos: cidade, estado.

O resultado referente ao levantamento da rede de coautores foi salvo em um arquivo contendo as coautorias formadas a cada ano. Os campos salvos neste arquivo foram ano da coautoria, código identificador de 16 (dezesesseis) posições do pesquisador, código identificador de 10 (dez) posições do pesquisador e código identificador de 16 posições do coautor do primeiro pesquisador.

O segundo item merece um destaque especial, uma vez que não há campo no Currículo Lattes para se colocar afiliação. Sendo assim, foi utilizado o campo de

endereço profissional com a finalidade de recuperar o nome da instituição. Já para os dados geográficos, foi utilizado o mesmo campo citado.

Todo desenvolvimento metodológico até este momento possibilitou a coleta dos dados referente ao financiamento para pesquisa em dengue no Brasil. Entretanto, para possibilitar a análise destes dados, foi necessário criar uma base de dados onde os mesmos estivessem relacionados entre si. O sistema gerenciador de banco de dados (SGBD) escolhido foi o PostgreSQL, ou Postgres como é mais conhecido. O Postgres é um SGBD de código aberto e gratuito. Ele funciona no sistema operacional Linux que fora escolhido como base para o desenvolvimento deste método e pode atender tanto a pequenos como gigantescos bancos de dados. O modelo de relacionamento da base de dados está descrito na seção de resultados.

A partir da base de dados gerada pelo processo descrito acima, foi possível iniciar a análise dos dados com objetivo de **identificar o padrão de financiamento para pesquisa em dengue no Brasil entre os anos 2005 e 2014**. As perguntas abaixo foram norteadoras para esta análise:

- Como se dá a distribuição do fomento à pesquisa em dengue no Brasil pelas categorias de análise: Estado; ano; instituição; pesquisador e rede de coautoria?
- Existe alguma relação entre o financiamento e a produção científica ao longo do tempo a partir dos dados do Diário Oficial da União?
- Como se apresenta a relação do financiamento encontrado neste estudo com os dados existentes das diversas fontes oficiais internacionais e nacionais – G-FINDER, CNPq, CAPES, FINEP e DECIT ?

De sorte a responder as duas primeiras perguntas optou-se pela utilização da análise visual (*visual analytics*) dos dados. Conforme já visto anteriormente, a visualização dos dados quando, em uma forma mais amigável que tabela, possibilita uma melhor análise do cenário estudado. Sendo assim, com o acesso a base de dados, foram geradas diversas visualizações distintas com o intuito de se explorar o financiamento para pesquisa em dengue no Brasil e possibilitar a análise dos mesmos.

Todas as visualizações geradas podem ser encontradas no link <http://157.86.8.9:8080/dou> ou <http://ctic009.icict.fiocruz.br:8000/dou> (Sistema FarejaDOU¹⁶) e clicando na opção **Análise** do menu principal.

¹⁶ Até o momento não foi possível disponibilizar uma URL definitiva para o Farejadou

Com a finalidade de explorar a relação entre financiamento e produção científica foi traçado um gráfico de linhas contendo montante de financiamento para aqueles pesquisadores ao longo do tempo assim como sua produção.

O procedimento adotado para comparar o resultado encontrado com os dados disponibilizados pelas instituições de financiamento – CNPq, CAPES, FINEP e DECIT – assim como pelo G-FINDER foi o mesmo. A página web que cada uma destas entidades disponibiliza com os dados de financiamento foi acessada e os dados foram baixados e tabulados de acordo com o nível de detalhamento que cada uma destas organizações apresenta.

Para o CNPq, a página acessada foi por meio do portal da instituição. Mais especificamente na opção do menu “Dados Abertos” dentro de “Acesso à Informação”. Foi possível baixar um arquivo compactado contendo arquivos no formato CSV datados de 2006 até 2015 com dados sobre o financiamento aos Projetos de Pesquisa. Tais arquivos possuem como principais campos para fins da análise: ano, estado, região, grande área, área e total geral, sendo este último o valor agregado de bolsas e orçamento para capital e consumo (OCC). O campo grande área foi utilizado para filtrar os financiamentos para Ciências Biológicas (pelo vetor), Ciências Humanas e Ciências Sociais, pela educação em saúde, sendo que estas duas últimas são quase esquecidas, ainda que apontadas como vital. Já pela área do conhecimento, a Saúde Coletiva, por conta da epidemiologia, foi utilizada como filtro.

Tanto a CAPES¹⁷ quanto a FINEP¹⁸ redirecionam a busca para o portal da transparência. Neste é possível recuperar o montante total executado por **natureza de despesa** ou **programa de governo** para um determinado ano escolhido. A CAPES possui dados de 2005 até o corrente ano. Em **natureza de despesas** a CAPES apresenta em duas linhas distintas valores para AUXILIO FINANCEIRO A PESQUISADORES, sendo uma com o código 33902000 e outra com o código 44902000.

A formação deste código, Figura 14, define onde o recurso foi alocado. O primeiro dígito apresenta a categoria econômica da despesa, sendo o número 3 para **despesas correntes** e o número 4 para **despesas de capital**. O segundo representa o grupo de natureza de despesa, sendo 3 para **outras despesas correntes** e 4 para **investimentos**. Interessante verificar que compra de equipamentos (despesas de capital) é considerado investimento, enquanto o financiamento da pesquisa propriamente dito é

¹⁷ <http://www3.transparencia.gov.br/jsp/execucao/execucaoTexto.jsf?consulta=1&consulta2=0&CodigoOrgao=26291>

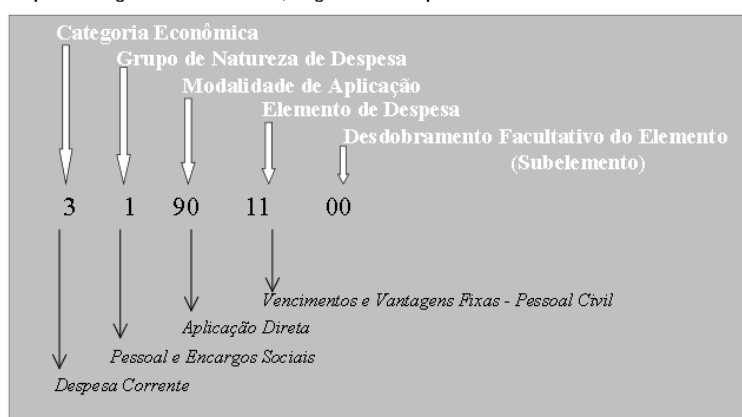
¹⁸ <http://www3.transparencia.gov.br/jsp/execucao/execucaoTexto.jsf?consulta=1&consulta2=0&CodigoOrgao=20502>

considerado “outras despesas correntes”. Ainda na composição do código, os dois dígitos seguintes representam a **modalidade de aplicação** a qual “indica se os recursos serão aplicados mediante transferência financeira [...] ou, então, diretamente pela unidade detentora do crédito orçamentário, ou por outro órgão ou entidade no âmbito do mesmo nível de Governo”. (BRASIL, 2016, p.61) Para estes casos o número 90 representa **aplicações diretas**. Os dois números seguintes, 20 (vinte) para aqueles códigos, apresenta que o **elemento de despesa** é justamente o **Auxílio Financeiro a Pesquisadores**.

Figura 14 – Formação do código de Natureza de Despesa e exemplo.

1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª
Categoria Econômica	Grupo de Natureza da Despesa	Modalidade de Aplicação	Elemento de Despesa	Subelemento			

Exemplo: código “3.1.90.11.00”, segundo o esquema abaixo:



Fonte: BRASIL, 2016

Para fins de análise, foram recuperados os valores referentes aos dois códigos para todos os anos de 2005 a 2014 inclusive. Estes valores foram agrupados por ano em uma tabela com o intuito de servir de base para a comparação da alocação de recursos ao longo do tempo com o que foi encontrado na mineração do Diário Oficial da União.

A CAPES ainda possui a ferramenta geocapes da qual é possível baixar a base de dados de concessão de bolsas desde o ano de 1995 até 2014, em um único arquivo no formato Excel. Embora seja uma base interessante, não se relaciona diretamente com este estudo.

Não foi possível recuperar os dados para a FINEP, uma vez que o Portal da Transparência apresenta apenas uma linha de código para cada ano verificado – de 2005 a 2014 inclusive. Código estes que não correspondem ao auxílio financeiro a pesquisadores.

Para o DECIT¹⁹, utilizou-se a busca avançada do site **Pesquisa Saúde** filtrando o resultado da busca pelos anos de 2005 a 2014 inclusive. Em seguida, foi solicitada a exportação dos resultados em formato de planilha Excel© contendo todos os campos disponíveis. Um destes campos é o link para o CV Lattes do pesquisador. Este link contém o código de 10 dígitos referente à identificação do pesquisador na plataforma de currículos. Este campo serviu de filtro para verificar se o pesquisador faz parte do grupo de pesquisadores identificados como pesquisadores em dengue. Sendo assim, os demais pesquisadores foram excluídos da tabela. A tabela ainda contém os campos de ano, unidade federativa, instituição, o parceiro administrativo, os valores de responsabilidade do DECIT e do parceiro administrativo entre outros. Esta tabela foi utilizada para a validação dos dados recuperados do DOU.

O capítulo seguinte apresenta os resultados encontrados em um formato descritivo.

¹⁹ <http://pesquisasaude.saude.gov.br/index.xhtml>

8. Resultados

No presente estudo os resultados são de duas ordens: um metodológico e outro do padrão de financiamento propriamente dito. Este estudo apresenta-se inovador, não só pelo objetivo de identificar o padrão de financiamento em nível federal para pesquisas em dengue, mas também pela metodologia desenvolvida para tal. Sendo assim as seções deste capítulo apresentarão respectivamente (i) as ferramentas tecnológicas decorrentes do estudo e (ii) os resultados fruto do desenvolvimento metodológico referentes ao conjunto de dados levantados e quanto ao financiamento para pesquisa em dengue encontrados.

8.1. Ferramentas tecnológicas

8.1.1. CP2T: convertendo de PDF para TXT

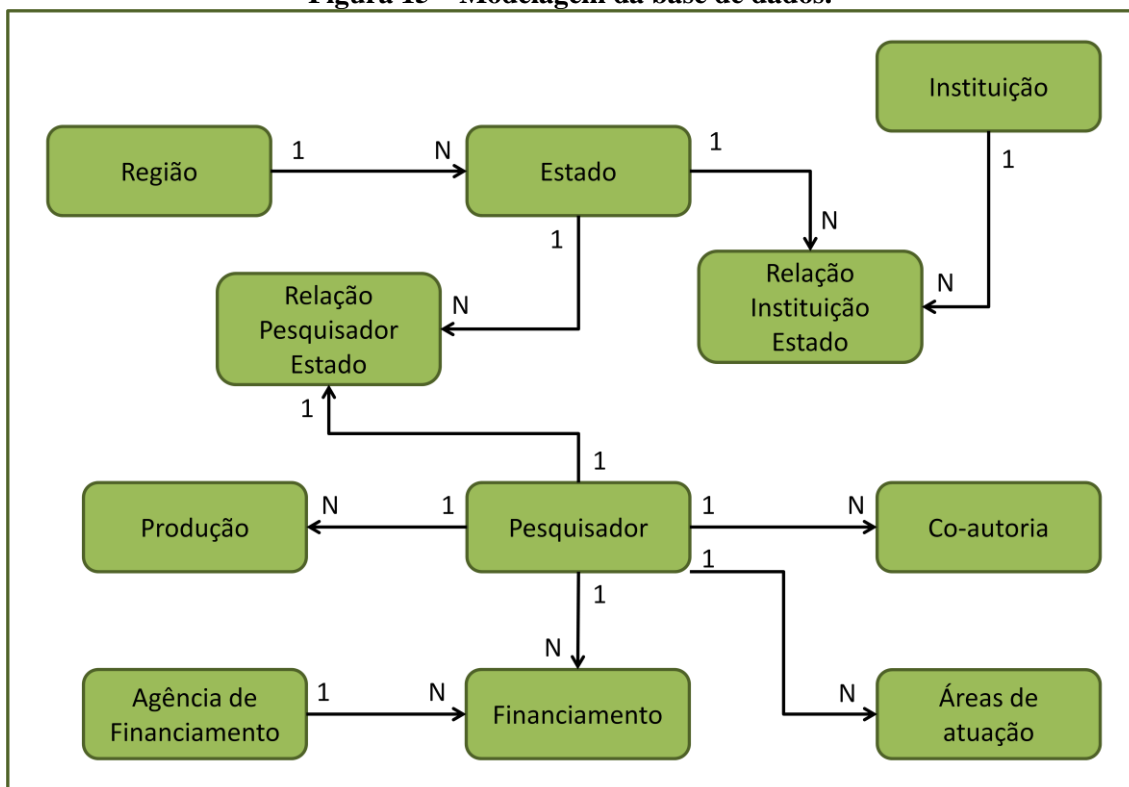
O primeiro resultado deste estudo está intrinsecamente vinculado ao método se confundindo com o mesmo. O processo de *download*, conversão e indexação do Diário Oficial da União (DOU) tem sua maior importância na etapa de conversão do formato *Portable Document Format* (PDF) para o formato texto (TXT) mantendo as tabelas existentes no DOU, conforme descrito no capítulo anterior. Com esta finalidade, foi desenvolvida uma ferramenta chamada *Convert PDF to Text* (CP2T) a qual disponibilizada no GitHub, portal de acesso mundial aberto para ferramentas desenvolvidas em diversas linguagens. O acesso à ferramenta se dá pelo link <http://github.com/pepcmarques/cp2t>. Tal ferramenta também possibilita a conversão de PDF para o formato *HyperText Markup Language* (HTML) e está livre para receber contribuições.

8.1.2. A base de dados

A análise se apoia em duas bases de dados. A primeira, sob a forma de arquivos de índice criados com a biblioteca Whoosh em Python com a finalidade de possibilitar a busca no D.O.U. e a segunda criada a partir da recuperação, identificação e consolidação dos dados do Currículo Lattes, a qual inclui o nome dos pesquisadores identificados e respectivos números de identificação de 10 e 16 dígitos, assim como sua filiação; Google Acadêmico contendo a quantidade de artigos por ano de um determinado pesquisador; e do Diário Oficial da União contendo o ano do financiamento, quem foi financiado, a agência financiadora e o valor financiado. A base

de dados foi criada em PostgreSQL com o conceito de relacionamento entre as tabelas conforme apresentado na figura a seguir.

Figura 15 – Modelagem da base de dados.



Fonte: Elaboração própria

Esta base de dados pode ser utilizada, por exemplo, para identificar o quanto foi financiado para as diferentes regiões do Brasil em um determinado período de tempo. É possível saber os valores financiados por região demográfica entre os anos de 2005 e 2009 executando o comando abaixo na interface do PostgreSQL no servidor onde se encontra a base de dados:

```
select e.id_regiao, cast(sum(aw.valor) as decimal(20,2)) from award as aw INNER JOIN pesquisador as p ON aw.id16 = p.id16 INNER JOIN rpesqest as r ON p.idx = r.id_pesquisador INNER JOIN estado as e ON r.id_estado = e.idx where aw.ano >= 2005 and aw.ano <= 2009 group by e.id_regiao order by e.id_regiao ;
```

Tal comando só pode ser executado por alguém com acesso ao servidor e por isso foi criada uma página web com várias possibilidades de busca e visualizações já disponibilizada, o FarejaDOU.

Cada uma das tabelas de dados representada na figura anterior, Figura 15, possui campos pertinentes às mesmas conforme descrito no Anexo 2.

8.1.3. FarejaDOU: A ferramenta de busca do DOU e visualização

Além da ferramenta descrita acima, para a busca e análise dos dados coletados foi construída uma ferramenta *web* utilizando um framework Python (Web2py). O portal desenvolvido encontra-se disponível na URLs <http://157.86.8.9:8000/dou> e <http://ctic009.icict.fiocruz.br:8000/dou>. A esta ferramenta deu-se o nome de FarejaDOU. Na tela principal do sistema, Figura 16, encontram-se um menu com as opções (a) Principal, (b) Busca, (c) Visualizações, (d) Produção e (e) Sobre. Ainda é possível encontrar a imagem de uma lupa que corresponde à opção Busca do menu e uma imagem de peças tridimensionais de quebra-cabeças que corresponde à opção Visualizações do menu.

Figura 16 – Tela principal do portal.



Fonte: Coutinho-Marques (2014)

Ao clicar na opção de Busca ou na Lupa o sistema abrirá uma página, Figura 17 a seguir, contendo campos para recuperação de páginas do DOU contendo o(s) termo(s) e opções de busca utilizadas.

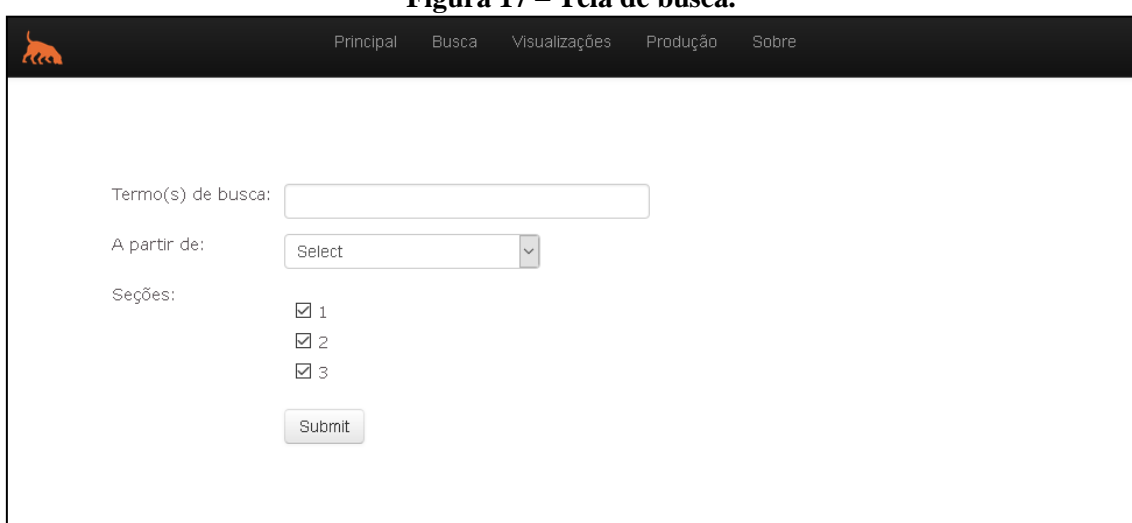
O FarejaDOU possui recursos como busca de termos, termo exato e, entre outros, proximidade de palavras. Ainda é possível filtrar pelo ano de início da busca e Seção do Diário Oficial da União – Seção 1, Seção 2 ou Seção 3.

Para a busca de termos, basta digitá-los separando-os por espaço (ex.: financiamento dengue). Para a busca do termo exato é necessário colocá-lo entre aspas (ex.: “financiamento para dengue”). A busca por proximidade se utiliza de um caractere especial que indica ao motor de busca que se buscará palavras próximas. Para tanto é necessário colocar o termo entre aspas seguido de til (~) e o número de palavras que se pode encontrar entre os termos (ex.: “termo concessão apoio”~2). Para este exemplo, a distância entre as palavras contidas entre aspas pode ser de até duas palavras, ou seja,

pode existir uma palavra entre elas, como por exemplo, poderíamos encontrar o resultado: “termo de concessão e apoio”, ou ainda “termo para concessão ao apoio”.

A escolha do ano inicial de busca se dá por meio de um *menu tipo drop down* que se inicia em 1990 e finda em 2017 (corrente ano). Já a escolha da Seção do DOU é feita por meio de *checkboxes*. Conforme apresentado no capítulo Um Novo Olhar Sobre a Política Científica, Seção 1 traz leis, decretos, resoluções, instruções normativas, portarias assim como outros atos normativos de interesse geral; a Seção 2 apresenta a vida da Administração Pública Federal e a Seção 3 possui contratos, editais, distribuição de recursos financeiros.

Figura 17 – Tela de busca.



The image shows a search interface with a dark header containing a logo and navigation links: Principal, Busca, Visualizações, Produção, and Sobre. The main content area is white and contains the following elements:

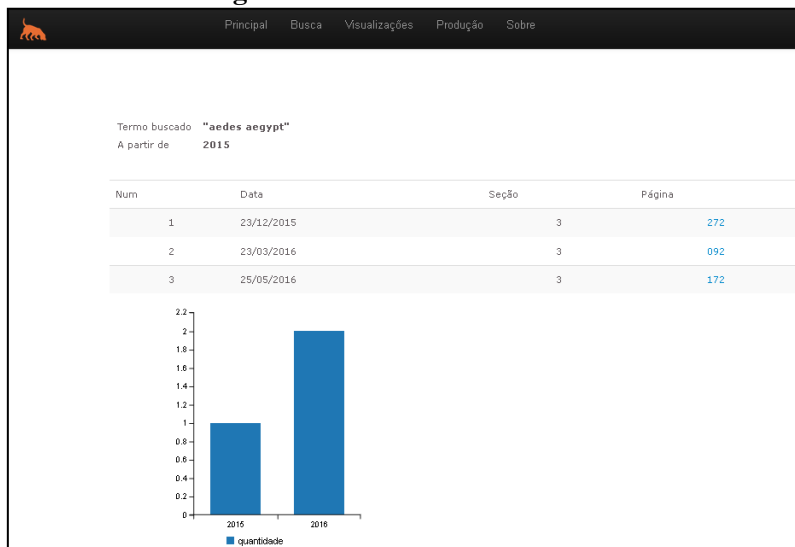
- A text input field labeled "Termo(s) de busca:".
- A dropdown menu labeled "A partir de:" with the text "Select" and a downward arrow.
- A section labeled "Seções:" with three checkboxes, each followed by a number: 1, 2, and 3.
- A "Submit" button.

Fonte: Coutinho-Marques (2014)

O resultado da busca é apresentado em uma tabela que contém um número sequencial, a data em que o termo foi encontrado, a respectiva Seção e o número da página. Esta última possui um link direto para a página em PDF no site da Imprensa Nacional.

A tela de resultados, Figura 18, possui ainda um gráfico de barras o qual apresenta a quantidade de vezes que o termo buscado foi encontrado por ano.

Figura 18 – Resultado da busca.



Fonte: Coutinho-Marques (2014)

O FarejaDOU também apresenta a visualização de gráficos e grafos dinâmicos que podem ser acessados clicando na opção **Análise** do menu ou na imagem de peças tridimensionais de quebra cabeças na página principal do sistema. Estas visualizações têm a finalidade de, além de auxiliar na análise dos dados deste estudo, servir como uma ferramenta que possibilite o *Visual Analytics* com os dados sobre financiamento para pesquisa em dengue no Brasil. O *menu* de visualizações, Figura 19 a seguir, possui cinco opções, a saber: (i) agência x ano, (ii) pesquisadores x ano, (iii) financiamento e produção (ano, agência) - distribuição geográfica, (iv) financiamento por rede de coautoria (ano e ranking) e (v) parceria entre as instituições (ano).

Figura 19 – Menu de Visualizações.



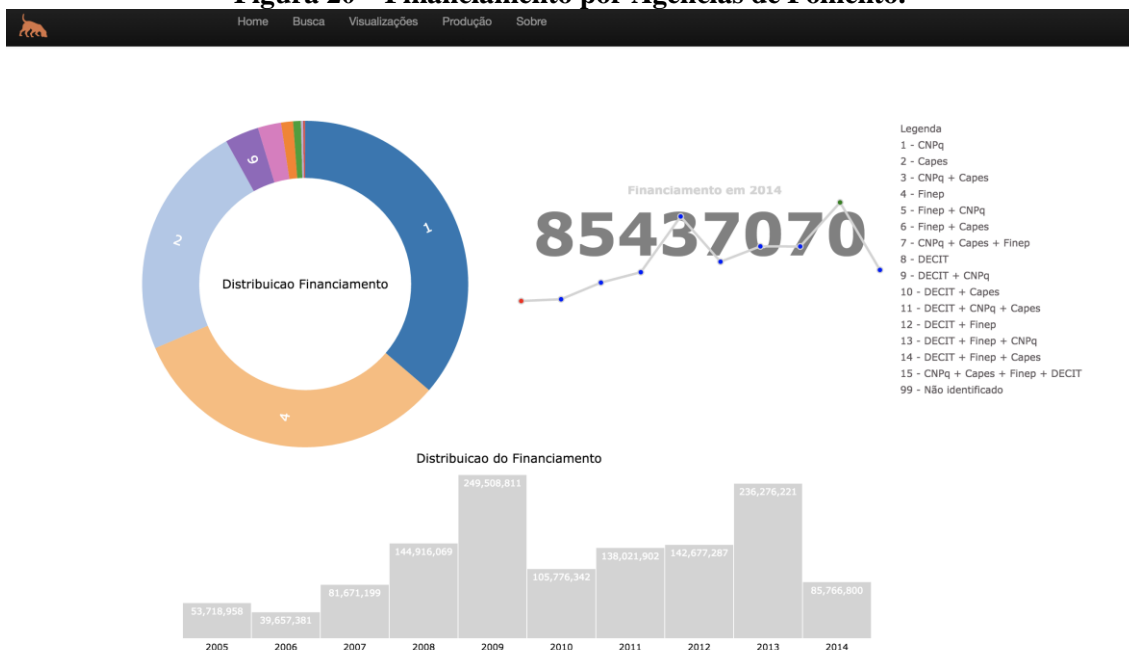
Visão: agência x ano
Visão: Pesquisadores x ano
Visão: Financiamento e produção (ano, agência) - distribuição geográfica
Visão: Financiamento por rede de coautoria (ano e ranking)
Visão: Parceria entre as instituições (ano)

Fonte: Coutinho-Marques (2014)

A visão do financiamento por agências de fomento apresenta inicialmente um grafo no formato de um *donut* contendo a distribuição do financiamento por agências de fomento, o valor total de financiamento no ano de 2014 (último ano que esta pesquisa

compreende), por cima deste colar um gráfico de linha com o financiamento ao longo dos dez anos estudados, sendo cada ponto a representação de um ano e o menor valor em vermelho, o maior valor em verde e os demais em azul. Abaixo é apresentado um gráfico de barras com os valores financiados por ano, Figura 20.

Figura 20 – Financiamento por Agências de Fomento.

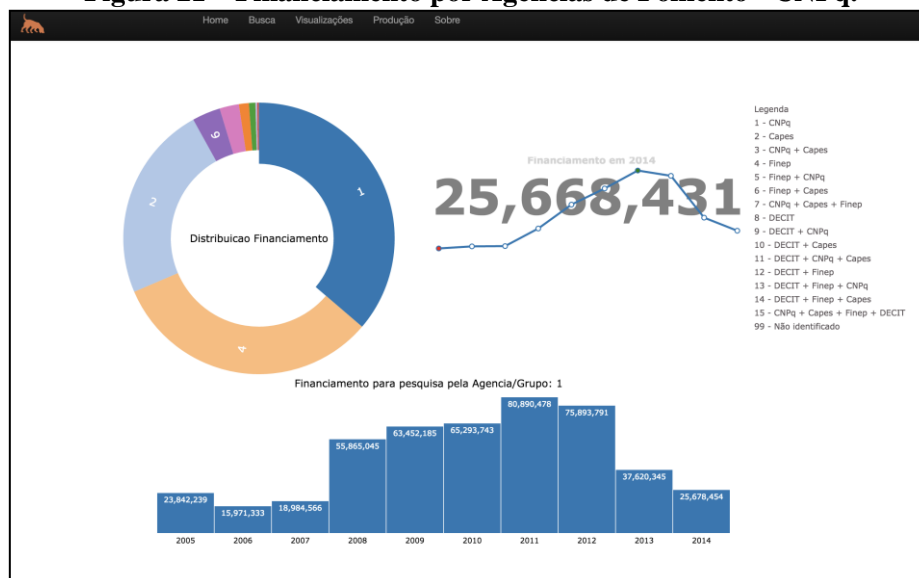


Fonte: Coutinho-Marques (2014)

Ao posicionar o ponteiro do mouse sobre um dos setores do grafo em forma de *donut* o mesmo irá aumentar de tamanho, ao clicar sobre o setor selecionado, ou seja, a agência financiadora (ou conjunto de agências financiadoras), os demais gráficos irão apresentar os dados daquele selecionado.

Em outras palavras, o valor apresentado para o ano de 2014 será o da agência (ou conjunto) selecionado, assim como o gráfico de linha e o de barras apresentarão os dados anuais daquele selecionado, Figura 21 a seguir.

Figura 21 – Financiamento por Agências de Fomento - CNPq.



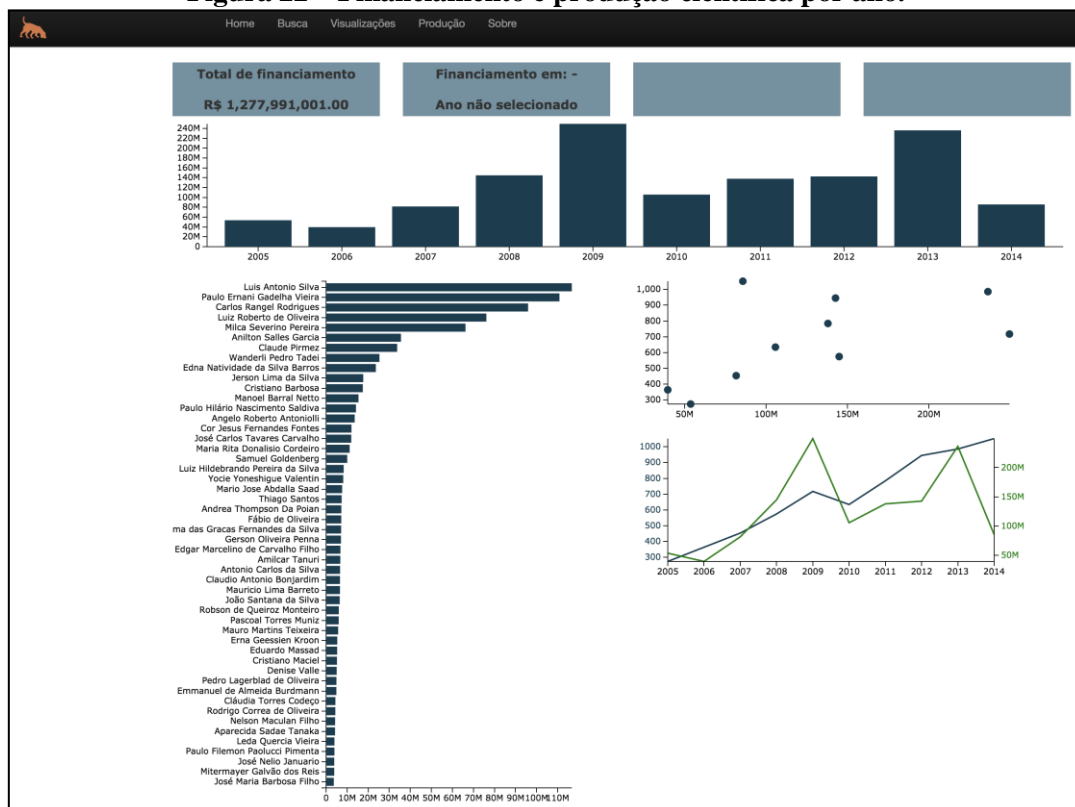
Fonte: Coutinho-Marques (2014)

Outra visualização desenvolvida para fins deste estudo também foi baseada no conceito de *dashboard* onde vários indicadores podem ser vistos em um mesmo ambiente. Esta visualização teve o intuito de identificar a relação do financiamento e da produção científica.

Inicialmente esta visualização apresenta os dados agregados para todos os anos. Um gráfico de barras no topo do *dashboard* apresenta a distribuição do financiamento para pesquisa em dengue ao longo dos dez anos levantados por este estudo. O valor total encontra-se no quadro superior esquerdo. O quadro imediatamente ao seu lado não apresenta valores neste início, somente quando se escolhe um dos anos. Um gráfico de barras na vertical lista os 50 pesquisadores em ordem decrescente de valores recebidos ao longo dos 10 anos. Ainda neste *dashboard* pode ser encontrado um gráfico tipo *scatter plot* onde cada ponto representa um ano do estudo com a relação financiamento pelo número de artigos identificados no Google Acadêmico. O eixo X deste gráfico representa o valor financiado, enquanto o eixo Y representa o número de artigos. Em outras palavras, quanto mais à direita e para cima estiver o “ponto azul” maior o volume de recursos recebidos assim como maior a quantidade de artigos publicados naquele ano. Posicionando o ponteiro do mouse sobre um dos “pontos azuis”, aparecerá uma pequena janela contendo o ano, o valor financiado naquele ano e a quantidade de artigos.

Também é apresentado neste *dashboard* um gráfico de linhas com dois eixos: um para quantidade de produção científica identificada e outro para valores de financiamento, Figura 22 a seguir.

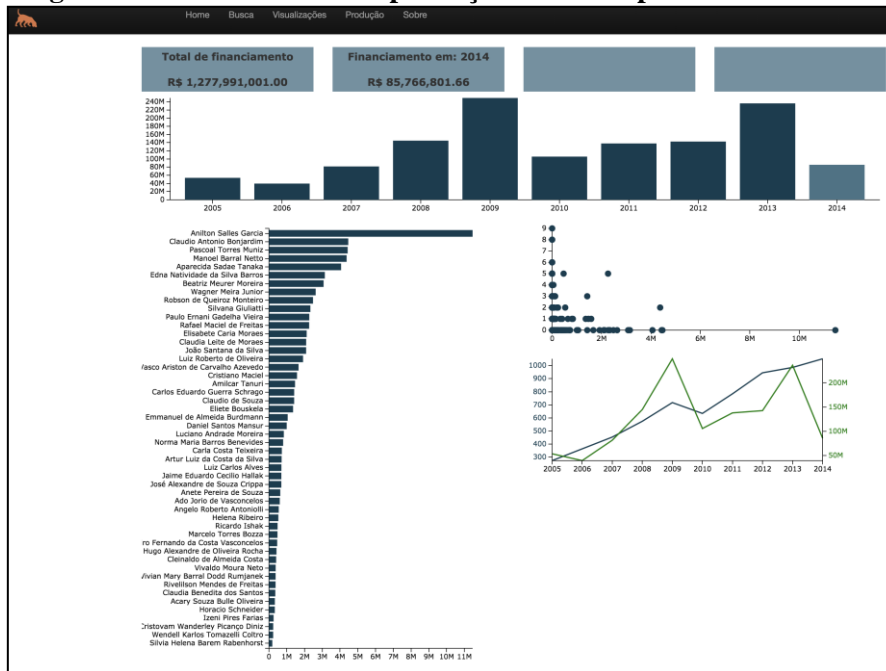
Figura 22 – Financiamento e produção científica por ano.



Fonte: Coutinho-Marques (2014)

Ao posicionar o ponteiro do mouse sobre uma das barras do primeiro gráfico (anos), a mesma mudará de cor indicando a seleção. Ao clicar na barra selecionada os gráficos acompanharão a seleção. Em outras palavras, o gráfico de barras verticais apresentará a lista dos 50 pesquisadores que mais receberam financiamento naquele ano em ordem decrescente de valores. O gráfico tipo *scatter plot* apresentará 50 bolas representando cada pesquisador. Neste momento, posicionando o ponteiro do mouse sobre uma das bolas aparecerá o nome do pesquisador, quanto ele recebeu naquele ano de financiamento e a quantidade de artigos que foram identificados no Google Acadêmico referente ao pesquisador, Figura 23 a seguir. Ressalta-se aqui que, neste momento, o gráfico de linhas continua mostrando o todo e que o segundo quadro superior passa a mostrar o montante de financiamento para aquele ano.

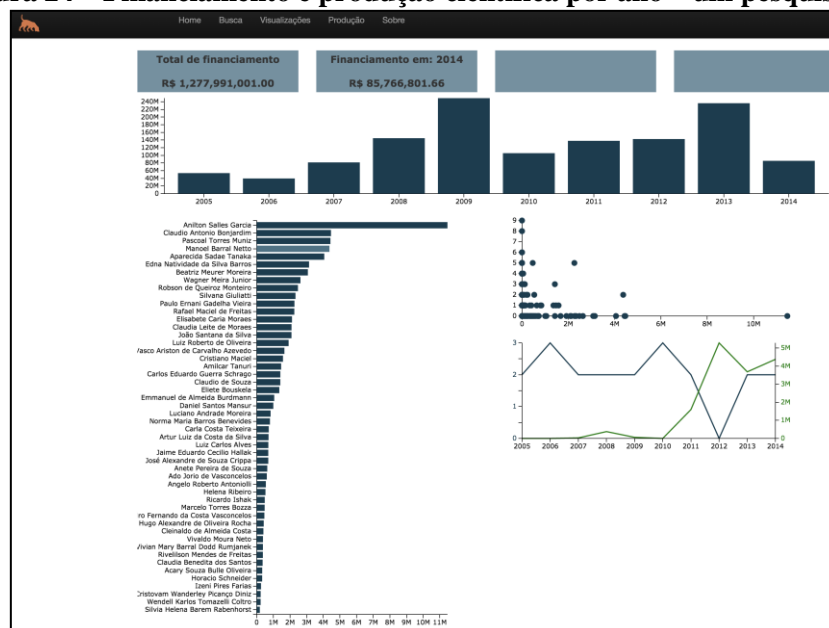
Figura 23 – Financiamento e produção científica por ano – ano 2014.



Fonte: Coutinho-Marques (2014)

O gráfico de barras verticais contendo os 50 pesquisadores que mais receberam financiamento também permite interação. Posicionando o ponteiro do mouse sobre a barra de um dos pesquisadores, a mesma mudará de cor indicando a seleção do pesquisador. Porém, é necessário clicar na mesma para que a interação com o gráfico de linhas ocorra. O gráfico de linhas apresentará o financiamento recebido pelo pesquisador selecionado assim como a quantidade de artigos de sua autoria no tema, indexados pelo Google Acadêmico, Figura 24 a seguir.

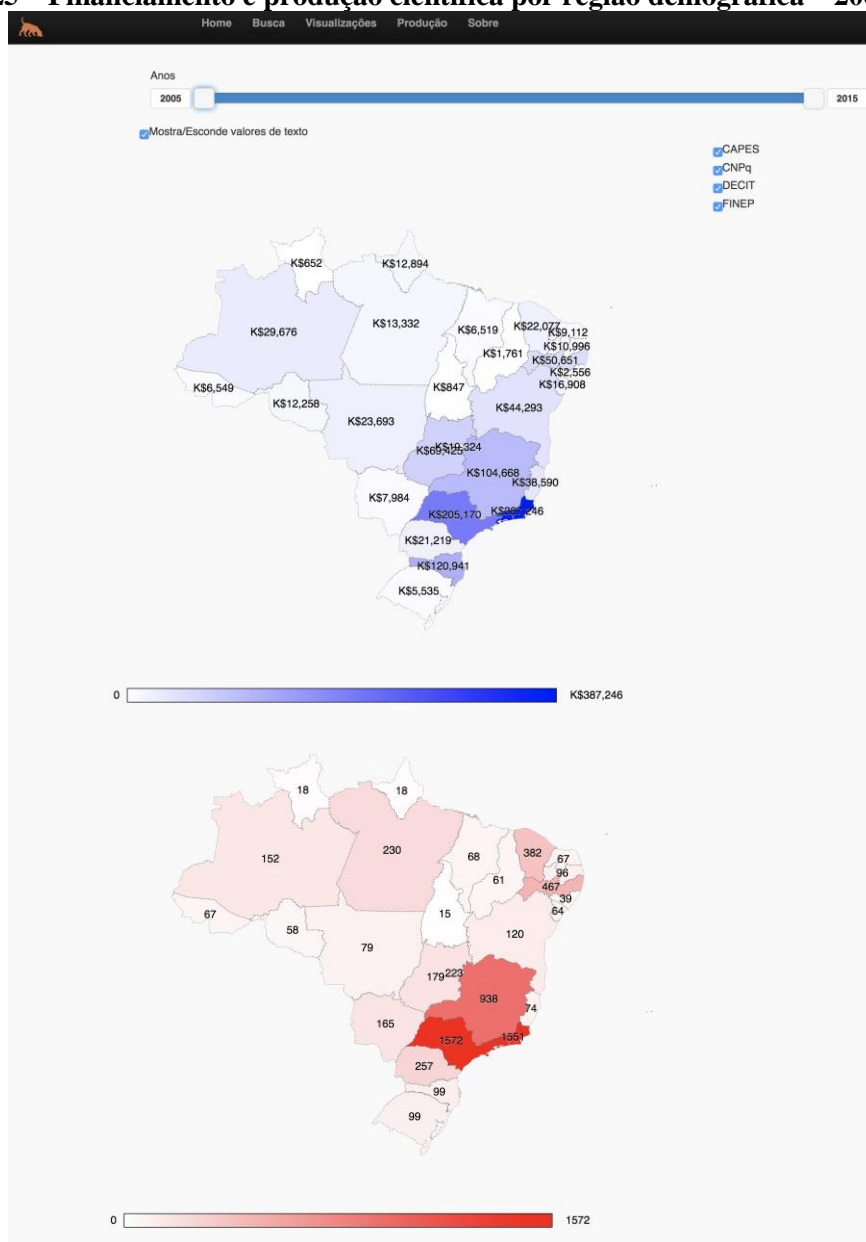
Figura 24 – Financiamento e produção científica por ano – um pesquisador.



Fonte: Coutinho-Marques (2014)

Com o objetivo de responder a pergunta de como se dá a distribuição geográfica do financiamento para a pesquisa em dengue assim como a produção científica no país, foi criada uma visualização contendo dois mapas do Brasil sendo que o primeiro apresenta o financiamento (em azul) e o segundo apresenta a produção científica. Ambos os gráficos são atualizados automaticamente quando são selecionados os anos inicial e final na barra superior (*sliders*). O gráfico contendo o financiamento também é atualizado quando se escolhe as agências de financiamento que aparecem à direita. Um outro *checkbox* à esquerda oculta ou exhibe os valores numéricos nos dois mapas, Figura 25 a seguir.

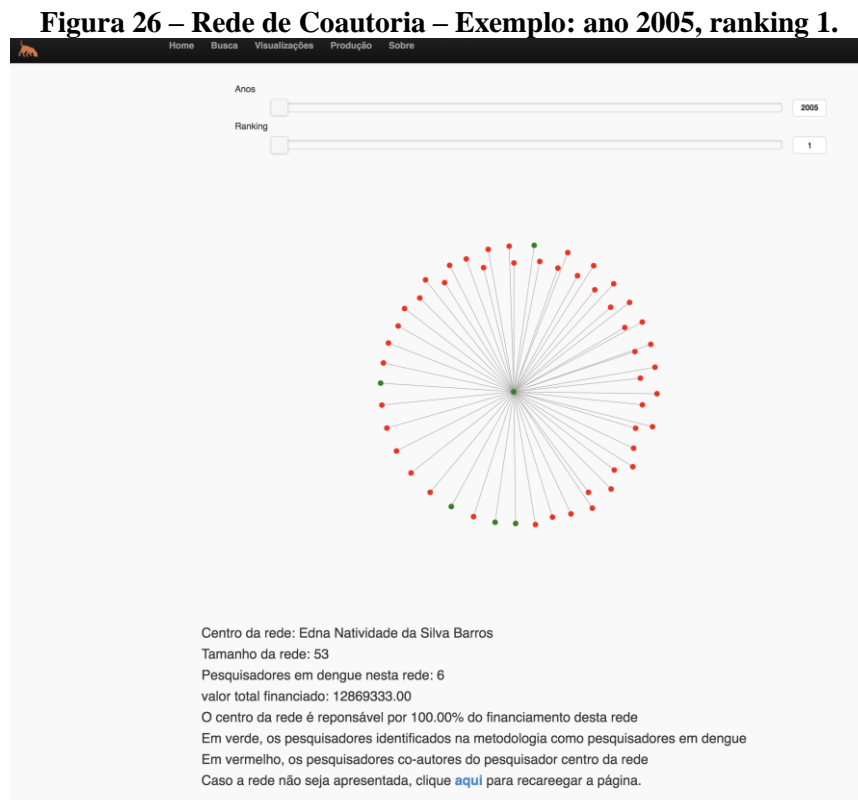
Figura 25 – Financiamento e produção científica por região demográfica – 2005 a 2014



Fonte: Coutinho-Marques (2014)

Um grafo de redes foi montado com o intuito de verificar quais redes de coautoria receberam um maior financiamento ao longo dos anos. Este grafo possui dois seletores que o atualizam automaticamente: um seletor de ano e um seletor de posicionamento no *ranking*. Este posicionamento é dado em ordem decrescente de valores recebidos por toda a rede, ou seja, somando os valores recebidos por cada nó da rede. Ressalta-se que a rede foi gerada a partir de dados coletados pelo ScriptLattes e não se limitou aos pesquisadores identificados na metodologia deste estudo. Sendo assim, para fins de melhor visualização, foram atribuídas cores aos nós da rede (pesquisadores) que receberam a cor verde por pertencer ao conjunto de pesquisadores inicialmente identificados e, vermelho para aqueles fora do conjunto. Ao posicionar o ponteiro do mouse sobre um nó da rede, é apresentado o nome do pesquisador ou ID com 16 dígitos correspondente ao Lattes caso seja um “nó vermelho” e o valor recebido por aquele pesquisador naquele ano.

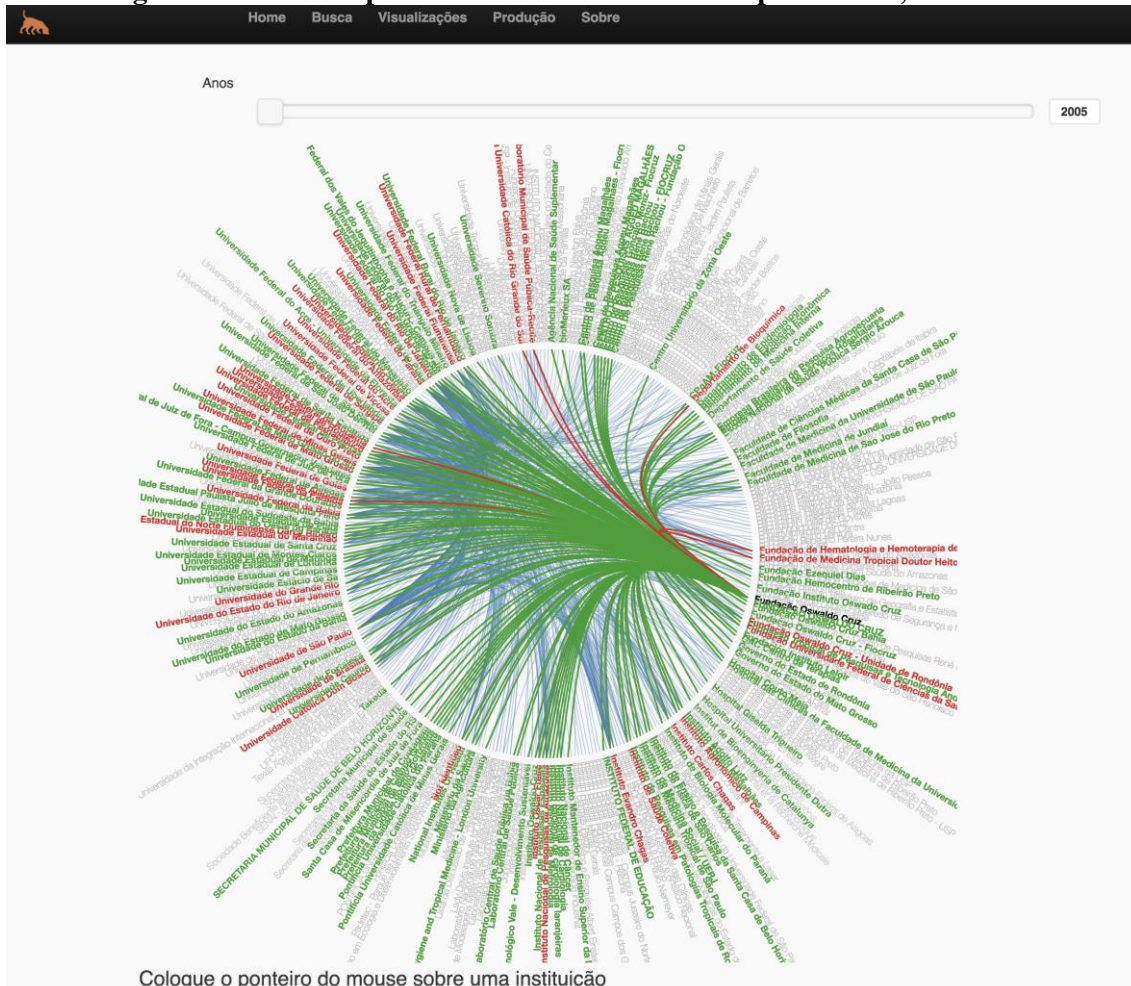
Outros dados são apresentados em forma de texto para compor a visualização, a saber: (i) o tamanho da rede, (ii) a quantidade de pesquisadores em dengue nesta rede, (iii) o valor total financiado e (iv) o percentual que o centro da rede representa na rede, Figura 26.



Fonte: Coutinho-Marques (2014)

O grafo de rede de parcerias institucionais responde a uma pergunta semelhante ao grafo descrito anteriormente, porém, neste caso, sobre as redes institucionais. O grafo foi montado a partir da rede de coautoria, porém levando em consideração as instituições. As arestas – ligação entre as instituições – somente foram efetivadas caso, além da produção em conjunto, também houvesse um financiamento²⁰. As instituições estão dispostas em ordem alfabética no sentido horário. Posicionando o ponteiro do mouse sobre uma instituição, a mesma ficará em negrito enquanto as instituições parceiras e as arestas mudarão de cor. São denotadas com a cor verde caso aquela instituição selecionada tenha iniciado a parceria e vermelho caso a outra instituição tenha iniciado a parceria. O conceito de iniciar a parceria se deu pelo fato de uma instituição ter recebido recursos naquele ano por meio de seus pesquisadores e outra não (Figura 27 a seguir).

Figura 27 – Rede de parcerias institucionais – Exemplo Fiocruz, ano 2005.



Fonte: Coutinho-Marques (2014)

²⁰ O ano não foi levado em consideração para o caso da produção em conjunto, somente para o financiamento.

8.2. Os dados identificados

O período estudado compreendeu desde o ano 2005 até o ano 2014 inclusive. Estes 10 anos contabilizaram um total de 967.736 páginas do Diário Oficial da União contando as seções 1, 2 e 3 do mesmo.

A média de páginas publicadas diariamente, pelo DOU cresceu pouco mais que 50% do primeiro ano estudado até 2014, sendo que a Seção 3 – usada como corpus deste estudo – teve sua média de páginas por dia mais que duplicada conforme verificado abaixo:

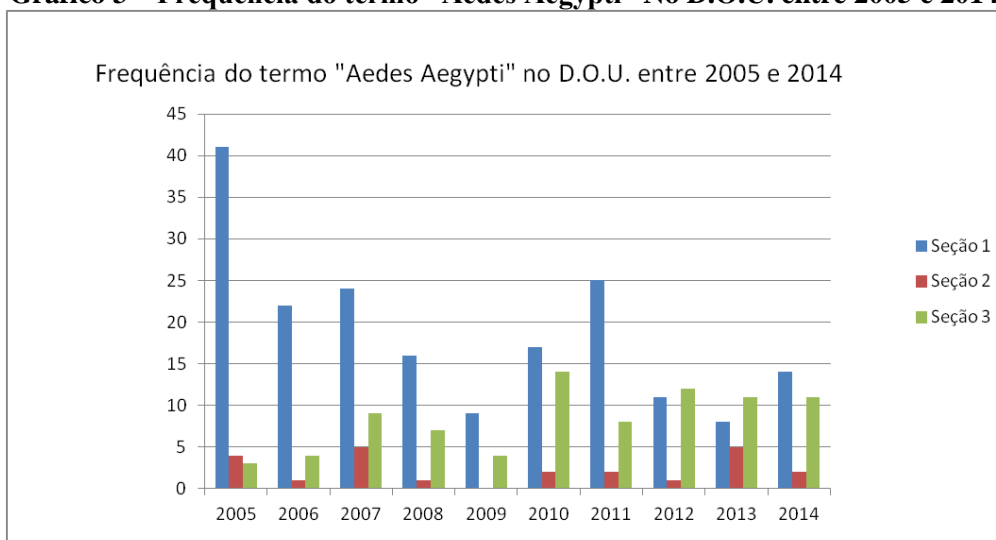
Quadro 9 – Quantidade de páginas do D.O.U. – corpus

Ano	Seção 1		Seção 2		Seção 3		DOU	
	Páginas	Média/dia	Páginas	Média/dia	Páginas	Média/dia	Páginas	Média/dia
2005	35.164	140,10	11.659	46,45	30.187	120,27	77.010	306,81
2006	28.452	114,27	11.212	45,03	37.432	150,33	77.096	309,62
2007	28.164	112,66	10.656	42,62	38.005	152,02	76.825	307,30
2008	31.008	122,08	11.932	46,98	44.796	176,36	87.736	345,42
2009	31.864	127,46	13.568	54,27	45.948	183,79	91.380	365,52
2010	34.480	137,37	14.768	58,84	52.312	208,41	101.560	404,62
2011	41.688	166,09	15.160	60,16	51.784	206,31	108.632	432,56
2012	42.112	167,78	15.404	61,37	57.224	227,98	114.740	457,13
2013	36.836	145,60	17.252	67,13	62.436	239,22	116.524	451,94
2014	35.104	138,75	17.428	68,89	63.701	251,78	116.233	459,42
Total	344.872	137,21	139.039	55,17	483.825	191,65	967.736	384,03

Fonte: Diário Oficial da União, elaboração própria

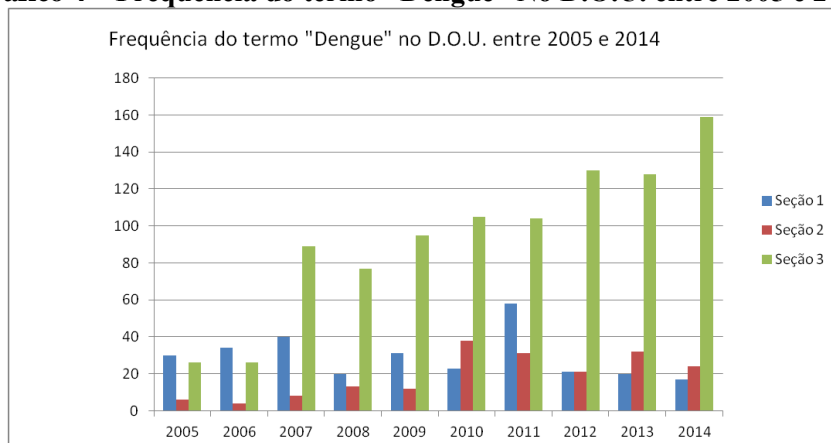
A busca pelos termos “aedes aegypti”, “dengue” e “mosquito” resultaram na distribuição nas três seções do DOU representada abaixo, Gráficos 3, 4 e 5, entre os anos de 2005 e 2014 inclusive.

Gráfico 3 – Frequência do termo “Aedes Aegypti” No D.O.U. entre 2005 e 2014



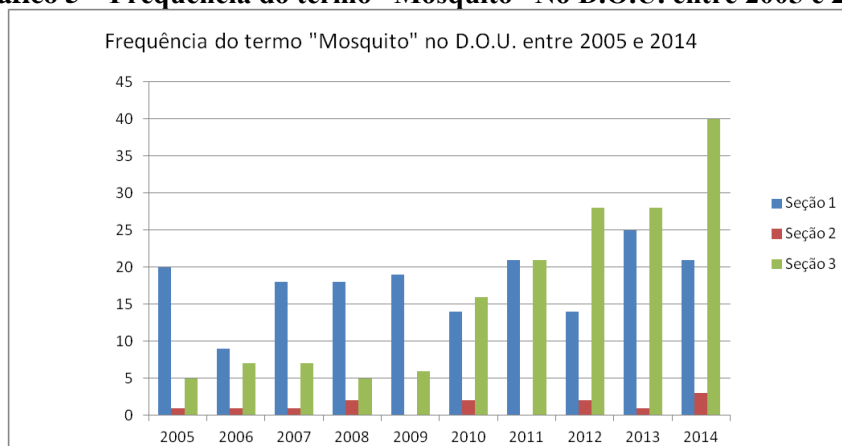
Fonte: Diário Oficial da União – Elaboração Própria

Gráfico 4 – Frequência do termo “Dengue” No D.O.U. entre 2005 e 2014



Fonte: Diário Oficial da União (elaboração própria)

Gráfico 5 – Frequência do termo “Mosquito” No D.O.U. entre 2005 e 2014



Fonte: Diário Oficial da União (elaboração própria)

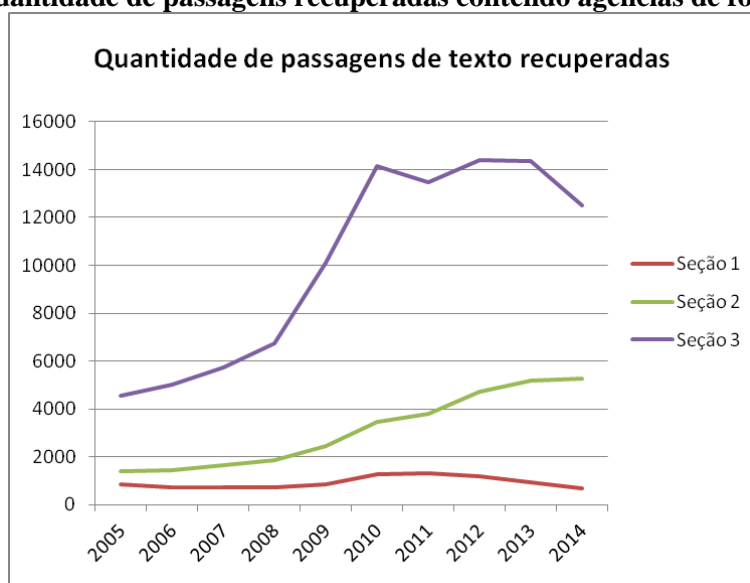
A nova estratégia de se identificar o padrão de financiamento para pesquisa em dengue visou recuperar as passagens de texto onde pudesse ser encontrada pelo menos uma das agências financiadoras. No Quadro 10 e no Gráfico 6 a seguir é representada a quantidade de passagens de texto encontradas organizadas por ano e seção do DOU.

Quadro 10 – Quantidade de passagens recuperadas contendo agências de fomento

ANO	Seção 1	Seção 2	Seção 3
2005	863	1.382	4.554
2006	738	1.423	5.006
2007	706	1.628	5.721
2008	733	1.867	6.733
2009	862	2.454	10.092
2010	1.286	3.474	14.163
2011	1.298	3.785	13.472
2012	1.200	4.716	14.414
2013	949	5.181	14.372
2014	678	5.280	12.503
Total	9.313	31.190	101.030

Fonte: Diário Oficial da União (elaboração própria)

Gráfico 6 – Quantidade de passagens recuperadas contendo agências de fomento nos anos



Fonte: Elaboração própria

Tendo em vista que os editais de financiamento encontram-se na Seção 3, passou-se a trabalhar somente com a referida seção. A distribuição das agências de fomento nas passagens de texto da Seção 3 do DOU pode ser visualizada abaixo:

Quadro 11 – Quantidade de passagens de texto pelas agências de fomento na Seção 3

Qtd.	CNPq	CAPES	FINEP	DECIT
33979	✓			
31434		✓		
27976			✓	
3494				✓
1899	✓	✓		
1555	✓		✓	
270	✓	✓	✓	
257	✓			✓
101		✓	✓	
35			✓	✓
20	✓		✓	✓
6		✓		✓
2	✓	✓		✓
2	✓	✓	✓	✓
0		✓	✓	✓

Fonte: Elaboração própria

Destaca-se que não foi encontrada nenhuma passagem de texto contendo a combinação de agências CAPES–Finep–Decit. Porém, a mesma foi representada no Quadro 11 com a finalidade de compor todas as combinações possíveis.

Estas passagens foram utilizadas para montar a lista de frequência em ordem decrescente dos termos. Foram identificados os termos: **projeto**, **apoio** e **edital** como

importantes dentro do contexto. Tal identificação se deu ao fato destas possuírem uma frequência representativa. Outras palavras, como por exemplo, “**data**” também apresentou uma grande frequência; entretanto esta pode ser usada em qualquer contexto e, se incorporada, criaria um distanciamento do foco aumentando a revocação de passagens, porém diminuindo a precisão.

Sendo assim verificou-se dentre as passagens da seção 3 identificadas anteriormente aquelas que possuíam, além do nome da agência de financiamento ou sua sigla, os três termos supracitados.

A partir das passagens de texto recuperadas foi possível caracterizar (i) chamada para financiamento, (ii) divulgação de resultado, (iii) homologação do resultado caracterizado pela divulgação da entrega da verba para pesquisa e (iv) prorrogação de prazo para o desenvolvimento de uma pesquisa. É sabido que ao término de um projeto de pesquisa o pesquisador responsável tenha que prestar contas, entretanto não foi possível, dentre as 100 passagens escolhidas aleatoriamente, identificar nenhuma notícia contendo dados que caracterizassem a prestação de contas.

O caso da prestação de contas fora resolvido conforme descrito na metodologia e encontrou-se um fluxo de publicação que reflete o modelo do padrão público de financiamento. Segue abaixo, Figura 28, o modelo identificado:

Figura 28 – Modelo do padrão público de financiamento



CALL Chamada para financiamento (edital ou convênio)

RES Resultado

AWR Recebimento da verba para pesquisa

PRO Termo aditivo (prorrogação)

CON Prestação de contas

Fonte: Elaboração própria

Frente a este modelo, foi realizada a classificação manual de 1.200 passagens de texto que resultou no padrão ouro (GOLD) referido na metodologia. Tal classificação resultou na distribuição apresentada no quadro a seguir.

Quadro 12 – Distribuição da classificação manual

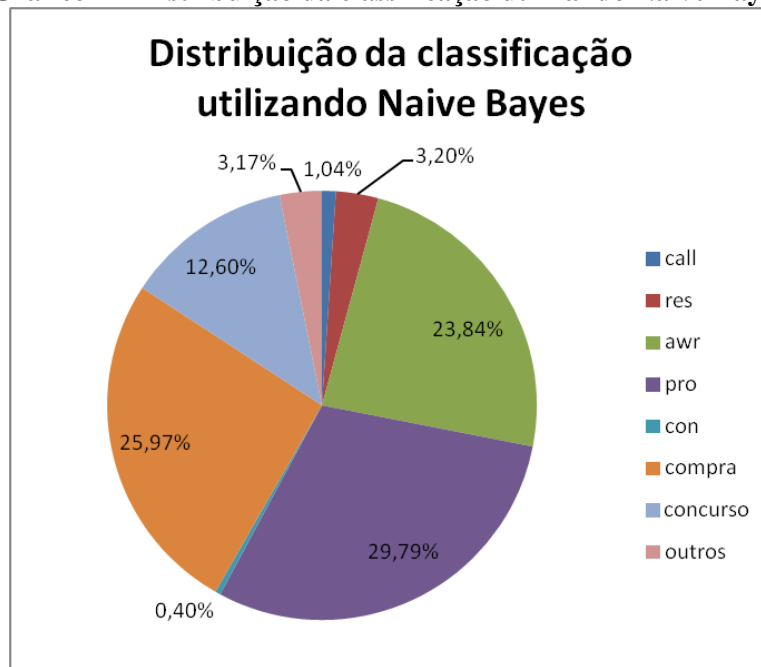
Classificação	Qtd.	Percentual
call	31	2,58
res	21	1,75
awr	287	23,92
pro	315	26,25
con	10	0,83
compra ²¹	299	24,92
concurso ²²	109	9,08
outro	128	10,67

Fonte: Elaboração própria

Ressalta-se que mesmo tendo sido escolha aleatória, a quantidade de passagens classificadas como “call” e “res” é bem menor que a quantidade de “awr”. Isto se atribui ao fato de que a relação não é de um para um, pois a distribuição de recursos (awr) não é realizada toda de uma única vez para os contemplados.

O processo de classificação automática das demais 99.830 passagens de texto (101.030 – 1200 = 99.830) pelo método *tf-idf* foi bem mais rápido que o processo classificado pelo método *Naive Bayes*, não tendo demorado nem um minuto. A seguir, Gráfico 7, encontra-se a distribuição encontrada pelo método de classificação baseada em *Naive Bayes*.

Gráfico 7 – Distribuição da classificação utilizando Naive Bayes



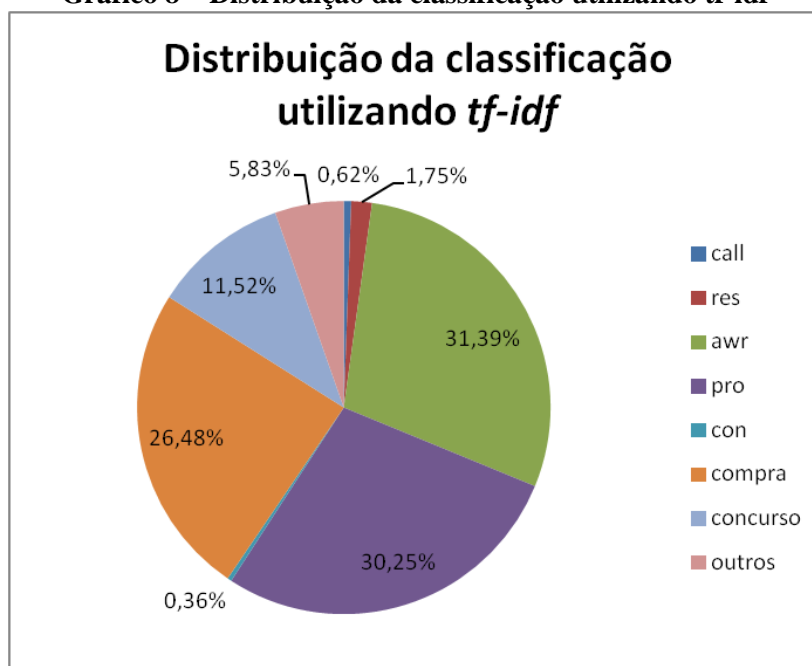
Fonte: Elaboração própria

²¹ A classificação **compra** foi mantida por se tratar de caracterização bem definidas pelas passagens de texto

²² A classificação **concurso** foi mantida por se tratar de caracterização bem definidas pelas passagens de texto

A distribuição da classificação realizada pelo método baseado em *tf-idf* foi bem similar à primeira conforme pode ser verificado no Gráfico 8 a seguir.

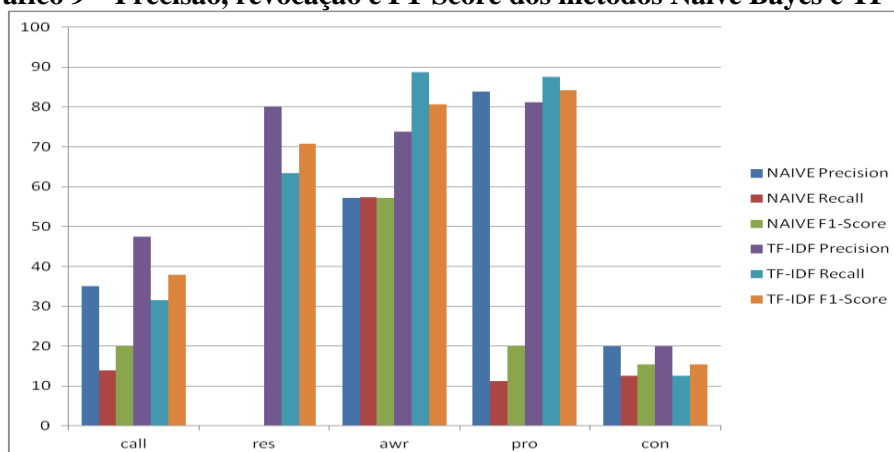
Gráfico 8 – Distribuição da classificação utilizando *tf-idf*



Fonte: Elaboração própria

Já o gráfico 9, apresenta-se o gráfico resultado da matriz de confusão.

Gráfico 9 – Precisão, revocação e F1-Score dos métodos Naive Bayes e TF-IDF



Fonte: Elaboração própria

Pouco mais de 20% das passagens de texto não receberam a mesma classificação para cada um dos dois métodos utilizados. Entretanto, apenas 1% não continham pares de classificação pertinentes ao tema de financiamento. Sendo assim, foram retirados da contabilização todos os pares que apresentavam as classificações de “compra”, “concurso” e “outros” em qualquer combinação que aparecesse tanto na classificação

por Naive Bayes quanto por *tf-idf* (ex. compra-concurso; outro-compra; concurso-outro *etc*).

Verificou-se para quatro das cinco etapas uma coincidência de termos igual ou maior que 93%, Quadro 13. Porém para a etapa de resultado do financiamento (res) encontrou-se uma coincidência de 61%. Este pequeno percentual de coincidência se dá pelo fato dos textos de publicação de resultados serem normalmente pequenos e com palavras com uma baixa frequência como o nome da agência de fomento e o link de publicação do resultado.

Quadro 13 – Comparação entre os métodos de classificação (Naive Bayes e *tf-idf*)

Etapa de financiamento	Coincidência entre os métodos de classificação (Naive Bayes e <i>tf-idf</i>)
call	93%
res	61%
awr	94%
pro	94%
con	97%

Fonte: Elaboração própria

Com a finalidade de verificar o porquê do percentual para divulgação de resultado ser menor do que os demais, foram verificados quais pares de classificação tinham resultado (res) presente, vide Quadro 14.

Quadro 14 – Pares de classificação que contém RES

Naive Bayes	TF-IDF	Quantidade
Res	awr	1033
Res	outro	1028
concurso	res	509
compra	res	120
Call	res	83
Res	compra	72
Res	pro	67
Res	concurso	58
Awr	res	2
outro	res	1

Fonte: Elaboração própria

Destes 2.973 pares de classificação onde aparece como classificado a divulgação de resultado de financiamento (RES) foram escolhidas aleatoriamente 100 passagens e verificadas manualmente frente à lista de termos gerada conforme descrito anteriormente. Verificou-se nestas passagens que 18 das 100 eram referentes à classificação de divulgação de resultado de financiamento. Os termos destas 18 passagens foram identificados e, comparando com a lista gerada, foi possível identificar que as sequências “torna público o resultado” ou “torna pública a divulgação do

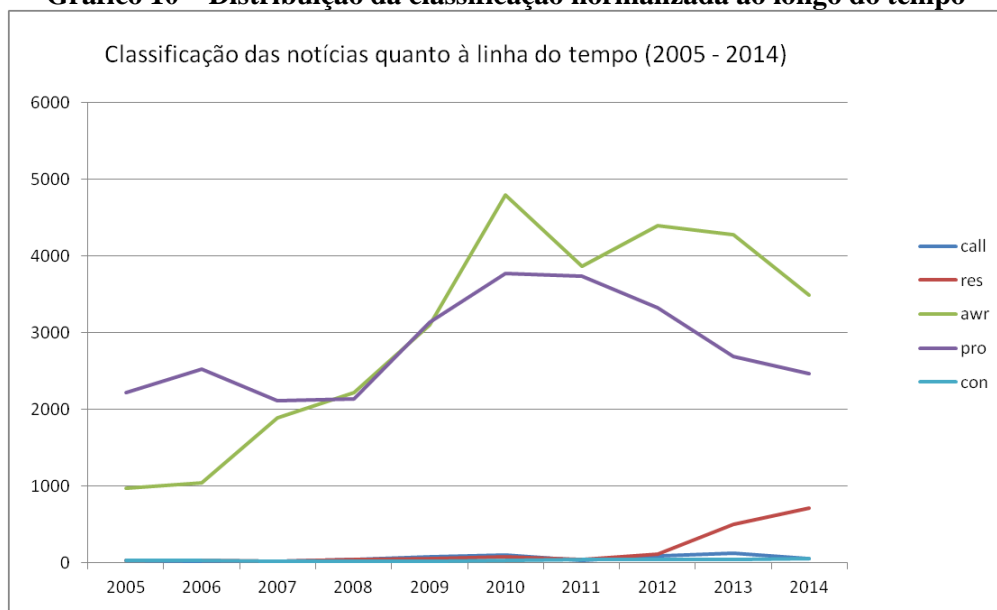
resultado” aparecia em todas. Utilizou-se dos termos “torna”, “públic” (sem a letra “o” ou “a”) e “resultado” para, de forma automática verificar os 2.973 pares de classificação. Um total de pouco mais que 20% foram constatados como pertencendo à classificação de divulgação de resultado (RES).

O cálculo de frequência foi feito apenas para o grupo RES e verificou-se, para este novo grupo de palavras, 82% de coincidência. Ressalta-se aqui que a publicação do resultado de uma chamada de financiamento não apresenta efetivamente nenhum resultado, apenas informado que o resultado se tornou público no site da agência de fomento.

Esta classificação normalizada pelo processo acima foi tomada como base para continuação da pesquisa.

A seguir, Gráfico 9, é apresentada a distribuição da classificação das passagens de texto ao longo do tempo após a normalização. Verifica-se que a quantidade de chamada é quase inalterada ao longo dos dez anos – 2005 a 2014 – entretanto isso não significa que o montante financiado não tenha tido variação ao longo deste período.

Gráfico 10 – Distribuição da classificação normalizada ao longo do tempo



Fonte: Elaboração própria

Usando tal normalização gerou-se a lista de frequência das palavras contidas nas passagens de texto conforme apresentado abaixo, Quadro 15, o qual contém as vinte palavras com maior número de ocorrência para cada classificação. Ressalta-se que RES apresenta uma lista de 18 palavras, pois foram retiradas as *stopwords* mais comuns – artigos e preposições – assim como o nome e siglas das agências de fomento.

Quadro 15 – Lista das 20 palavras com maior frequência para cada classificação

CALL	RES	AWR	PRO	CON
propostas	torna	data	termo	processo
recursos	resultado	vigência	cnpj	Prazo
página	público	cnpj	aditivo	recolhimento
chamada	chamada	assinatura	data	Data
aprovadas	propostas	termo	objeto	contas
internet	aprovadas	objeto	prazo	presente
edital	encontram-se	concessão	concessão	Deste
partir	link	valor	cpf	Fica
cronograma	pública	projeto	assinatura	Dias
união	encontra-se	espécie	beneficiário	CPF
oficial	aprovada	concedente	prorrogação	comprovante
diário	proposta	partir	espécie	Mil
anexo	faixa	meses	concedente	notificação
data	universal	execução	fundação	Desta
parte	juízo	cpf	vigência	Débito
critérios	edital	financeiro	signatários	Trinta
valor	prorrogando	apoio	pesquisa	Cofres
serem	página	recursos	apoio	atendimento
apoio		pesquisa	partes	incerto
tecnologia		auxílio	projeto	Sabido

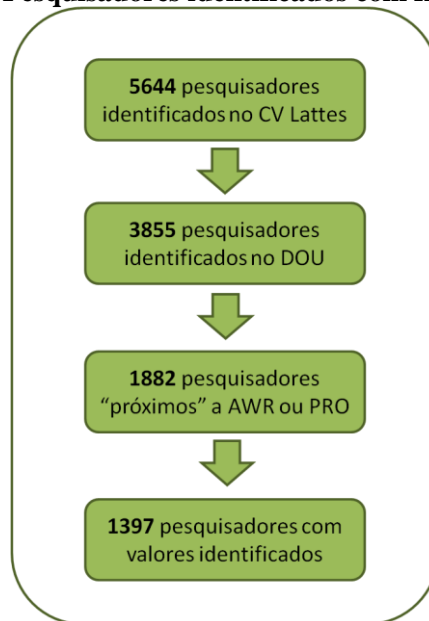
Fonte: Elaboração própria

Estas listas apresentadas no Quadro 15 acima servem como orientação na busca por passagens de texto que representam as etapas do financiamento para pesquisa.

De modo a verificar para quem e quanto foi financiado utilizou-se o Currículo Lattes para identificar os pesquisadores que atuam neste campo. A busca recuperou do Currículo Lattes um total de **5644** nomes de pesquisadores que possuem o termo “dengue” em alguma parte do respectivo currículo. Destes, **3855** foram encontrados na Seção 3 do Diário Oficial da União, porém apenas **1882** foram encontrados próximos²³ de notícias classificadas como AWR ou PRO. Foram identificados valores monetários associados a **1397** pesquisadores. A seguir encontra-se um esquema visual, Figura 29, do que fora descrito neste parágrafo.

²³ distância de no máximo três páginas entre o nome do pesquisador e notícias classificadas como AWR ou PRO

Figura 29 – Pesquisadores identificados com financiamento

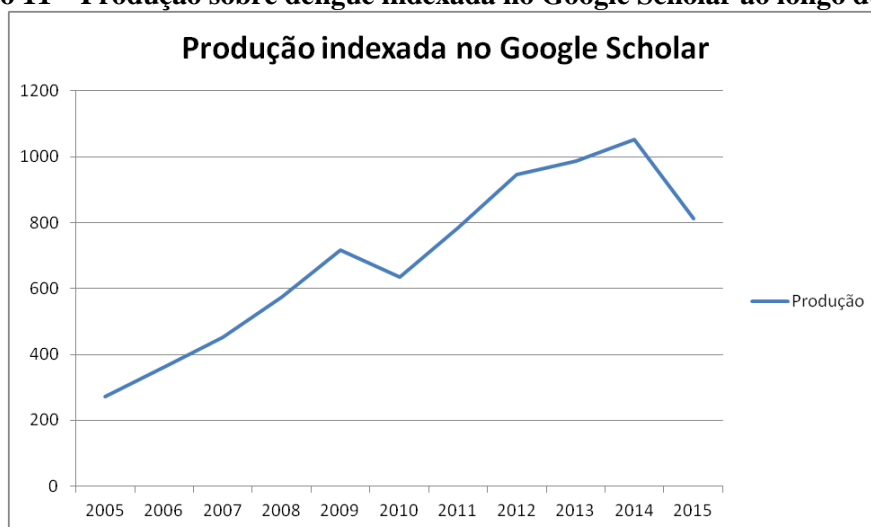


Fonte: Elaboração própria

Para cada um dos mesmos **5644** pesquisadores identificados no CV Lattes foi realizada uma busca na plataforma Google Scholar utilizando o termo “dengue” pelo nome do pesquisador dentre os anos de 2005 e 2015 inclusive. Foram encontrados **2302** daqueles pesquisadores com produção ao longo dos anos.

A seguir, Gráfico 10, é possível verificar a distribuição da quantidade de produção contendo dengue como termo de busca; entre os anos de 2005 e 2015 dos pesquisadores identificados no Currículo Lattes.

Gráfico 11 – Produção sobre dengue indexada no Google Scholar ao longo do tempo



Fonte: Elaboração própria

Foi possível identificar que **1097** pesquisadores identificados inicialmente no Currículo Lattes não receberam financiamento por agência federal entre o período de 2005 a 2014, porém possuíram algum tipo de produção em dengue indexada pelo Google Acadêmico. Também foram identificados **523** pesquisadores do mesmo conjunto inicial que não tiveram produção indexada pelo Google Acadêmico, porém receberam financiamento em nível federal. Por outro lado, foi identificado um total de **2819** pesquisadores que não possuem nem financiamento encontrado em nível federal e nem produção nesta temática indexada no Google Acadêmico.

A distribuição do financiamento pelas agências de fomento ao longo dos 10 anos estudados é apresentada a seguir:

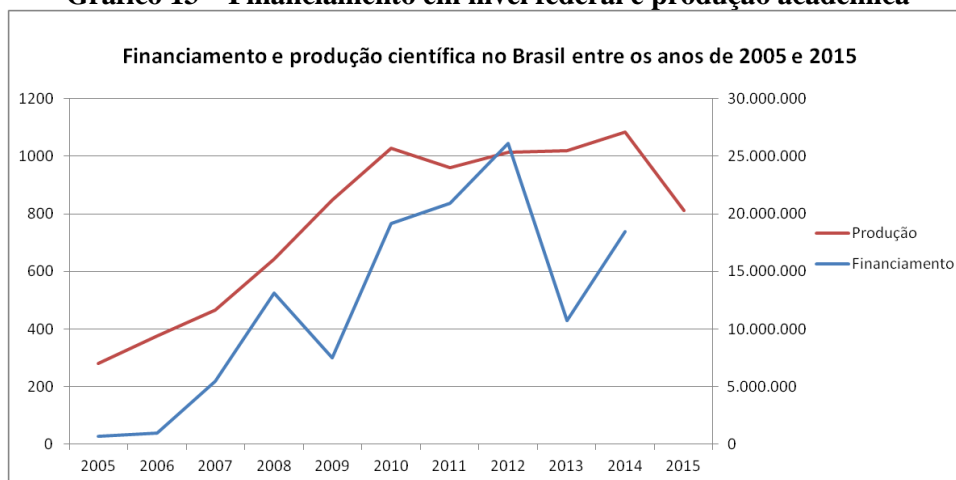
Gráfico 12 – Financiamento em nível federal para pesquisa em dengue



Fonte: Elaboração própria

Ao sobrepor o gráfico de quantidade de produção ao longo do tempo ao financiamento em nível federal para pesquisa em dengue encontramos:

Gráfico 13 – Financiamento em nível federal e produção acadêmica



Fonte: Elaboração própria

Também foi identificado que existem pesquisadores que receberam financiamento por mais de uma agência de financiamento num mesmo ano, sendo este total de **308** pesquisadores. Porém, ressalta-se aqui que no ano de 2008, por exemplo, o pesquisador “Wanderli Pedro Tadei” recebeu financiamento de cinco diferentes combinações de agências financiadoras – códigos 1, 4, 5, 6 e 12 neste estudo – e, ao verificar o referido pesquisador, identificou-se que o mesmo ocupava o cargo de Diretor Substituto no Instituto Nacional de Pesquisas da Amazônia (INPA). Para este caso, alguma verba destinada ao instituto foi certamente atribuída ao pesquisador por este estar como responsável pela instituição.

Tendo isto em vista, verificam-se a seguir, Quadro 16, as dez instituições que mais receberam financiamento diretamente e/ou por meio de seus pesquisadores.

Quadro 16 – Financiamento recebido entre 2005 e 2014 pelas instituições (top 10)

Financiamento em reais (R\$)	(%)	Instituição
247.611.788,47	30,62	Fundação Oswaldo Cruz
174.285.880,33	21,55	Universidade Federal do Rio de Janeiro
80.626.510,15	9,97	Universidade Estadual Paulista Júlio de Mesquita Filho
66.329.039,28	8,20	Pontifícia Universidade Católica de Goiás
56.168.448,77	6,95	Universidade Federal de Minas Gerais
52.896.214,18	6,54	Universidade de São Paulo
40.028.099,23	4,95	Universidade Federal do Espírito Santo
39.436.065,89	4,88	Universidade Federal de Pernambuco
26.554.241,24	3,28	Instituto Nacional de Pesquisas da Amazônia
24.703.517,43	3,05	Universidade Estadual de Campinas
808.639.804,97	100,00	Total

Fonte: Elaboração própria

O caminho realizado permitiu identificar um total de 808 milhões de reais distribuídos pelas dez instituições que mais receberam recursos financeiros entre os anos de 2005 e 2014.

Foram verificados os valores atribuídos para os pesquisadores vinculados à instituição posicionada em primeiro lugar, no Quadro 16 acima. Verificou-se que o pesquisador **Paulo Ernani Gadelha Vieira** foi responsável por pouco menos que **50%** do financiamento atribuído àquela instituição. Uma busca por seu nome no portal da Fundação Oswaldo Cruz (Fiocruz) verificou-se que o pesquisador é presidente da mesma desde o ano de 2010. De mesma sorte, **Carlos Rangel Rodrigues** ocupou desde 2006 o cargo de Diretor de Farmácia da **Universidade Federal do Rio de Janeiro**. Ressalta-se aqui que o valor total atribuído a instituições não identificadas foi de R\$209.70.284,18; o que ocuparia a décima primeira posição neste *ranking*. Destes,

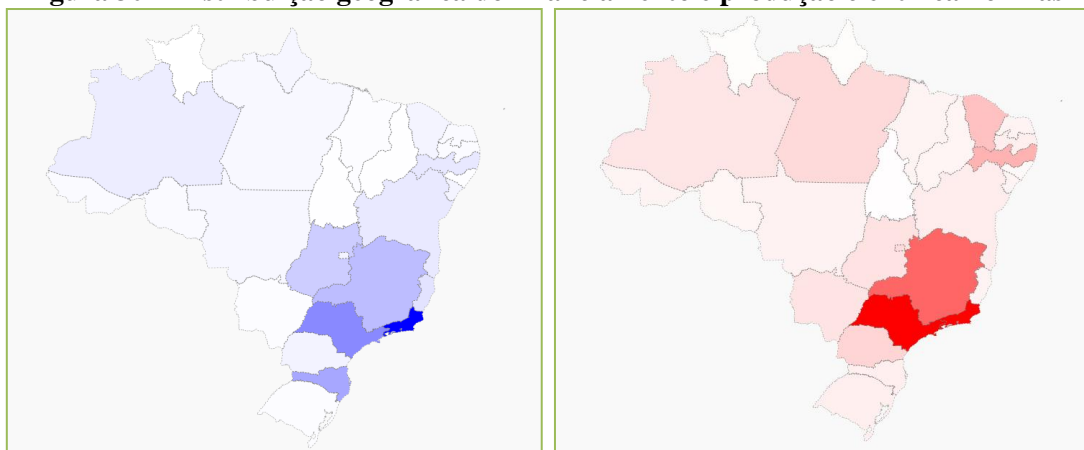
R\$17.437.250,00 foram atribuídos a Cristiano Barbosa o qual ocupava o cargo de Secretário Extraordinário Substituto da SES-GE/MJ.

Esta distribuição de valores se apresenta basicamente nas instituições públicas. Entretanto entre as dez instituições que mais receberam recursos entre os anos de 2005 e 2014 encontra-se a **Pontifícia Universidade Católica de Goiás**. A pesquisadora **Milca Severino Pereira** foi responsável por quase 99,94% deste valor. Na ocasião, a referida pesquisadora se encontrava em cargo de gestão da Secretaria Estadual de Educação de Goiás.

Quanto à distribuição do número de pesquisadores a região Sudeste conta com quase 50% do número de pesquisadores nesta temática. Seguido pela região nordeste com pouco menos de 20% e uma distribuição semelhante entre as demais regiões Centro-Oeste, Sul e Norte.

Para fins de verificação da distribuição do financiamento pelos Estados Brasileiros entre os anos de 2005 e 2014, foi desenhado um mapa contendo a distribuição conforme apresentado abaixo, Figura 30.

Figura 30 – Distribuição geográfica do financiamento e produção científica no Brasil



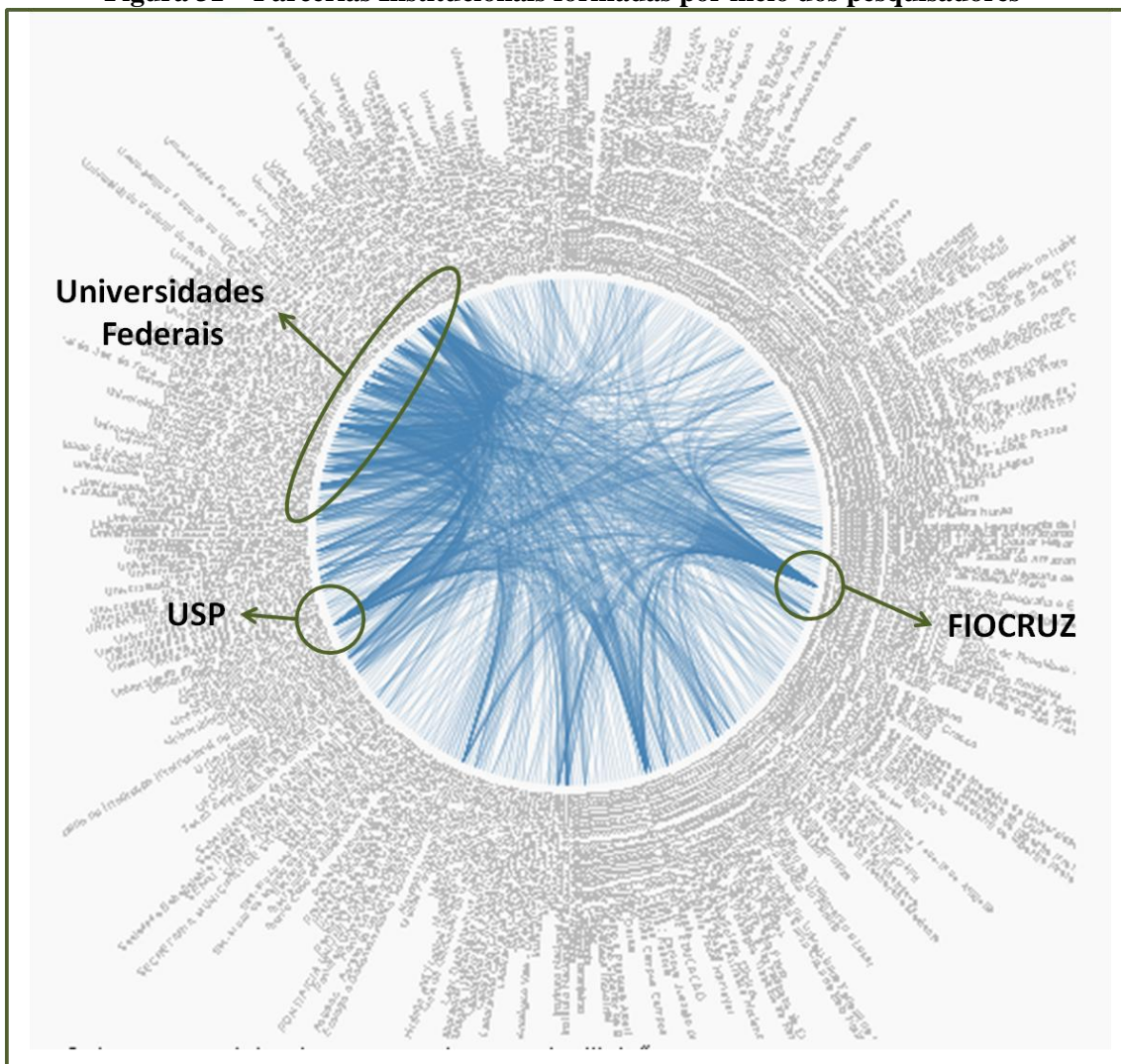
Fonte: Elaboração própria

Verifica-se claramente uma concentração de financiamento, em escala de azul à esquerda, assim como da produção científica, em escala de vermelho à direita, na região sudeste ao longo dos 10 anos que compõem este estudo – 2005 a 2014.

As parcerias institucionais que se deram por meio das pesquisas desenvolvidas por pesquisadores também podem ser visualizadas por meio de um grafo de redes desenvolvido para tal finalidade, Figura 31 a seguir. Este grafo possibilita visualizar as parcerias que se formaram ao longo dos anos caso uma das instituições envolvidas tenha recebido algum financiamento por meio de seus pesquisadores. Verifica-se abaixo o ano

de 2005, porém é possível ter este mesmo grafo para os demais anos, os quais praticamente repetem a mesma situação.

Figura 31 – Parcerias Institucionais formadas por meio dos pesquisadores

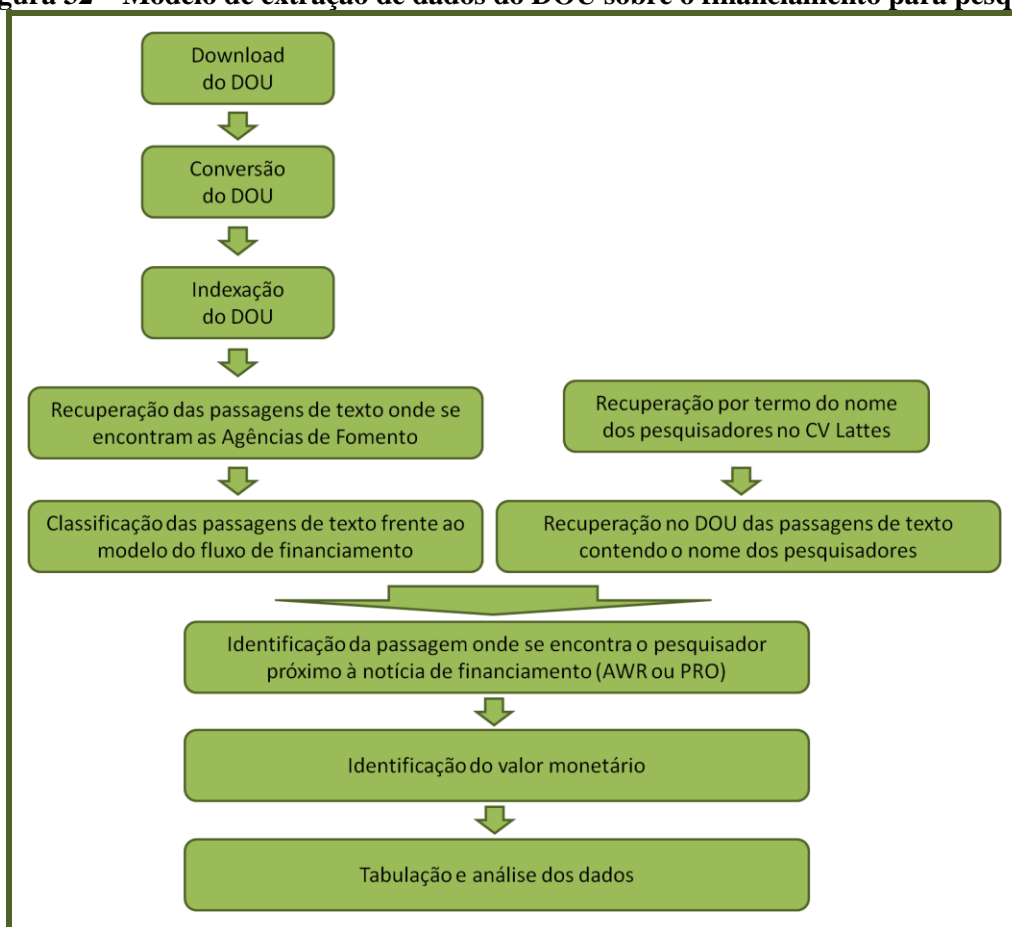


Fonte: Elaboração própria

8.3. Modelo de extração de dados do DOU sobre o financiamento para pesquisa

O modelo apresentado no início do capítulo que descreve a metodologia, Figura 6, inclui a captura dos dados do Google Acadêmico, entretanto este item, Figura 32 a seguir, apresenta o modelo de extração de dados do Diário Oficial da União sobre o financiamento para pesquisa em dengue tendo o Currículo Lattes como suporte para identificação dos pesquisadores a serem buscados no DOU.

Figura 32 – Modelo de extração de dados do DOU sobre o financiamento para pesquisa



Fonte: Elaboração própria

9. Discussão

Este capítulo discute os resultados encontrados quanto ao financiamento para pesquisa em dengue no Brasil frente aos dados disponíveis pelas diversas ferramentas das agências de fomento e instituições internacionais como o G-Finder. Neste capítulo serão também apresentados os limites encontrados no desenvolvimento desta pesquisa assim como aponta para possíveis desdobramentos para o aprofundamento do conhecimento sobre o financiamento para pesquisa no Brasil. Isto posto, este capítulo se encontra organizado em duas partes sendo que a primeira abriga as questões do financiamento e a segunda reúne limites e desdobramentos.

9.1. Financiamento para pesquisa em dengue no Brasil

É inegável que o financiamento para pesquisa é de extrema importância para o desenvolvimento científico e tecnológico de um país. Entretanto, a relação entre o financiamento e os resultados a serem avaliados não são claros. A “linha de produção” onde entra recursos por uma ponta – financiamento para pesquisa – e sai “produto” na outra, em forma de artigos científicos e patentes na outra, explicitando o modelo linear, não atende mais às demandas colocadas para o desenvolvimento do país.

Porém, poucos são os estudos que explicitam este modelo. Em outras palavras, os poucos dados que se encontram sobre o financiamento para pesquisa não são agregados facilmente por conta dos diferentes níveis de detalhamento em que são apresentados pelas diversas agências de fomento.

Os valores encontrados para o financiamento para pesquisa em dengue mostram que do ano de 2005 ao ano de 2009 houve um crescimento de três vezes e meia no valor financiado, quando em 2010 os valores caíram pela metade.

A partir de 2010 os valores voltaram a crescer até o ano de 2013, quando atingiu quase o mesmo valor de 2009. Já para 2014 uma nova queda foi registrada, caindo mais que a metade do valor do ano anterior, vide Gráfico 11 descrito na seção anterior. Há que se perguntar se tal padrão se relaciona com o calendário eleitoral presidencial no Brasil. O Brasil teve eleições presidenciais no ano de 2010 e 2014.

Do período estudado – 2005 a 2014 – a região Sudeste ocupou o primeiro lugar em alocação de recursos em quase todos os anos, com exceção de 2009, quando a região Sul tomou seu lugar deixando-a em segundo. O fato de a região Sudeste ocupar o primeiro lugar se justifica pela distribuição de pesquisadores no tema. Com cerca de 50% dos pesquisadores nesta temática estarem vinculados à instituições desta região,

parece previsível que o financiamento seja direcionado em maior quantidade para estes. Porém, em contra ponto, a região Sul tendo **seis vezes** menos pesquisadores que a primeira ocupou, como descrito, a primeira posição no ano de 2009. Foi verificado, porém, que o pesquisador “Luis Antonio Silva” foi responsável por quase toda verba, especialmente a institucional e não de recursos destinados diretamente a ele. Este foi responsável por verba institucional e não por verba destinada a ele para desenvolvimento de pesquisa. Neste caso, se descontarmos o valor atribuído a ele, a região Sudeste passaria a primeiro neste ano também e a região Sul ficaria em quarta posição. Ressalta-se que a região Sul ocupou a quinta posição em todos os outros anos.

Em relação às agências de fomento, o CNPq foi o órgão responsável sozinho por 36% do volume monetário financiado ao longo dos dez anos deste estudo, seguido pela Finep com 32% e pela CAPES com 23%. Para este estudo foram consideradas quatro agências de fomento e todas as combinações possíveis entre elas, vide Quadro 11 apresentado anteriormente. A própria natureza das agências explica tal distribuição. Entre outros, ao CNPq compete “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” (CNPQ, [s.d.]).

A Finep possui como missão a promoção e “o desenvolvimento econômico e social do Brasil por meio do fomento público à Ciência, Tecnologia e Inovação em empresas, universidades, institutos tecnológicos e outras instituições públicas ou privadas”. (FINEP, [s.d.]). Embora financie instituições de pesquisa e, por conseguinte, pesquisadores, a Finep distribui seus recursos provenientes do Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT) também para empresas. Sabendo que um grande número de pesquisadores encontra-se em instituições públicas, é possível inferir que grande parte do financiamento desta agência de fomento seja destinada para empresas.

Já a CAPES, enquanto fundação do Ministério da Educação (MEC) promove a formação e capacitação de pesquisadores das diversas áreas do conhecimento. Em outras palavras, esta agência de fomento não tem como foco o financiamento de pesquisas.

A CAPES, a partir dos dados do Portal da Transparência, apresenta um crescimento nos valores de auxílio a pesquisadores e no total de suas despesas, entre os anos de 2005 e 2014, conforme pode ser verificado no Quadro 17 a seguir.

Quadro 17 – Auxílio a Pesquisadores pela CAPES nos anos de 2005 a 2014

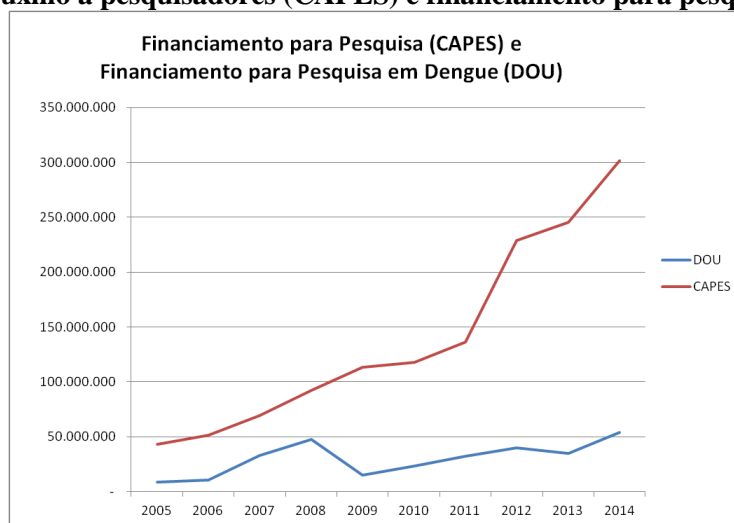
Ano	Valor (Auxílio a Pesquisadores)	Valor Total (para todas as despesas)
2005	42.851.846,75	600.916.981,98
2006	51.345.111,70	640.901.519,75
2007	69.457.215,19	694.379.871,55
2008	92.274.771,08	932.802.306,70
2009	113.249.133,46	1.158.104.704,32
2010	118.076.586,90	1.497.242.836,49
2011	136.522.110,61	1.978.496.540,11
2012	228.593.012,84	2.823.862.680,24
2013	245.526.437,45	4.260.668.892,87
2014	301.724.557,69	5.404.354.543,46

Fonte: Portal da Transparência (2016)

Entretanto, os dados levantados a partir do Diário Oficial da União apontam em outra direção. Verifica-se uma queda no ano de 2009 para esta agência de fomento e depois se inicia um novo crescimento.

O financiamento encontrado para os pesquisadores de dengue realizados pela CAPES foi, em média, vinte e cinco por cento menos que o financiamento apresentado pela própria agência no Portal da Transparência, muito embora se apresente um afastamento ao longo do tempo. Estes 25% são uma ótima marca para a pesquisa em dengue. O ano de 2009 foi o ano que apresentou o menor percentual, com 13% enquanto no ano de 2008 a fatia do financiamento para pesquisadores em dengue foi de 51%.

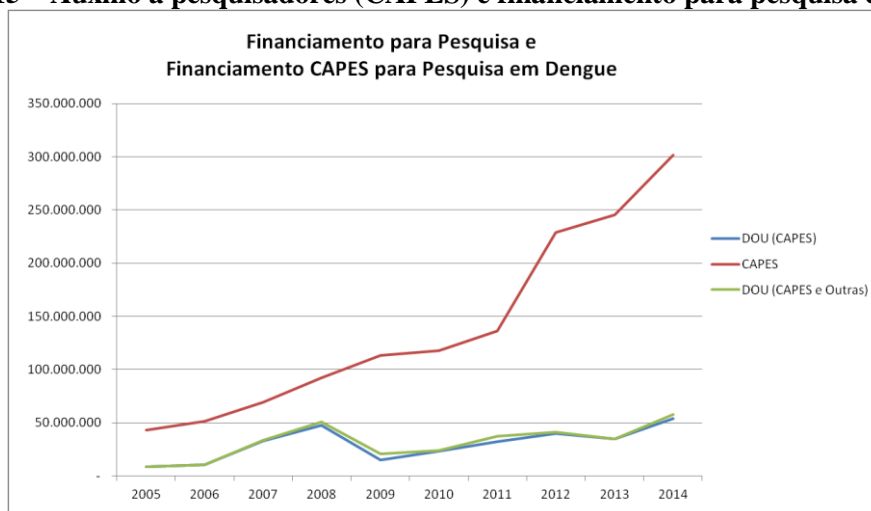
Gráfico 14 – Auxílio a pesquisadores (CAPES) e financiamento para pesquisa em dengue



Fonte: Esta pesquisa; BRASIL, ([s.d.])

Mesmo agregando os financiamentos conjuntos da CAPES com outras agências de fomentos o crescimento não é nem de longe significativo conforme pode ser verificado no Gráfico 14, a seguir.

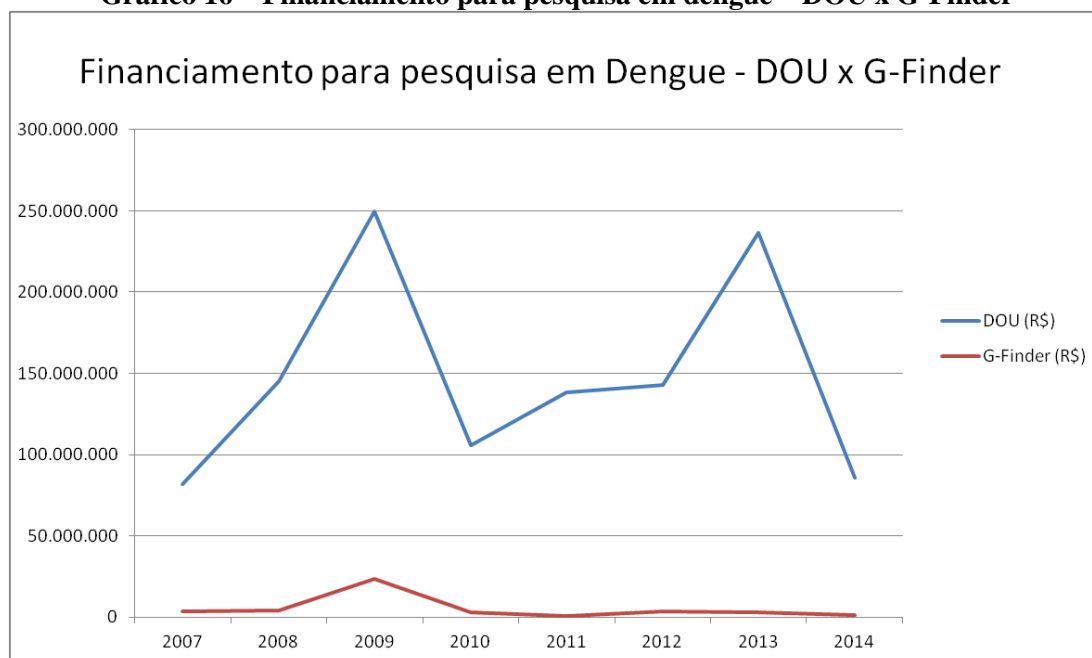
Gráfico 15 – Auxílio a pesquisadores (CAPES) e financiamento para pesquisa em dengue



Fonte: Esta pesquisa; BRASIL, ([s.d.])

A comparação entre o financiamento destinado para pesquisadores em dengue no Brasil entre os anos 2007 e 2014 é muito superior ao encontrado no site do G-Finder como pode ser verificado no Gráfico 15 a seguir.

Gráfico 16 – Financiamento para pesquisa em dengue – DOU x G-Finder



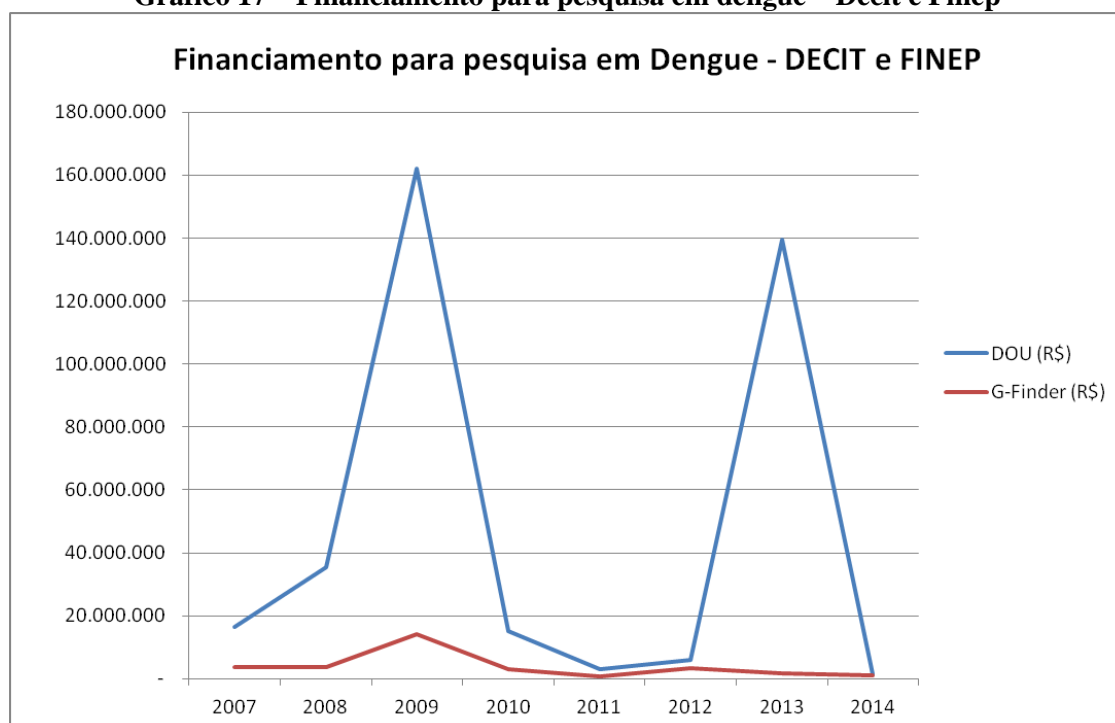
Fonte: Esta pesquisa, Policy Cures; Bill & Mellinda Gates Foundation, ([s.d.])

Sobre esta visão é importante ressaltar que o G-Finder é um levantamento realizado por uma organização internacional e que os dados encontrados no mesmo são auto declaratórios pelas instituições que participaram do levantamento. Sendo assim, é possível que não constem todos os financiamentos neste site. Adiciona-se a isto o fato

de que 26,5% do montante apresentado pelo G-Finder é referente ao financiamento proveniente das Fundações de Amparo à Pesquisa (FAPs) ou outros órgãos locais como Secretarias de Saúde do Estado.

Por outro lado, das quatro agências de financiamento deste estudo – CNPq, CAPES, FINEP e DECIT – somente as duas últimas constam dos dados apresentados pelo G-Finder. Neste caso, se considerarmos apenas as duas agências, de forma isolada das demais, isto é, sem considerar financiamentos conjuntos, verifica-se que embora haja distância entre os valores, o padrão de variação ao longo dos anos se mantém semelhante. Verifica-se ainda que para os anos de 2011 e 2012 o *gap* entre as duas fontes é bem menor que para os demais anos.

Gráfico 17 – Financiamento para pesquisa em dengue – Decit e Finep

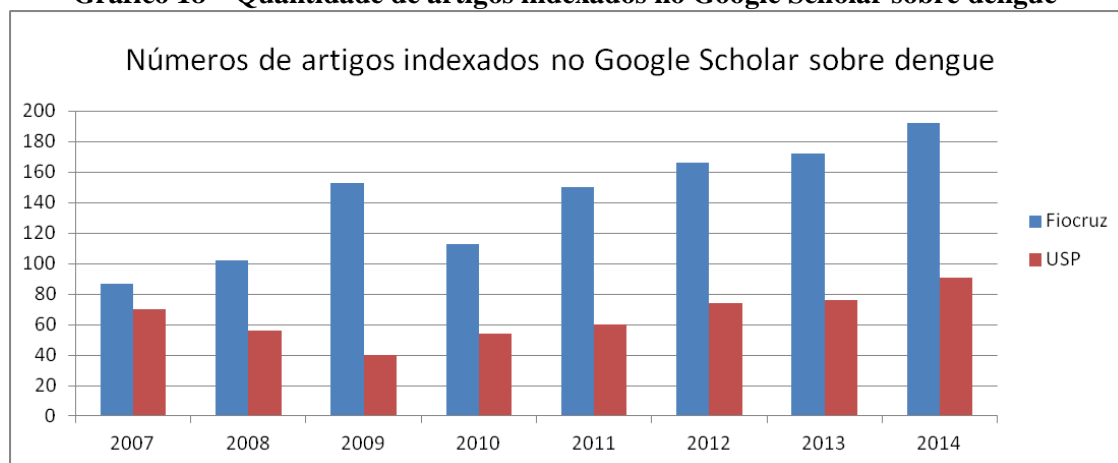


Fonte: Esta pesquisa, Policy Cures; Bill & Mellinda Gates Foundation, ([s.d.])

Embora o critério metodológico do G-Finder seja reconhecido, é importante ressaltar novamente que os dados são provenientes de um questionário aplicados às instituições e que estas declaram o que é possível, pois muitas vezes nem estas possuem os dados que refletem a realidade. É impossível imaginar que grandes instituições de nível federal como a Fundação Oswaldo Cruz (Fiocruz) e a Universidade de São Paulo (USP) não pesquisem dengue com regularidade. Os dados apresentados pelo portal do G-Finder demonstram que não houve financiamento para pesquisa em dengue no ano de 2010 recebido pela Fiocruz. Já para a USP, somente consta financiamento para pesquisa

em dengue no ano de 2008. Adiciona-se a esta incoerência o fato de pesquisadores de ambas as instituições terem publicados artigos sobre o tema dengue.

Gráfico 18 – Quantidade de artigos indexados no Google Scholar sobre dengue



Fonte: Esta pesquisa

Já o CNPq, conforme descrito anteriormente, fornece seus dados agregados por áreas e subáreas de conhecimento ao decorrer dos anos desde 2006. No Quadro 18 a seguir pode ser encontrada a distribuição do financiamento pelas áreas (i) Ciências Biológicas, (ii) Ciências Humanas, (iii) Ciências Sociais Aplicadas e pela subárea (iv) Saúde Coletiva. Entende-se que a área das Ciências Biológicas compreende o estudo do vetor, as áreas das Ciências Humanas e Ciências Sociais Aplicadas, embora quase esquecidas, são de grande importância por compreenderem a educação em saúde. Já a subárea da Saúde Coletiva tem, para esta pesquisa, sua importância por conta da epidemiologia.

Quadro 18 – Financiamento do CNPq para pesquisa segundo CNPq – 2006 a 2014

Ano	Ciências Biológicas	Ciências Humanas	Ciências Sociais Aplicadas	Saúde Coletiva
2006	423.897.719,00	142.748.236,00	85.577.685,00	122.041.877,00
2007	317.224.457,00	97.189.633,00	71.529.239,00	46.745.419,00
2008	1.520.593.195,00	326.105.301,00	211.610.963,00	150.360.596,00
2009	593.669.819,00	130.626.578,00	86.040.019,00	56.418.574,00
2010	1.000.139.621,00	198.844.833,00	166.703.802,00	88.461.177,00
2011	347.512.206,00	180.553.928,00	104.422.205,00	38.554.124,00
2012	685.403.845,00	249.835.316,00	147.390.902,00	67.238.822,00
2013	1.137.484.177,00	259.096.555,00	348.101.712,00	175.503.642,00
2014	1.257.508.145,00	428.368.967,00	290.605.469,00	175.969.430,00
	7.283.433.184,00	2.013.369.347,00	1.511.981.996,00	921.293.661,00

Fonte: CNPq, ([s.d.])

Referente aos dados levantados do DOU, para as mesmas áreas do conhecimento, temos no Quadro 19 a seguir o que fora financiado para dengue pelo

CNPq e agências parceiras – CAPES, DECIT e/ou FINEP – em conjunto com a primeira.

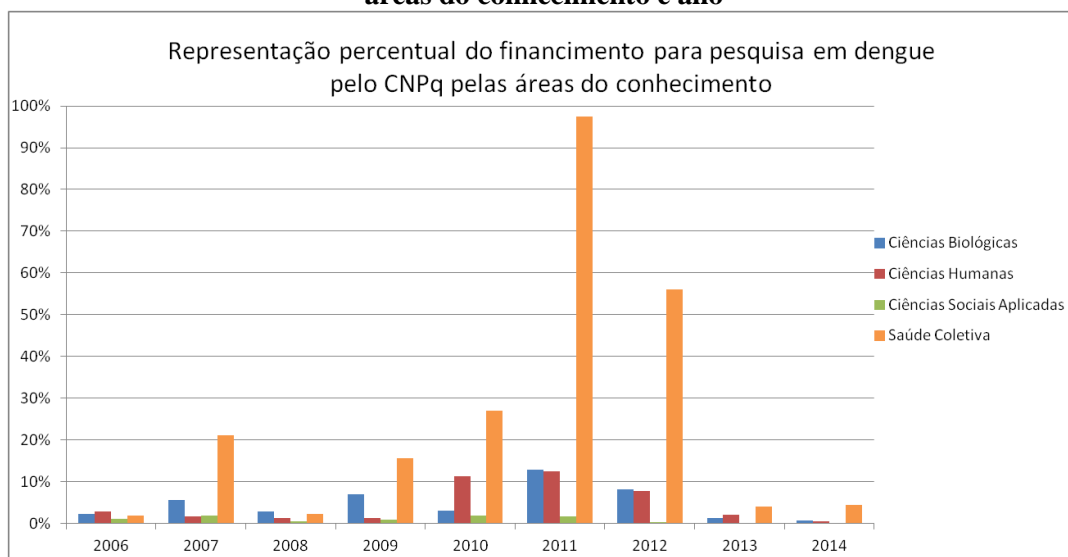
Quadro 19 – Financiamento do CNPq para pesquisa segundo DOU – 2006 a 2014

Ano	Ciências Biológicas	Ciências Humanas	Ciências Sociais Aplicadas	Saúde Coletiva
2006	10.062.429,64	4.071.754,22	902.543,88	2.223.121,74
2007	17.649.854,30	1.658.194,38	1.419.308,40	9.846.525,87
2008	42.535.282,61	4.280.350,43	1.088.752,79	3.575.877,51
2009	41.826.266,44	1.731.537,27	760.211,42	8.794.064,96
2010	30.268.974,11	22.344.514,01	2.985.938,89	23.881.901,59
2011	44.649.960,33	22.661.740,36	1.692.741,11	37.570.533,59
2012	55.372.144,63	19.453.966,82	467.154,16	37.667.411,33
2013	13.902.204,65	5.361.653,66	14.000,00	6.949.715,45
2014	8.417.723,34	2.316.410,56	0,00	7.756.321,75
	264.684.840,05	83.880.121,71	9.330.650,65	138.265.473,79

Fonte: Diário Oficial da União

Para fins de comparação os valores da segunda tabela foram divididos pelos da primeira para verificar o quanto o financiamento em dengue representa sobre o financiamento total para uma determinada área do conhecimento, Gráfico 18 a seguir. Chama a atenção alguns anos para a saúde coletiva onde a representatividade ultrapassa os 10%. Porém em especial o ano de 2011 onde os valores encontrados chegam a quase 100% do que fora financiado pelo CNPq na subárea da Saúde Coletiva.

Gráfico 19 – Representatividade do financiamento para pesquisa em dengue do CNPq por áreas do conhecimento e ano



Fonte: Diário Oficial da União

Pelo menos 80% dos valores tidos como financiados para dengue no ano de 2011 se agrupam em um conjunto de 31 pesquisadores do total de 332 pesquisadores

que receberam financiamento naquele ano. Destes 31 pesquisadores, pelo menos 10 ocupavam cargo de gestão naquela época o que pode caracterizar o recebimento de financiamento institucional e não para o desenvolvimento de pesquisa em dengue. Este tipo de caracterização se torna difícil também pelo fato de inúmeros pesquisadores atuarem em mais de uma área do conhecimento.

Segundo dados apresentados pelo DECIT seu pico de valor financiado para pesquisa em dengue foi em 2009, vide Gráfico 2 apresentado na seção Pesquisa em Saúde no Brasil. Porém, para os dados recuperados do DOU o pico para esta agência em conjunto com as parceiras foi em 2011.

A lista de 134 pesquisadores financiados pelo DECIT entre os anos de 2005 e 2014 para pesquisa em dengue serviu para iniciar a validação da metodologia. Desta lista inicial 47 não possuíam valores nem para o DECIT e nem para a agência parceira. Em outras palavras, embora o DECIT tenha apresentado que o pesquisador recebeu alguma verba, a referida agência não apresentou os valores deste financiamento. Dos 87 que sobraram, 15 não possuíam link para o CV. Da lista de 72 pesquisadores restantes, foram escolhidos aleatoriamente cinco pesquisadores com o intuito de verificar se os mesmos haviam sido recuperados pela metodologia.

O Quadro 20 a seguir apresenta os achados:

Quadro 20 – Pesquisadores utilizados para conferência dos dados do DECIT

Caso	Ano	Pesquisador	Parceiros	Valor	Confere
1	2007 ¹	Mauro Martins Teixeira	CNPq	150.000,00	SIM
2	2007 ¹	Jerson Lima da Silva	CNPq	123.000,00	NÃO
3	2008	Marcelo Dias Baruffi	CNPq	40.000,00	NÃO
4	2011	Christovam Barcellos	CNPq	50.000,00	NÃO
5	2009	Silvania Sousa do Nascimento	CNPq	128.822,72	NÃO

¹ O ano apresentado pelo DECIT é o ano do edital, o pesquisador recebeu a verba em janeiro do ano seguinte

Fonte: Elaboração Própria

O termo “Confere” na tabela é pertinente se as agências, valor e data correspondem no que foi recuperado pela metodologia do Diário Oficial da União ao que fora disponibilizado pelo DECIT. No caso 2, o pesquisador em questão fora identificado pela metodologia tendo recebido duas vezes o valor descrito. De volta ao Diário Oficial da União (DOU) foi verificado que a notícia havia sido publicada duas vezes sendo a primeira em 15/01/2007 iniciando a notícia na página 7 e em 18/01/2007, coincidentemente na mesma página.

No caso 3, embora o referido financiamento encontre-se no DOU de 26/12/2008, na página 7, o pesquisador não se encontra na base de dados de pesquisadores que

estudam sobre dengue. Isto significa que o motor de busca do CV Lattes não recuperou este pesquisador. O termo dengue consta de seu currículo.

Para o caso 4, o pesquisador aparece duas vezes no ano de 2011, porém somente em uma das vezes tem algum valor associado ao seu nome. Seu nome e valores são encontrados em uma tabela, porém em duas linhas diferentes como pode ser verificado abaixo, Figura 33.

Figura 33 – Notícia do D.O.U. referente a financiamento

Espécie: Termo de Concessão e Aceitação de Apoio Financeiro a Projeto de Pesquisa, Encomenda, - CONCEDENTE: CNPq; OBJETO DO TERMO: Concessão de auxílio financeiro a projeto de natureza científica, tecnológica ou de inovação - VIGÊNCIA: 24 (vinte e quatro meses) a partir da data de publicação no Diário Oficial da União, assinatura: 12/12/2011, SIGNATÁRIO: pelo CNPq; José Aureliano Fonseca Matos - Coordenador Geral de Execução do Fomento Substituto - BENEFICIÁRIO: abaixo:				
BENEFICIÁRIO	PROCESSO	TOTAL CUSTEIO + CAPITAL	EMPENHO NÚMERO	BOLSAS VIGÊNCIA QUANTIDADE
Celia Regina Pierantoni	552745/11-1	687.993,69 C 100.000,00 K	2011NE020759 2011NE020757	EV-3, 18 m, 02 EV-2, 18m, 01 EV-3, 09m, 04
Christovam Barcellos	552746/11-8	450.000,00 C 50.000,00 K	2011NE020755 2011NE020760	—
Belita Koiller	490160/11-5	99.680,00 C	2010NE026214	
Paulo Mol Júnior	552736/11-2	—	—	DTI-A,36 m, 35

Fonte: DOU – 19/12/2011 – Seção 3 – Página 20

A metodologia descrita nesta tese somente recuperou o valor de R\$ 450.000,00 tendo ignorado o segundo valor. Ressalta-se aqui que os dados apresentados pelo DECIT constam o valor de R\$50.000,00 de responsabilidade do DECIT e R\$ 0,00 de responsabilidade do parceiro, neste caso CNPq. Cabe ressaltar ainda que na notícia deste financiamento, não consta o DECIT. Concluindo este caso, este poderia ser uma questão de coincidência de valores, o que significaria que o financiamento do DECIT não foi publicado no DOU. Ou poderia ser ainda uma falta do nome do DECIT na publicação da notícia de financiamento.

Foram verificados nas ferramentas de busca do FarejaDOU e do DOU no site da Imprensa Nacional e o nome da pesquisadora do caso 5 só aparece uma única vez no DOU do ano de 2009, cujo valor foi recuperado corretamente.

Em comparação com o G-FINDER, os valores são muitos distantes do que foi recuperado a partir do DOU. Uma vez que o Estado não unifica as fontes de dados sobre financiamento para pesquisa e que ter um olhar sumarizado é impossível, a aposta no Diário Oficial da União parece acertada.

Entretanto, trata-se de uma validação complexa de ser realizada a partir do momento em que os dados são desagregados e apresentam níveis de detalhamento diferente entre as diversas agências de fomento.

O atual formato do fomento à pesquisa no Brasil é plural, descentralizado e direcionado à pesquisa básica. Embora para Guimarães (2004), um grande número de agências de fomento seja um instrumento de proteção contra possíveis favorecimentos pessoais, possibilitando maior transparência do processo, a descentralização exige um mecanismo de coordenação adequado entre as múltiplas instâncias de fomento, o que parece ser uma lacuna no caso do Brasil. Este talvez seja uma das explicações para uma baixa capacidade de articulação entre as ações de fomento científico-tecnológico e a política de saúde. O investimento na pesquisa básica, por sua vez, parece não dar conta das necessidades de inovação em saúde no Brasil. É urgente que se atente para a pesquisa estratégica, com aplicação prática direta.

Os próprios nomes em que as instituições governamentais – agências de fomento, Ministério de Planejamento, Orçamento e Gestão etc – apresentam o financiamento refletem a visão que estes têm sobre os recursos financeiros disponibilizados para o desenvolvimento das pesquisas. Algumas apresentam como “**auxílio a pesquisadores**”, outras como “**despesas**”. O primeiro termo, baseado no senso comum, leva a acreditar que é um favor que as agências de fomento fazem ao pesquisador. Já o segundo termo, embora usual para os recursos que saem do caixa, apresenta no senso comum um recurso que não retorna de nenhuma forma. O financiamento a pesquisa deveria ser considerado como um “**investimento**”, termo que traria à mente a ideia de um resultado mensurável no qual se mediria o retorno sobre o valor investido.

Uma possível modificação de ordem mais concreta no modelo de financiamento com a finalidade de auxiliar em uma melhor distribuição dos recursos financeiros deveria visar a desvinculação do recurso financeiro a uma rubrica definida. Em outras palavras, os editais de fomento não deveriam ditar as regras de distribuição dos recursos financeiros entre os possíveis gastos: material permanente, material de consumo, serviço de terceiro.

O desenvolvimento desta pesquisa enfrentou desafios, mas também abre portas para novos estudos.

9.2. Limites e possibilidades

Este estudo se deparou com algumas questões que dificultaram a realização do mesmo e que podem ser trabalhadas mais profundamente em estudos futuros.

Um dos grandes desafios enfrentado neste estudo foi de caráter técnico. A disponibilização do Diário Oficial da União (DOU) em arquivos no formato PDF é um limitador frente ao processo de recuperação de dados no mesmo. O processo de conversão descrito neste estudo pode servir de base para futuros estudos na conversão de textos do formato PDF para texto (TXT). No entanto, a diagramação do DOU torna o processo extremamente complexo o que pode acarretar na perda de dados quando da tentativa de organizar o texto sequencialmente. Países como os Estados Unidos da América disponibilizam sua publicação oficial, o *Federal Register*, nos formatos *Extensible Markup Language* (XML) e texto (TXT) o que possibilita a recuperação dos dados de forma mais fácil. O movimento técnico-político na direção de organização e disponibilização do DOU nestes formatos seria de imensa valia para os estudos das políticas públicas.

Outra questão que serviu como delimitador deste estudo e que deve ser mais bem aprofundada é a identificação dos pesquisadores que estudam sobre uma determinada temática, no caso deste estudo a dengue. Dos 5644 pesquisadores identificados no CV Lattes, 4247 não foram identificados como tendo recebido algum tipo de financiamento pelas agências de fomento de nível federal; e 3343 não publicaram nenhum artigo indexado no Google Scholar. Apenas 775 pesquisadores do grupo inicial identificado receberam financiamento e possuem publicação sobre o tema no Google Scholar. Por se tratar de uma base declaratória que possui campos no formato aberto, o CV Lattes, embora se torne uma fonte de dados riquíssima em termos de volume de dados e de possibilidades, também se torna de difícil cruzamento entre os inúmeros currículos. Mesmo assim, partir do Currículo Lattes para identificação de pesquisadores de uma determinada temática não deixa de ser uma boa opção. Todavia, filtrar pelo termo na própria busca avançada do CV Lattes, aumenta o número de pesquisadores recuperados, mas diminui a precisão, pois se recupera todos os pesquisadores que possuem o termo – no caso deste estudo “dengue” – em seu CV, podendo este termo estar relacionado a uma palestra ou evento em que o pesquisador tenha participado.

Possíveis alternativas para aumentar a precisão na identificação de pesquisadores de uma determinada temática poderiam ser identificá-los (i) por meio dos Grupos de Pesquisa, (ii) pela produção científica contida em Repositórios Institucionais ou em bases como o Scielo ou ainda (iii) pela mineração do CV dos pesquisadores nos campos referentes a projetos de pesquisa. A primeira alternativa serviria de passo anterior à

busca dos currículos na Plataforma Lattes. O Diretório de Grupos de Pesquisa (DGP) disponível em <http://lattes.cnpq.br/web/dgp> serviria para a identificação de quem, de fato, pesquisa sobre a temática. Já para a segunda alternativa, os Repositórios Institucionais são de acesso livre e permitem a recuperação por título, palavras-chave, ou termos em qualquer parte do texto. Ainda no segundo item, para busca na base Scielo, é possível usar a própria ferramenta de busca, pelo termo em palavras-chave, título ou resumo por exemplo. Para os dois movimentos é importante automatizar o processo para a recuperação do nome dos pesquisadores. Neste processo, a dificuldade estaria em relacionar o nome do autor para a busca na Plataforma Lattes. Por fim, o terceiro caminho poderia ser usado em complemento ao descrito nesta tese. Em outras palavras, após um primeiro filtro pelo termo referente à temática e *download* dos currículos, utilizaria o mesmo termo para minerar os currículos no campo de projetos.

Dos 5644 pesquisadores identificados, foi encontrado um homônimo. Este limite, em pesquisas futuras deverá ser tratado caso a pesquisa tenha como categoria de análise o ente pesquisador. Este limite acarreta na possível alocação de verba para outro pesquisador, instituição, região demográfica.

As visualizações apresentadas por este estudo foram de grande valia para a análise dos dados. Porém, embora tragam certa interação com o usuário para escolha de intervalo de anos, por exemplo, já são pré-programadas e não permitem com que o usuário do mesmo façam suas próprias interações. Para tanto, seria importante que pesquisadores e desenvolvedores trabalhassem em conjunto para proporcionar uma plataforma mais interativa neste sentido.

Os suplementos do Diário Oficial da União, a saber: (i) Agência Nacional de Vigilância Sanitária (Anvisa), (ii) Plano Plurianual (PPA) e (iii) Orçamento da União, podem e devem ser utilizados como fonte de dados. As Seções 1 e 2 do Diário Oficial da União também podem vir a ser importantes fontes de dados para pesquisa. A Seção 2, por exemplo, traz notícias sobre o exercício de cargos comissionados, ou seja, quando um profissional passou a exercer um determinado cargo com, por exemplo, Direção e Assessoramento Superior (DAS) ou Função Gratificada (FG) ou ainda quando foi exonerado do mesmo. Este tipo de dado pode ser utilizado para filtrar o nome do pesquisador caso o mesmo esteja ocupando um cargo de gestão no período estudado.

Por fim, ressalta-se que técnicas de processamento de linguagem natural em conjunto com a análise visual dos dados tornam possível a recuperação e análise de dados do Diário Oficial da União e possibilita um aprofundamento em estudos das

Políticas Públicas no Brasil. A metodologia aqui desenvolvida e empregada é de possível reprodução e outros campos e temáticas de estudo podem ser beneficiados por ela.

REFERÊNCIAS

- ABBOTT, R. **Big Data and Pharmacovigilance: Using Health Information Exchanges to Revolutionize Drug Safety**. Rochester, NY: Social Science Research Network, 26 set. 2013. Disponível em: <<http://papers.ssrn.com/abstract=2246217>>. Acesso em: 4 out. 2013.
- About Visual Analytics**. Disponível em: <<https://www.viva-viva.ca/index.php/about/about-va>>. Acesso em: 21 jun. 2014.
- AÑEZ, G. et al. Economic impact of dengue and dengue hemorrhagic fever in the State of Zulia, Venezuela, 1997-2003. **Revista Panamericana de Salud Pública**, v. 19, n. 5, p. 314–320, maio 2006.
- ARANHA, C.; PASSOS, E. A Tecnologia de Mineração de Textos. **Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi:10.5329/RESI**, v. 5, n. 2, 2006.
- ARIELY, D. **Dan Ariely - Big data is like teenage sex: everyone talks... | Facebook**. Disponível em: <<https://www.facebook.com/dan.ariely/posts/904383595868>>. Acesso em: 15 jan. 2017.
- BARRETO, M. L.; TEIXEIRA, M. G. Dengue no Brasil: situação epidemiológica e contribuições para uma agenda de pesquisa. **Estudos Avançados**, v. 22, n. 64, p. 53–72, dez. 2008.
- BOKHARI, S. M. Z. **Understanding Amazon.com by the Numbers**. Disponível em: <<http://www.scdigest.com/assets/FIRSTTHOUGHTS/13-03-15.php?cid=6834>>. Acesso em: 19 jan. 2016.
- BRANCO, A.; SILVA, J. **Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese**. . In: PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC-2004). European Language Resources Association (ELRA), 2004Disponível em: <<http://aclasb.dfki.de/nlp/bib/L04-1354>>. Acesso em: 27 jul. 2015
- BRASIL. Portaria nº 1.347. . 2002 a.
- BRASIL. **Política nacional de ciência, tecnologia e inovação em saúde**. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/Politica_Portugues.pdf>. Acesso em: 16 jun. 2014.
- BRASIL. **Ciência, tecnologia e inovação em Saúde**. Brasília: Ministério da Saúde, 2008.
- BRASIL. **Manual técnico de orçamento MTO**. 2016. ed. Brasília: [s.n.].
- BRASIL. **CT-Saúde - Arquivos**, [s.d.]. Disponível em: <http://www.mct.gov.br/index.php/content/view/1417/CT___Saude.html#lista>
- BRASIL. **Painel Lattes**. Disponível em: <<http://estatico.cnpq.br/painelLattes/>>. Acesso em: 11 abr. 2016b.

- BRASIL. **Portal da Transparência**. Disponível em:
<<http://www.portaldatransparencia.gov.br/>>. Acesso em: 28 dez. 2016c.
- BRASIL, M. DA C. E T. E C. DE G. E E. E. **Documento Básico do Fundo Setorial da Saúde**, 2002b. Disponível em: <http://www.mct.gov.br/upd_blob/0017/17844.pdf>. Acesso em: 6 nov. 2014
- BRASIL. MINISTÉRIO DA SAÚDE. SECRETARIA DE CIÊNCIA, T. E I. E. D. DE C. E T. **Pesquisa para Saúde: Por que pesquisa em saúde?** Brasília: [s.n.].
- BURKE, M. A.; FRANCISCO, A. DE; GLOBAL FORUM FOR HEALTH RESEARCH (ORGANIZATION). **Monitoring financial flows for health research 2005: behind the global numbers**. Geneva: Global Forum for Health Research, 2006.
- BURLE, L. L. A Política de Comércio Exterior e a Abertura do Mercado de Capitais: 1990 - 92. **Revista de Administração Pública**, v. 27, n. 3, p. 98–114, 1993.
- CARMO, A. F. C. DO; SHIMABUKURO, M. H.; ALCÂNTARA, E. H. DE. AVALIAÇÃO DA QUALIDADE DE DADOS AMBIENTAIS POR MEIO DE TÉCNICAS DE ANALÍTICA VISUAL. **Boletim de Ciências Geodésicas**, v. 22, n. 3, p. 542–556, set. 2016.
- CHOPRA, A.; PRASHAR, A.; SAIN, R. Natural Language Processing. v. 1, p. 131–134, 2013.
- CNPQ. **Competências - Portal CNPq**. Disponível em:
<<http://www.cnpq.br/web/guest/competencias/>>. Acesso em: 8 jan. 2016a.
- CNPQ. **Demanda e Atendimento - Portal CNPq**. Disponível em:
<<http://cnpq.br/demanda-e-atendimento>>. Acesso em: 29 dez. 2016b.
- COLLIER, N. et al. BioCaster: detecting public health rumors with a Web-based text mining system. **Bioinformatics (Oxford, England)**, v. 24, n. 24, p. 2940–2941, 15 dez. 2008.
- CONDE, M. V. F.; ARAÚJO-JORGE, T. C. DE. Modelos e concepções de inovação: a transição de paradigmas, a reforma da C&T brasileira e as concepções de gestores de uma instituição pública de pesquisa em saúde. **Ciência & Saúde Coletiva**, v. 8, n. 3, p. 727–741, 2003.
- CORLEY, C. D. et al. Text and Structural Data Mining of Influenza Mentions in Web and Social Media. **International Journal of Environmental Research and Public Health**, v. 7, n. 2, p. 596–615, fev. 2010.
- COUTINHO-MARQUES, P. E. P. **FarejaDOU**. Aplicativo. Disponível em:
<<http://ctic009.icict.fiocruz.br:8000/dou>>. Acesso em: 16 dez. 2016.
- DAMASCENO, C. C. S. et al. **Inteligência em fontes abertas: um estudo sobre extração de informações do diário oficial da união**, 2011. Disponível em:
<http://intertwitter.googlecode.com/svn/trunk/documentacao/Monografia_final.doc>. Acesso em: 14 maio. 2015

DIEBOLD, F. X. **A Personal Perspective on the Origin(s) and Development of “Big Data”: The Phenomenon, the Term, and the Discipline, Second Version**. Rochester, NY: Social Science Research Network, 26 nov. 2012. Disponível em: <<http://papers.ssrn.com/abstract=2202843>>. Acesso em: 19 jan. 2016.

DING, Y.; ZHENG, Z.; RONG, M. Improved SPRINT Algorithm and its Application in the Physical Data Analysis. **TELKOMNIKA Indonesian Journal of Electrical Engineering**, v. 12, n. 9, p. 6909–6920, 1 set. 2014.

EDCON. **Making Sense of Big Data with Visual Analytics** Las Vegas, Nevada, USA, 2014. Disponível em: <<https://www.youtube.com/watch?v=aQolBdLEtq4>>. Acesso em: 13 jan. 2017

FAN, W.; BIFET, A. Mining Big Data: Current Status, and Forecast to the Future. **SIGKDD Explor. Newsl.**, v. 14, n. 2, p. 1–5, Abril 2013.

FEALING, K. H. **The science of science policy a handbook**. Stanford, Calif.: Stanford Business Books, 2011.

FILHO, C. et al. use of big data in healthcare in Brazil: perspectives for the near future. **Epidemiologia e Serviços de Saúde**, v. 24, n. 2, p. 325–332, jun. 2015.

FINEP. **Finep - CT-Saúde, o que é?**, 27 dez. 2012. Disponível em: <http://www.finep.gov.br/pagina.asp?pag=fundos_ctsaude>. Acesso em: 11 jun. 2014

FINEP. **Sobre a Finep**. Disponível em: <<http://www.finep.gov.br/a-finep-externo/sobre-a-finep>>. Acesso em: 8 jan. 2016.

FIOCRUZ. **Rede Dengue, Zika e Chikungunya**. Rede Dengue, Zika e Chikungunya. Disponível em: <<http://rededengue.fiocruz.br/>>. Acesso em: 20 dez. 2016.

FUNASA. **Programa Nacional de Controle da Dengue: amparo legal à execução das ações de campo - imóveis fechados, abandonados ou com acesso não permitido pelo morador**. 2. ed. Brasília: [s.n.].

GADELHA, C. A. G. O complexo industrial da saúde e a necessidade de um enfoque dinâmico na economia da saúde. **Ciência & Saúde Coletiva**, v. 8, n. 2, p. 521–535, 2003.

GARTNER, INC. **What Is Big Data? - Gartner IT Glossary - Big Data**. Disponível em: <<http://www.gartner.com/it-glossary/big-data>>. Acesso em: 19 jan. 2016.

GU, B. **Recognizing named entities in biomedical texts**. Thesis—[s.l.] School of Computing Science - Simon Fraser University, 2008.

GUIMARÃES, M. C. S. **Tecnologia como conhecimento: o público e o privado; o social e o econômico**. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 1998. Tese de doutorado

GUIMARÃES, R. Bases para uma política nacional de ciência, tecnologia e inovação em saúde. **Ciência & Saúde Coletiva**, v. 9, n. 2, p. 375–387, jun. 2004.

- GUIMARÃES, R. Pesquisa em saúde no Brasil: contexto e desafios. **Revista de Saúde Pública**, v. 40, n. spe, p. 3–10, ago. 2006.
- GUIMARÃES, R. F. N. Sair da resistência e partir para novas competências. **Ciência & Saúde Coletiva**, v. 15, n. 4, p. 1913–1915, jul. 2010.
- GUIMARÃES, R.; SERRUYA, S. J.; DIAFÉRIA, A. O Ministério da Saúde e a Pesquisa em Saúde no Brasil. **Gazeta Médica da Bahia**, v. 78, n. 1, 10 jul. 2008.
- HALASA, Y. A.; SHEPARD, D. S.; ZENG, W. Economic Cost of Dengue in Puerto Rico. **The American Journal of Tropical Medicine and Hygiene**, v. 86, n. 5, p. 745–752, 1 maio 2012.
- HAY, S. I. et al. Big Data Opportunities for Global Infectious Disease Surveillance. **PLoS Medicine**, v. 10, n. 4, abr. 2013.
- KEIM, D. **Mastering the information age: solving problems with visual analytics**. Goslar: Eurographics Association, 2010.
- LANE, J. Assessing the Impact of Science Funding. **Science Innovation**, v. 324, n. 5932, p. 1273–1275, 2009.
- LANE, J. Let's make science metrics more scientific. **Nature**, v. 464, n. 7288, p. 488–489, 25 mar. 2010.
- LANEY, D. **3D Data Management: Controlling Data Volume, Velocity, and Variety**. [s.l.] META Group, fev. 2001. Disponível em: <<http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>>.
- LIDDY, E. D. **Natural Language Processing**. iSchool Faculty Scholarship. 2001.
- LOPES, I. L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ciência da Informação**, v. 31, n. 1, 16 abr. 2002.
- MARGARIDA BASILIO. **Formação e classes de palavras no português do Brasil**. 3. ed. São Paulo: Editora Contexto, 2011.
- MENA-CHALCO, J. P.; CESAR JUNIOR, R. M. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, dez. 2009.
- MOLNAR, A. et al. Using Visualization to Derive Insights from Research Funding Portfolios. **IEEE Computer Graphics and Applications**, v. 35, n. 3, p. 91-c3, maio 2015.
- MOREL, C. M. et al. Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases. **PLoS Negl Trop Dis**, v. 3, n. 8, p. e501, Agosto 2009.

- OLIVEIRA, A. DE; BIANCHETTI, L. CNPq: política de fomento à pesquisa nos governos Fernando Henrique Cardoso (FHC). **Perspectiva**, v. 24, n. 1, p. 161–182, 2006.
- OLSON, S. et al. (EDS.). **Measuring the impacts of federal investments in research: a workshop summary**. Washington, D.C: National Academies Press, 2011.
- PANG, T. et al. Knowledge for better health: a conceptual framework and foundation for health research systems. **Bulletin of the World Health Organization**, v. 81, n. 11, p. 815–820, 2003.
- POLICY CURES. **Measuring global health R&D for the post-2015**. Australia: Policy Cures, 2015. Disponível em: <<http://polycures.org/downloads/Post%202015%20indicators%20report%20WEB.pdf>>. Acesso em: 18 ago. 2016.
- POLICY CURES; BILL & MELLINDA GATES FOUNDATION. **G-Finder**. Disponível em: <<https://gfinder.polycures.org/PublicSearchTool/searchDisease>>. Acesso em: 28 dez. 2016.
- RAMIREZ-JIMENEZ, R. et al. Risks of Dengue Secondary Infection Associated with *Aedes aegypti* in Home Environments in Monterrey, Mexico. **Southwestern Entomologist**, v. 38, n. 1, p. 99–108, mar. 2013.
- RUIVO, B. “Phases” or “paradigms” of science policy? **Science and Public Policy**, v. 21, n. 3, p. 157–164, 1 jun. 1994.
- SACRIPANTI, A. Judo match analysis, a powerful coaching tool, basic and advanced tools. **arXiv:1308.0457 [physics]**, 2 ago. 2013.
- SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**. New York: McGraw-Hill, 1983.
- SELIGMAN, S. J. Constancy and diversity in the flavivirus fusion peptide. **Virology Journal**, v. 5, n. 1, p. 27, 2008.
- STOLTE, C.; TANG, D.; HANRAHAN, P. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. **IEEE Transactions on Visualization and Computer Graphics**, v. 8, n. 1, p. 52–65, mar. 2002.
- SUAYA, J. A. et al. Cost of dengue cases in eight countries in the Americas and Asia: a prospective study. **The American journal of tropical medicine and hygiene**, v. 80, n. 5, p. 846–855, maio 2009.
- TALIBERTI, H.; ZUCCHI, P. Custos diretos do programa de prevenção e controle da dengue no Município de São Paulo em 2005. **Revista Panamericana de Salud Pública**, v. 27, n. 3, p. 175–180, mar. 2010.
- TERRY, R. F. et al. Creating a global observatory for health R&D. **Science**, v. 345, n. 6202, p. 1302–1304, 12 set. 2014.

TIGRE, P. B. **Gestão da Inovação. A economia da tecnologia no Brasil**. 1. ed. Rio de Janeiro: Campus, 2006.

USA. **Transparency and Open Government | The White House**. Disponível em: <https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment>. Acesso em: 27 jul. 2016.

VALDEZ, B.; LANE, J. The science of science policy: a federal research roadmap. **Policy, O. o. S. a. T. Washington, DC, Executive Office of the President**, 2008.

VIERGEVER, R. F.; HENDRIKS, T. C. C. The 10 largest public and philanthropic funders of health research in the world: what they fund and how they distribute their funds. **Health Research Policy and Systems**, v. 14, n. 1, dez. 2016.

VIVA. **Visual Analytics is the New Orange** Vancouver Institute for Visual Analytics, , 2015. Disponível em: <<https://www.viva-viva.ca/images/VIVA/Documents/VIVA-Whitepaper-NewOrange-083115.pdf>>. Acesso em: 17 jul. 2016

WANG, H.; WENBO, Q.; SHEN, Q. Table tennis video data mining based on performance optimization of artificial fish swarm algorithm. **Computer Modeling and New Technologies**, v. 18, n. 12A, p. 584–588, 2014.

WANG, L.; LIU, W. Online publishing via pdf2html EX. **TUG Boat - COMMUNICATIONS OF THE T E X USERS GROUP**, v. 34, n. 3, p. 372, 23 out. 2013.

WHO. **Dengue and severe dengue**, 2014. Disponível em: <<http://www.who.int/mediacentre/factsheets/fs117/en/>>. Acesso em: 22 maio. 2014

YAU, N. **Jobs Charted by State and Salary** **FlowingData**, 2 jul. 2014. Disponível em: <<http://flowingdata.com/2014/07/02/jobs-charted-by-state-and-salary/>>. Acesso em: 13 jan. 2017

YOUNG, A. J. et al. Global trends in health research and development expenditures – the challenge of making reliable estimates for international comparison. **Health Research Policy and Systems**, v. 13, n. 1, p. 7, 2015.

ANEXO 1- Texto original para submissão do CV no Canadian Common CV

The Canadian Institutes of Health Research (CIHR) is committed to respecting the privacy rights of individuals in accordance with the provisions of the Privacy Act and the policies of the government of Canada. If you decide to proceed and register with the Common Curriculum Vitae system (CCV), which is administered by CIHR, and provide certain personal information about you, CIHR will keep and use the information for the purposes that you have authorized.

CIHR is not liable for data, including username and password, that are provided to and stored by third parties, which are then used to update the CCV system.

All data that you enter in the CCV application is readable, but not writable, by the support team of each CCV subscribing institution. This is required in order to offer proper assistance when requested.

Each time you submit your CV, you will be asked whether you wish to have CIHR submit your personal information to other institutions as part of your application for funding. If you do, you will be consenting to CIHR making your personal information available to the designated institutions for them to use in evaluating your application. Once the information is successfully forwarded to them, the designated institutions will be responsible for the management and protection of the information that they receive, in accordance with federal and provincial privacy laws.

When submitting your CV for the first time, you will be asked whether you wish to include some of your personal information in provincial and national researcher directories available to the public. More information will be provided at that time when your consent is requested. You will also be able to grant or revoke this consent using the 'Consent' tab at any time.

Finally, from time to time institutions conduct searches to identify candidates for peer review, focus groups or surveys and you will be asked whether you wish to be included in this database that Institutions will be able to search against certain criteria. Here again you will be given more information when asked whether you wish to be included in the database and you will also be able to grant or revoke this consent using the 'Consent' tab at any time.

Any personal information collected by CIHR is protected from disclosure in accordance with the Privacy Act and the policies of the government of Canada. For inquiries concerning treatment of personal information, you may contact the Name of

the Program, email address at CIHR, or Individuals may contact CIHR's Access to Information Coordinator (located at 160 Elgin Street, Room 97, Address Locator 4809A, Ottawa, ON, K1A 0W9, Canada / Phone: 613-948-2284), for access to their personal information pursuant to the provision of the Privacy and Access to Information Acts. Please note that once CIHR provides your personal information to an institution, CIHR can no longer control the use that that institution makes of the information and is not responsible for any unauthorized or improper use by the institution. In such a case you will have to deal directly with the institution.

The information collected for the Common CV includes:

Contact information, academic information, credentials and recognitions, languages competency, research classification information, research interest and experience, professional and volunteer activities, contributions.

I have read and understood the above statement and I hereby consent to the collection and storage of the information that I provide CIHR, and to its use for the purposes that I may authorize.

ANEXO 2- Dicionário de dados das tabelas do banco de dados

Table "public.agencias"

Column	Type	Description
idx	integer	índice interno
agencia	character varying(512)	agência de fomento
sigla	character varying(512)	sigla da agência de f.
codigo	integer	código da agência de f.

Indexes:

"agencias_pkey" PRIMARY KEY, btree (sigla)

Exemplo de conteúdo:

idx		3
agencia		Financiadora de Estudos e Projetos
sigla		FINEP
codigo		4

Table "public.area"

Column	Type	Description
id	integer	índice interno
id16	character varying(512)	id lattes do pesquisador
grande	character varying(512)	grande área
area	character varying(512)	área
sub	character varying(512)	subárea

Indexes:

"area_pkey" PRIMARY KEY, btree (id)

Exemplo de conteúdo:

id		6
id16		8613000100699719
grande		Ciências da Saúde
area		Medicina
sub		Patologia Clinica

Table "public.award"		
Column	Type	Description
id	integer	indice interno
agencias	integer	codigo das agencias
data	date	data de publicação do financiamento
id16	character varying(512)	id lattes pesquisador
valor	double precision	valor financiado
ano	integer	ano de publicação do financiamento

Indexes:

"award1_pkey" PRIMARY KEY, btree (id)

Exemplo de conteúdo

id	4
agencias	1
data	2012-06-28
id16	1807193235853557
valor	85000
ano	2012

Table "public.coauthorship"		
Column	Type	Description
id	integer	indice interno
idx	integer	indice interno
id16	character varying(512)	id lattes do pesquisador
id10	character varying(512)	id lattes do pesquisador
id16coauthor	character varying(512)	id lattes do pesquisador
ano	integer	ano da parceria

Indexes:

"coauthorship_pkey" PRIMARY KEY, btree (id)

Exemplo de conteúdo:

id	5
idx	5
id16	8613000100699719
id10	K4282472U9
id16coauthor	7944807402012600
ano	2005

Table "public.estado"		
Column	Type	Description
idx	integer	indice interno
estado	character varying(512)	Sigla do Estado
id_regiao	character varying(512)	Região do país

Indexes:

"estado_pkey" PRIMARY KEY, btree (idx)

Exemplo de conteúdo:

idx	5
estado	BA
id_regiao	Nordeste

Table "public.instituicao"		
Column	Type	Description
idx	integer	indice interno
instituicao	character varying(512)	instituição de ensino/pesquisa

Indexes:

"instituicao_pkey" PRIMARY KEY, btree (idx)

Exemplo de conteúdo:

idx	412
instituicao	Fundação Oswaldo Cruz

Table "public.pesquisador"		
Column	Type	Description
idx	integer	indice interno
id10	character varying(512)	id lattes do pesquisador
id16	character varying(512)	id lattes do pesquisador
nome	character varying(512)	nome do pesquisador
modificado	date	data da última modificação do cv
id_instituicao	integer	id instituição

Indexes:

"pesquisador_pkey" PRIMARY KEY, btree (idx)

Exemplo de conteúdo:

```
idx          | 5
id10         | K4104428T7
id16         | 6343168720295275
nome         | Eliane de Alcântara Teixeira
modificado  | 2016-04-25
id_instituicao | 369
```

Table "public.regiao"

Column	Type	Description
idx	integer	índice interno
regiao	character varying(512)	região do país

Indexes:

"regiao_pkey" PRIMARY KEY, btree (idx)

Exemplo de conteúdo:

```
idx | 4
regiao | Sudeste
```

Table "public.rinstest"

Column	Type	Description
id	integer	índice interno
id_instituicao	integer	id da instituição
id_estado	integer	id do estado

Indexes:

"rinstest_pkey" PRIMARY KEY, btree (id)

Exemplo de conteúdo:

```
id          | 3
id_instituicao | 3
id_estado   | 19
```


Table "public.rpesqest"

Column	Type	Description
id	integer	índice interno
id_pesquisador	integer	id interno do pesquisador
id_estado	integer	id interno do estado

Indexes:

"rpesqest_pkey" PRIMARY KEY, btree (id)

Exemplo de conteúdo:

```
id          | 3
id_pesquisador | 3
id_estado   | 26
```

Table "public.scholar"

Column	Type	Description
id	integer	índice interno
id16	character varying(512)	id lattes do pesquisador
qtd	integer	quantidade de artigos publicados
ano	integer	ano de publicação

Indexes:

"scholar_pkey" PRIMARY KEY, btree (id)

Exemplo de conteúdo:

```
id      | 5
id16    | 1807193235853557
qtd     | 2
ano     | 2012
```