**UNIVERSIDADE DE BRASÍLIA**
**FACULDADE DE CIÊNCIA DA INFORMAÇÃO**

# DIRETRIZES PARA UMA POLÍTICA DE GESTÃO DE DADOS CIENTÍFICOS NO BRASIL

MAÍRA MURRIETA COSTA

Brasília
2017

**UNIVERSIDADE DE BRASÍLIA**

**FACULDADE DE CIÊNCIA DA INFORMAÇÃO**

# DIRETRIZES PARA UMA POLÍTICA DE GESTÃO DE DADOS CIENTÍFICOS NO BRASIL

MAÍRA MURRIETA COSTA

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília como requisito para obtenção do Doutorado em Ciência da Informação.

Orientador: Prof. Dr. Murilo Bastos da Cunha

**Área de concentração:** Organização da Informação.

**Linha de pesquisa:** Biblioteca digital.

Brasília

2017

Ficha Catalográfica

C837d Costa, Maíra Murrieta.
Diretrizes para uma política de gestão de dados científicos no Brasil / Maíra Murrieta Costa; orientador Murilo Bastos da Cunha. – Brasília: 2017.
288 f. : il

Tese (Doutorado em Ciência da Informação) – Universidade de Brasília. Faculdade de Ciência da Informação.
Possui figuras, quadros, gráficos, tabelas e referências.

1. Big data. 2. E-science. 3. Data deluge. 4. Cyberinfraestructure. 5. Dados científicos. 6. Dados de pesquisa. 7. Gestão de dados científicos. 8. Política de informação. 9. Biblioteca digital.
I. Cunha, Murilo Bastos da. II. Título.

## FOLHA DE APROVAÇÃO

**Título:** "Diretrizes para uma política de gestão de dados científicos no Brasil"

**Autor (a):** Maíra Murrieta Costa
**Área de concentração:** Gestão da Informação
**Linha de pesquisa:** Comunicação e Mediação da Informação

Tese submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Doutor** em Ciência da Informação.

Tese aprovada em: 18 de agosto de 2017.

**Profº Drº Murilo Bastos da Cunha**
Presidente (UnB/PPGCINF)

**Profº Drº Rubens de Oliveira Martins**
Membro Externo (MPOG)

**Profª Drª Luana Farias Sales**
Membro Externo (CNEN - RJ)

**Profº Drº Jayme Leiro Vilan Filho**
Membro Interno (UnB/PPGCINF)

**Profº Drº Fernando William Cruz**
Suplente (UnB/PPGCINF)

À minha família que me
acompanhou durante essa trajetória.

## AGRADECIMENTOS

O aluno de doutorado, para concluir sua tese, precisa, antes de mais nada, do apoio de seu companheiro e de seus familiares, pois, a trajetória é árdua e muitas vezes solitária. Por isso deixo meu agradecimento mais do que especial ao Térzio Roberto, meu marido e companheiro nesse longo trajeto, ele não me deixou desistir. Minha outra companheira incansável nessa jornada sem dúvida nenhuma foi minha mãe, tanto pelo apoio, encorajamento como pelas inúmeras vezes que se dispôs a ler o texto e corrigir o necessário.

A minha família é parte integrante dessa história. Deixo aqui meus sinceros agradecimentos ao meu pai pelo apoio e subsídios políticos, aos meus sogros, às minhas irmãs, à minha avó, à minha linda sobrinha e ao meu cunhado Guilherme Araújo.

Às minhas grandes amigas durante o doutorado, Sônia Boeres, Thais Araújo, Cacilda Pereira. Ao meu amigo José Antônio Machado do Nascimento. Todos estiverem presente ao longo desses quatro anos e meio. Me incentivaram e me ajudaram em tudo que foi possível.

Às outras grandes amigas, que me acompanharam de longe, mas deixaram sua marca nesses quatro anos, Monica Bezerra Alves e Mara Gomes. Aos meus colegas do MCTI – Giancarlo Muraro, Sérgio Knorr Velho, Hideraldo Almeida, Edilson Pedro, Fernanda Magalhães.

Não posso deixar de agradecer aos poucos mais preciosos amigos que fiz em Ann Arbor – Michigan – meu orientador Victor Rosenberg, além de grande professor, uma pessoa de coração inestimável, John Brinks, Peter Aetema e ainda à família Alen que me acolheu e tornou suportável a distância do Brasil.

Algumas pessoas tiveram um valor inestimável na reta final, durante a análise dos dados. São elas – Márcio Medeiros, colega de doutorado da Sociologia que tanto me ajudou com o Nvivo. Rejane Miranda e Simone Cerveira pelo auxílio no uso do SPSS, mas, mais do que isso, porque a experiência delas me propiciou ver situações que eu provavelmente não enxergaria sozinha. Ao Alan Freitas por todo apoio na criação e vetorização de imagens. Agradeço também à Luciana Oliveira e ao Leonardo Silva Oliveira consultores de normalização deste documento.

Também agradeço aos professores da Faculdade de Ciência da Informação da Universidade de Brasília com os quais eu tive a oportunidade de aprender e amadurecer a complexidade da ciência da informação durante as aulas, dentre eles, Jayme Leiro pela oportunidade única de aprofundar meus conhecimentos na Bibliometria.

Ao meu grande orientador, incansável tutor, agradeço por acreditar em mim, por me proporcionar a experiência do doutorado sanduíche, por ter tido toda a paciência nos momentos em que eu não conseguia produzir o necessário.

À CAPES que tornou possível a participação no doutorado sanduíche em um programa de primeiro mundo, a School of Information – University of Michigan.

Também agradeço a querida Martinha Araújo e à Vivian Miatelo, sempre sorridentes e dispostas a me ajudar com os procedimentos burocráticos do doutorado.

## RESUMO

Contextualiza a organização social da ciência contemporânea. Discute aspectos sobre a ciência colaborativa do Século XXI, a internacionalização e a virtualização da ciência que culminaram com a explosão de dados científicos coletados *on-line*, dando origem ao fenômeno de *big data* e *e-science*. Discute o surgimento dos termos guarda-chuva *e-science* e *cyberinfrastructure.* Contextualiza a literatura sobre dados de pesquisa/ dados científicos e argumenta sobre a necessidade da estruturação de políticas públicas que norteiem a gestão dos dados científicos oriundos da *e-science*, visto que, do ponto da gestão da informação, faz-se necessário apontar soluções para um tratamento adequado dos dados científicos de forma a viabilizar o processo de armazenamento, organização, busca, recuperação e difusão dos dados coletados. Nesse aspecto, relembra que a preocupação da informação científica está na origem da ciência da informação, logo discute o papel do profissional da informação no tratamento dos dados oriundos da *e-science*. Também são contextualizados na revisão de literatura a política brasileira de ciência e tecnologia, bem como a política de informação. O estudo apresenta como objetivo geral – *é elaborar um esboço de diretrizes governamentais para a gestão de dados científicos no Brasil.* Para tanto, desenharam-se os seguintes objetivos específicos: OE 1) Identificar os países desenvolvidos que possuem ações de governo para a gestão de dados científicos, OE 2) Analisar as ações de governo de países desenvolvidos sobre a gestão de dados científicos nos países identificados, OE 3) Identificar os principais problemas e as soluções inerentes à construção de uma política estruturada para a gestão de dados científicos, OE 4) Identificar a postura das agências de fomento no Brasil com relação ao tema, OE 5) Identificar o posicionamento dos pesquisadores brasileiros envolvidos com o tema. Discursa sobre a primeira fase da pesquisa, marcada pela necessidade de nortear a busca bibliográfica de forma a identificar as políticas nacionais de gestão de dados científicos nos países mais avançados em e-science. Informa que para tanto foi realizada uma pesquisa descritiva e de levantamento (*survey*), que utilizou a bibliometria, um método quantitativo baseado em análises estatísticas, para análise de dados. O estudo analisou o termo *e-science* nas bases de dados Library Information Science Abstracts e Library Information, Science Technology Abstracts. Discorre sobre a metodologia da pesquisa, classificando-a como exploratória, com características quantitativas e qualitativas na coleta de dados, portanto uma pesquisa mista. Informa que a amostra da pesquisa é não probabilística, formada pelo critério de intencionalidade. Foram entrevistados 40 pesquisadores doutores envolvidos com a gestão de dados científicos no Brasil. Além deles, foi aplicado questionário a 22 servidores de agências de fomento e fundações de amparo à pesquisa no Brasil. No aspecto qualitativo da análise dos dados, tece considerações sobre a abordagem de investigação utilizada – a *Grounded Theory*. O estudo argumenta que o Brasil carece de uma política explícita que norteie as ações do Estado em termos de gestão e preservação dos dados científicos, bem como diretrizes para reutilização dos dados em questão. Em termos de iniciativas nacionais, tem-se apenas a referente a informação geoespacial já estabelecida em Decreto 6.666 de 2008. Além desta, a outra iniciativa de destaque é a informação sobre biodiversidade trabalhada no Portal Brasileiro da Biodiversidade. Também apresenta o edital da FAPESP em *e-science* como iniciativa relevante na área. Apresenta um *framework* com itens considerados de extrema relevância para a elaboração de um conjunto de diretrizes que venham a servir de elementos norteadores para a elaboração de uma política para a gestão de dados científicos no Brasil. Conclui que uma política de gestão de dados precisa abordar aspectos tais como: regras de compartilhamento e reuso dos dados, prazo de carência para algumas categorias de dados, prazo de armazenamento para algumas classes de dados, padrões de metadados e interoperabilidade destes. Além disso, deve exigir do pesquisador um plano de gestão de dados quando a pesquisa for fomentada pelo governo, bem como definir os requisitos necessários para a implementação do DOI para dados a exemplo das questões relacionadas no *framework.*



**Palavras-Chave:** *Big data.* Biblioteca digital. *Data deluge. E-science. Cyberinfrastructure.* Dados científicos. Dados de pesquisa. Gestão de dados científicos.

## ABSTRACT

It contextualizes the social organization of contemporary science. It discusses aspects of the collaborative science in the 21st century, the internationalization and virtualization of science that culminated in the explosion of scientific data online collected, giving rise to the phenomenon of big data and e-science. It discusses the emergence of the terms umbrella e-science and cyberinfrastructure. It contextualizes the literature on research data / scientific data and argues about the need for structuring public policies to guide the management of scientific data from e-science, as, from the point of information management, it is necessary to point out solutions for an adequate treatment of scientific data, to enable the process of storing, organizing, search, retrieving and dissemination of collected data on a large scale. In this aspect, it recalls that the concern of scientific information is at the origin of information science, thus, it discusses the role of the information professional in the treatment of data from e-science. The Brazilian science and technology policy, as well as the information policy are also contextualized in the literature review. The study presents, as general objective – Elaborate a draft of guidelines for the management of scientific data in Brazil. To do so, the following specific objectives were designed: OE 1) Identify the developed countries that have government actions for the management of scientific data; OE 2) Analyze government actions of developed countries on the management of scientific data in the identified countries; OE 3) Identify the main problems and the solutions inherent to the construction of a structured policy for the management of scientific data; OE 4) Identify the position of funding agencies in Brazil concerning this subject; OE 5) Identify the opinion of Brazilian researchers involved with the subject and. The thesis discusses the first phase of the research; characterized by the need to guide the bibliographic search to identify the national policies of scientific data management in countries that e-science is more advanced. It reports that a descriptive and survey research was carried out, using bibliometrics, a quantitative method based on statistical analysis, for data analysis. The study analyzed the term e-science in the following database: Library Information and Science Abstracts e Library Information, Science Technology Abstracts. It discusses the methodology of the research, classifying it as exploratory, with quantitative and qualitative characteristics in the data collection, as a result, a mixed research. It informs that the research sample is non-probabilistic, formed by the criterion of intentionality. Forty PhD researchers involved in the management of scientific data in Brazil were interviewed. Additionally, twenty-two federal employees working in the funding agencies answered a questionnaire. In the qualitative aspect of the data analysis, it makes considerations on the research approach used – a Grounded Theory. The study argues that Brazil lacks an explicit policy to guide State actions in terms of management and preservation of scientific data, as well as guidelines for reusing the data. In terms of national initiatives, there is only reference to geospatial information already established in 2008 by the Decree n. 6.666. In addition to this, the other outstanding initiative is the information on biodiversity worked on the Brazilian Biodiversity Portal. It also presents the FAPESP e-science call for proposals as a relevant initiative in this field. It come up with a framework of items considered to be of extremely relevance for the elaboration of a set of guidelines that will serve as guiding elements for the elaboration of a policy for the management of scientific data in Brazil. I concluded that a data management policy needs to address aspects such as rule of data sharing and its reuse, grace period for some data categories, shelf life for some classes of data, metadata standards, and interoperability for them. Moreover, a data management plan must me demanded from the researcher when the research is fostered by the government, as well as to define the necessary requirements for implementation of DOI for data, such as the issues related to the framework.

**Keywords**: Big data. Digital library. Data deluge. E-science. Cyberinfrastructure. Scientific data. Research data. Scientific data management.

## RÉSUMÉ

Met en perspective l'organisation sociale de la science contemporaine. Aborde les aspects sur la science de collaborative du XXIe siècle, l'internationalisation et la virtualisation de la science qui abutirent à l'explosion de données scientifiques ramassées en ligne, ce qui a donné lieu au phénomène de big-data et e-science. Débat sur l'apparition des termes génériques e-science et cyberinfrastructure. Situe la littérature sur les données de recherche/données scientifiques et soutien le besoin d'une structuration des politiques publiques pour guider la gestion des données scientifiques résultantes du e-science, vue que, pour la gestion de l'information, il faut proposer des solutions pour un traitement propre des données scientifiques pour le processus de stockage, organisation, recherche, récuperation et dissémination des données ramassées, en grande échelle. Sur ce point, rappele que le soucis sur l'information scientifique est à l'origine de la science de l'information, donc aborde le rôle du professionnel de l'information pour traiter les données résultantes du e-science. La politique brésilienne pour la science et technologia et celle pour l'information sont aussi mises en perspective dans la revision de la littérature. L'étude présente pour objectif général – concevoir un ébauche de directives pour la gestion des données scientifiques au Brésil. Pour l'atteindre, les objectifs spécifiques (OS) suivants ont été conçus: OS1) Identifier les pays développés qui ont de politiques pour la gestion des données scientifiques, OS2) Examiner politiques pour la gestion des données scientifiques des pays développés, OS3) Identifier les principaux problèmes et les solutions liés à la construcion d'un politique structurée pour la gestion des données scientifiques, OS4) Identifier l'attitude des agences de promotion du Brésil sur la tématique, OS5) Identifier le positionnement de chercheur brésiliens intéressés par la tématique et. Aborde la première phase des travaux, marqué par le besoin de guider la recherche bibliographique pour identifier les politiques nationales de gestion de données scientifiques dans les pays les plus avancés en e-science. Rapporte que pour cela, una recherche descriptive et d'enquête (survey) a été ménée, que la bibliométrie, une méthode quantitative basée sur des analyses statistiques, a été utilisée pour analyser les données. L'étude a examiné le terme e-science dans les bases de données de la Library Information Science Abstracts et Library Information, Science Technology Abstracts. Parle de la méthodologie de la récherche et la classe comme exploratoire, avec des aspects quantitatifs e qualitatifs pour le ramassage des données, par conséquent une recherche mixte. Fait savoir que l'échantillonage de la recherche est non-probabilistique, mais crée par le critère d'intention.40 chercheurs docteurs impliqué par la gestion des données scientifiques au Brésil ont été interviewé. En outre, un questionnaire a été employé a 22 fonctionnaires d'agences de promotion et fondations d'appui à la recherche du Brésil. Pour l'aspect qualitatif de l'analyse des données, expose des considérations sur l'approche de l'investigation utilisée - la Groudend Theoty. L'étude argumente que le Brésil est dépourvu d'une politique claire pour guider les actions de l'État pour la gestion e conservation des données scientifiques, ainsi que des directives pour la réutilisation de ces données. Quant aux initiatives nationales, il y a seulemente celle sur l'information géospatial établi par le Décret 6.666 de 2008. S'ajoute a cette initiative celle de l'information sur la biodiversité définie au Portal Brasileiro da Biodiversidade. L'avis public de la FAPESP sur e-science est aussi utilisée comme initiative importante dans le domaine. Présente un framework avec des élements considérés d'extrème importancepour l'élaboration d'un ensemble de directives que puissent guider l'élaboration d'une politique pour la gestion de données scientifiques au Brésil.

Conclut q'il faut qu'une politique de gestion de données considère certains aspects comme: des normatifs de partage et réutilisation de données, un délai de carence pour quelques catégories de données, une périuode de stockage pour quelques rangs de données, des standards de métadonnées et son interopérabilité. En plus, il faut démander au chercheur un plan de gestion des données pour les recherces promue par le gouvernement, ainsi que la définition de

exigences pour le déploiment du DOI pour les données, à l'exemple des questions sur le framework.

**Mots-Clés**: Big data. Bibliothèque digitale. Data deluge. E-science. Cyberinfrastructure. Données scientifiques. Données de recherché. Gestion de données scientifiques.

## LISTA DE FIGURAS

## LISTA DE GRÁFICOS

## LISTA DE QUADROS

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

A evolução da ciência está altamente relacionada com o aprimoramento do instrumental tecnológico que permitiu a realização de observações de fenômenos em geral. Para Bell (2011), as teorias científicas do Século XX foram baseadas em dados geralmente disponíveis em cadernos científicos pessoais. Já no início do Século XXI, emergiu, de forma crescente, uma questão: os dados oriundos de pesquisas que são coletados por meio de sensores especializados, telescópios, satélites e ensaios de laboratórios. Há autores, como, por exemplo, Green (2011), Fox e Hendler (2011), que destacam a transformação pela qual passará a pesquisa científica em razão da criação e disponibilidade de um grande volume de dados *online*, bem como em razão da possibilidade de colaboração entre cientistas em diferentes partes do mundo.

A pesquisa colaborativa presente no Século XXI é descrita como aquela que tem a "capacidade de gerar e armazenar dados em uma escala sem precedentes e muito além da capacidade humana de análise" (CÉSAR JÚNIOR, 2011). O volume de dados, de toda a ordem e origem, em escala sem precedente, foi denominado *big data.* No âmbito cientifico surge o termo *e-science* que se refere a uma nova forma de fazer ciência, cuja principal característica é a produção de um grande volume de dados que precisa estar *online* para facilitar a colaboração entre os pesquisadores. Como veremos mais à frente (na revisão de literatura, capítulo 2.3), o último termo, no que tange à sua conceituação ao longo do tempo, parece estar se alterando, ou originando termos mais específicos tais como dados científicos ou de pesquisa.

*Big data* é um termo mais amplo, refere-se a um grande volume de dados e ao conjunto de soluções tecnológicas para tratar esses dados digitais. Relaciona-se com a percepção e compreensão de informações analisadas em grande escala, utilizadas geralmente em aplicações comerciais (como, por exemplo, na Amazon para sugerir qual livro o usuário deve comprar), na prospecção de cenários futuros, em campanhas publicitárias, em campanhas de eleição, dentre outros. Para Mayer-Schönberger e Cukier (2013), o *big data* representa "uma nova fonte de valor econômico e informação". A filosofia do *big data* é *deixe os dados falarem*.

Dentre os exemplos sobre análise de *big data*, destaca-se o artigo publicado, em fevereiro de 2009, pela empresa Google, na revista *Nature* (GINSBERG *et al*., 2009), relatando a análise de 50 milhões dos termos mais pesquisados na plataforma de pesquisa e a sua posterior comparação com os dados do Center for Disease Control (CDC), no período de 2003 a 2008. A análise da Google mostrou ser mais eficaz que as estatísticas do governo americano sobre a disseminação da gripe H1N1 em 2009, pois revelou onde o vírus estava se espalhando com

mais velocidade que o sistema de informações do CDC (BOLLIER, 2010; MAYER-SCHÖNBERGER; CUKIER, 2013).

De acordo com Mayer-Schönberger e Cukier (2013, p. 2) o método utilizado pela Google não se fundamenta no contato com os médicos, ou mesmo na coleta de saliva, "ele se baseia em *big data* – a capacidade de uma sociedade de obter informações de maneira nova a fim de gerar ideias úteis e bens e serviço de valor significativo".

A *e-science*, por sua vez, também é retratada no âmbito da produção de um grande volume de dados e da necessidade de avanço da ciência. Dentre os termos relacionados com a *e-science*, destacam-se na literatura: ciência orientada a dados, computação fortemente orientada a dados, ciberinfraestrutura, dados científicos, dados de pesquisa ou quarto paradigma (ALVARO *et al*, 2011; CÉSAR JÚNIOR, 2011; HEY; TREFETHEN, 2003; MARCUM; GEORGE, 2010).

César Júnior (2011, p. 7) argumenta que "em muitos momentos a ciência precisou aguardar o aparecimento de tecnologias apropriadas de medição de fenômenos de interesse [...] para que conceitos e teorias realísticas pudessem ser propostos, refutados, aprimorados". Porém, o avanço recente da tecnologia de sensores nas mais variadas áreas (medicina, biologia, física, ciência sociais etc.) e das escalas (das nano escalas às astronômicas) levou a um deslocamento no gargalo para o avanço científico. Na opinião do autor, em vez de a ciência não avançar devido à escassez de dados, hoje em dia, ao contrário, ela frequentemente encontra dificuldades em avançar por seu excesso.

No que diz respeito ao crescimento exponencial de publicações, deve-se retomar a percepção de Solla-Price (1976) que defendeu que, após a Segunda Guerra Mundial, a quantidade de publicações em cada campo do conhecimento, cresceu exponencialmente, duplicando a cada dez ou quinze anos. Estudos recentes, sobre o fenômeno de crescimento da literatura científica, trabalham o volume de informação em papel (ou analógica) *versus* a quantidade de informação em ambiente digital ou virtual. A respeito do assunto, estudo realizado por Hilbert e Lopez (2012) revelou que, no ano 2000, apenas um quarto da informação armazenada no mundo era digital. Os outros três quartos correspondiam à informação analógica (papel, filmes, vinis, fitas magnéticas).

Porém, o ano de 2002 marcou o início da era da informação digital, pois foi o primeiro a ter dados digitais armazenados em uma quantidade maior do que a dos armazenados analogicamente. Já em 2007, apenas 7% dos dados armazenados eram analógicos. A previsão dos autores era de que em 2013 o volume de informação armazenada no mundo equivaleria a 1200 *exabytes*, sendo que destes apenas 2% serão analógicos. Sobre o assunto, Mayer-

Schönberger e Cukier (2013, p. 5) argumentam que o mecanismo de busca Google "processa mais de 24 *petabytes* ao dia, volume milhares de vezes maior que todo o material impresso na Library of Congress". A questão que neste momento se apresenta é: como tratar essa proliferação de dados?

Trazendo a discussão para o âmbito do volume de dados produzidos pelas pesquisas colaborativas do Século XXI, Corrêa (2016) defende que essa dimensão (volume) é em *terabytes* a *zetabytes*. Além disso, acredita que a reutilização dos dados tem se ampliado. Nesse cenário, o autor defende que há uma necessidade de estruturar a gestão de dados científicos, incluindo processos de preservação e reutilização desses dados.

Esse ambiente de dilúvio de dados levou Bell (2011, p. 11) a comentar que ″a origem remota dos dados, assim como o acesso comunitário a dados distribuídos, são apenas alguns dos desafios [da *e-science*]″. Na visão de Bell (2011, p. 12), no Século XXI:

> a maior parte do vasto volume de dados científicos capturados por novos instrumentos 24 horas por dia, todos os dias, junto com a informação gerada nos mundos artificiais dos modelos computacionais, deverá permanecer para sempre num estado submetido à curadoria e acessível para o público para análise contínua.

Dentre as áreas que têm se deparado com o dilúvio de dados, Dozier e Gail (2011, p. 41) relatam os problemas enfrentados pela "ciência emergente das aplicações ambientais" que procura estudar os fenômenos das mudanças climáticas, dentre eles o aquecimento global, e apresentar respostas para questões como, por exemplo, *"*quais são as implicações das mudanças regionais em recursos hídricos para as tendências demográficas, a produção agrícola e a produção de energia?".

A necessidade do homem em saúde e bem estar também revolucionou a medicina, cujo exemplo mais divulgado sobre compartilhamento foi o Projeto Genoma Humano e a descoberta de sequenciar o DNA de uma pessoa. Dentre os desafios contemporâneos, Buchan, Winn e Bishop (2011, p. 114) comentam que "a integração de diferentes escalas de dados, de variáveis do nível molecular ao populacional e de níveis diferentes de precisão da medida de fatores é um grande desafio para a ciência da saúde com uso intensivo em dados".

Em se retomando as colocações de Gray (2007), um telescópio é compartilhado por vários cientistas (entre 20 e 50). Assim, entende-se que áreas como a física e a astronomia compartilham instrumentos de coleta de dados, além de compartilharem os próprios dados coletados. Em contrapartida, Hunt, Baldocchi e Van Ingen (2011, p. 49) argumentam que "na ciência ecológica os dados são gerados por uma ampla variedade de grupos, usando uma ampla

variedade de padrões de dados e de metodologias de amostragem ou simulação". A questão é que os dados obtidos por diferentes fontes são imprescindíveis para obter respostas a problemas complexos.

Hey e Trefethen (2002), bem como Álvaro *et al* (2011) que comentam sobre experimentos em partículas físicas e sobre o grande colisor de *hádrons*[1], conduzidos no Laboratório da Organização Europeia para Pesquisa Nuclear (CERN[2]), envolvem a colaboração de mais de mil físicos de mais de cem instituições internacionais. Foi estimado que esse projeto gerasse muitos *petabytes*[3] de dados ao ano. Em face ao exposto, é pertinente ressaltar a constatação de Lyman e Varian (2003) de que meros dois *petabytes* equivalem ao conteúdo de todas as bibliotecas universitárias dos Estados Unidos.

Sob essas circunstâncias é que Gray (2007, p. 17) afirmou que "*e-science* é o ponto onde a TI (tecnologia da informação) encontra os cientistas". O fato é que autores como Hey e Trefethen (2003), bem como Gray (2007) e Mayer-Schönberger e Cukier (2013) têm destacado a importância da tecnologia da informação na forma de se fazer ciência. Para esses autores, os desafios tecnológicos incluem a necessidade de melhor captar, analisar, modelar, visualizar e preservar as informações científicas, tornando os sistemas de computação vitais para o moderno ambiente de pesquisa.

Na Ciência da Informação, a *e-science* traz implicações relevantes sobre a comunicação científica, afinal os dados oriundos da *e-science* são de fato dados científicos primários. Também traz implicações para os serviços e produtos de informação, bem como afeta diretamente as bibliotecas digitais, porque exige reflexões sobre a preservação digital e o planejamento das bases de dados.

Importante recordar que o surgimento da ciência da informação, independentemente de suas origens (Bélgica, antiga União Soviética e EUA), teve como motivação principal "a preocupação com volumes crescentes de informação científica que desafiavam as tecnologias tradicionais de controle" (MUELLER, 2007, p. 125). Dando continuidade ao assunto, Mueller (2007, p. 143) comenta que "a preocupação [com a] informação cientifica está na origem da ciência da informação". Além disso, a autora nos lembra que para Saracevic (1996) a ciência da informação está inexoravelmente ligada à tecnologia.

---

[1] É o maior acelerador de partículas existente do mundo. É considerado como um dos grandes marcos de engenharia da humanidade. Foi construído pela Organização Europeia para Pesquisa Nuclear (CERN).

[2] Em francês – Conseil Européen pour la Recherche Nucléaire, com sede em Genebra na Suíça.

[3] *1 petabyte* equivale a 1.024 *terabytes*. 1 *terabyte*, por sua vez equivale a 1.024 *gigabytes*. As medidas referem-se a capacidade de armazenamento de um dado dispositivo.

Importante recordar que as preocupações de Bush (1945), com o problema relativo ao armazenamento e acesso à informação científica e tecnológica, que cresceu vertiginosamente durante a Segunda Guerra Mundial, geraram e ainda geram inúmeras pesquisas na Ciência da Informação. O interessante é observar a atualidade e relevância das argumentações de Bush (1945) que ganharam um novo fôlego e enfoque em função do surgimento da *e-science.* Ressalta-se que o foco desta tese será a gestão de dados científicos.

A partir desse contexto, cabe retomar a colocação de Muller (1995, p. 64) que comenta que "os cientistas, no seu esforço para fazer avançar a ciência, necessitam ter acesso constante ao conhecimento já registrado e, nesse processo, farão referências em seus próprios trabalhos a ideias ou resultados de pesquisas de autores que os precederam." Gray (2007) corrobora esse entendimento ao expor que o objetivo dos cientistas é codificar a sua informação para que possam trocá-la com outro cientista.

Em consonância com Muller (2007), Gray (2007) argumentou que todos os dados científicos precisam estar *online*, afinal a internet oferece condições para o cientista disponibilizar muito mais do que apenas o seu artigo, a rede permite "unificar todos os dados científicos e toda a literatura" em um mundo virtual interativo. No Brasil, essa perspectiva vem sendo trabalhada por Sayão e Sales (2014) por meio do conceito de *publicação ampliada*.

Toda essa problemática de volume de dados científicos *online* e a necessidade de compartilhamento desses dados, no âmbito da Ciência da Informação trazem reflexões sobre "como tratar, armazenar, dar acesso e preservar esses dados científicos primários?". Certamente, a análise do fenômeno envolve aspectos legais, como, por exemplo, por quanto tempo armazenar dados produzidos no âmbito das ciências da saúde, ciências ambientais dentre outras? Nesse aspecto, infere-se que a teoria arquivística pode trazer importantes contribuições. Outro aspecto relevante se refere à portabilidade dos dados armazenados, ou seja, a sua preservação digital. Qual a infraestrutura tecnológica adequada para armazenar esse volume de dados de forma a garantir sua preservação?

A medida que a tecnologia da informação avança e as barreiras do livre acesso à informação científica e tecnológica se rompem, o cientista se depara com dúvidas complexas e com a necessidade de refletir sobre: como posso utilizar os dados coletados por outro cientista?; ou, ainda, qual o limite ético da reutilização desses dados?

Com referência ao tratamento técnico da informação, como se dará a curadoria desses dados? Onde as teorias de classificação, catalogação e indexação podem contribuir nesse cenário? A respeito do assunto, Borgman (2013) comenta que os dados oriundos de pesquisas se tornaram um fator crítico para a comunicação acadêmica, gestão da informação e política de

investigação. A autora observa que dados valiosos de pesquisa muitas vezes não são captados, citados ou reutilizados. Nesse sentido, Borgman (2013) considera que o desafio dos bibliotecários consiste em identificar quais as partes desses dados devem ser mantidas em um acervo, qual o caminho certo para mantê-los, quais os instrumentos e quais os serviços certos para tornar esses dados úteis.

A grande parte das indagações acima precisam ser refletidas em diversos níveis da esfera governamental de um país. Assim, parece coerente afirmar que um conjunto de respostas às indagações precisa ser elaborado por meio de uma política pública para a gestão de dados científicos.

Em razão do exposto, é prudente refletir sobre a atual estrutura dos dados científicos no Brasil e como essa gestão tem evoluído. Além disso, é pertinente identificar os principais repositórios de dados científicos do País, bem como os atores estratégicos envolvidos na gestão desses dados. Do ponto de vista estritamente técnico, faz-se necessário apontar soluções para um tratamento adequado dos dados científicos de forma a viabilizar o processo de armazenamento, organização, busca, recuperação e difusão dos dados. Caso contrário os dados coletados podem se tornar inelegíveis ou, o que seria mais drástico, se perder em um grande volume de dados, por falta de tratamento e preservação adequados.

A partir desses fatos, surgem perguntas que permeiam este projeto: quais os países que já dispõem de uma política de gestão de dados científicos? Quais os principais problemas e desafios inerentes à construção e implantação de uma política estruturada de gestão de dados científicos? Dentre os países que já possuem tal política, quais as similaridades? Por fim, como o Brasil está se preparando para esse novo contexto de tratamento da informação oriunda das pesquisas científicas?

A reflexão dessa conjuntura levou a construção do objetivo geral desta tese, a saber: elaborar um esboço de diretrizes para a gestão de dados científicos no Brasil.

A complexidade do tema da tese exigiu que em determinados momentos a pesquisa assumisse um caráter extremamente quantitativo, foi o caso do estudo bibliométrico que norteou a condução de toda a pesquisa. Após essa etapa, a pesquisa adotou a abordagem de investigação qualitativa proposta pela *Grounded Theory*. Ainda assim, alguns aspectos da tese precisaram ser avaliados com o apoio da estatística descritiva, portanto a tese assume um caráter de pesquisa misto e aplicada.

A tese está estruturada da seguinte forma, o primeiro capítulo apresenta uma introdução ao tema e a justificativa de relevância da pesquisa, apresentando sequencialmente seu objetivo geral e seus objetivos específicos. O segundo capítulo apresenta a fundamentação teórica da

tese, nesse aspecto se discute a ciência contemporânea e a literatura sobre *big data*, *e-science*, gestão de dados científicos, políticas de ciência e tecnologia e políticas de informação. O terceiro capítulo apresenta o estudo bibliométrico que norteou a condução da pesquisa e discute os resultados da análise quantitativa da literatura sobre *e-science* nas bases de dados LISA e LISTA. Na sequência é apresentada a metodologia da pesquisa, onde são discutidas considerações sobre a abordagem de investigação utilizada – a *Grounded Theory*. Além disso, são indicados os procedimentos metodológicos de coleta de dados. Por fim, o terceiro capítulo discute de forma sucinta os *softwares* utilizados como apoio na análise de dados, quais sejam o SPSS para o aspecto quantitativo e o NVivo para o aspecto qualitativo. A análise de dados é apresentada no quarto capítulo sempre contrapondo os resultados alcançados na coleta de dados com a literatura revisada sobre o tema. Já o quinto e último capítulo expõe a conclusão do estudo e apresenta algumas considerações para pesquisas futuras no tema.

Por se tratar de um tema extremamente novo, o leitor, por vezes**,** pode se confundir conceitualmente entre o que é *e-science*/*cyberinfrastructure* e o que são os dados de científicos, algumas vezes também denominados dados de pesquisa.

Em sentido *latu*, *e-science* e *cyberinfrastructure* são termos de grande abrangência que se referem à uma nova forma de se fazer ciência, bem como à infraestrutura tecnológica necessária para apoiar a pesquisa científica do Século XXI, como, por exemplo, a computação em *grid*[4] e bancos de dados que suportem *petabytes* de dados não estruturados, com fluxo constante. No sentido *strictu*, *e-science* refere-se a uma nova forma de se fazer ciência, enquanto *cyberinfraestructure* refere-se a infraestrutura tecnológica que viabiliza a *e-science*. O termo *e-science* é predominantemente utilizado pelo Reino Unido, enquanto o termo *cyberinfraestructure* é utilizado nos Estados Unidos, conforme será apresentado e esclarecido durante a revisão de literatura.

Os dados científicos e/ou de pesquisa, objeto desta tese, são aqueles coletados em grande volume, por sensores, telescópios, satélites, dentre outros instrumentos e que exigem a infraestrutura tecnológica acima comentada para processamento e análise. Isto posto, cabe

[4]A computação em *grid* é um modelo computacional capaz de alcançar uma alta taxa de processamento de dados dividindo as tarefas entre diversas máquinas. Os grids são compostos por recursos heterogêneos, reunindo desde clusters e supercomputadores, até desktops e dispositivos móveis. Essas máquinas podem estar em uma rede local ou em uma rede de longa distância, o que, por sua vez, forma uma máquina virtual. O processamento de dados pode ser executado no momento em que as máquinas não estão sendo utilizadas pelo usuário, assim evitando o desperdício de processamento da máquina utilizada. De acordo com Buyya (2005) "algumas destas aplicações estão relacionadas ao termo e-science, que denota a pesquisa realizada de forma colaborativa em escala global. Este ambiente de e-science envolve o compartilhamento de instrumentos científicos, dados distribuídos, visualização remota e interpretação colaborativa de dados e resultados, se adequando perfeitamente às características de uma infraestrutura de computação em grade". Dentre as inciativas nacionais de computação em grid destaca-se o LNCC. Disponível em: <http://www.portalgrid.lncc.br>.

ressaltar que ainda não há um consenso na literatura quanto ao uso da expressão *dados científicos* ou *dados de pesquisa*. Os autores Hey e Hey (2006), Bell (2011) e Rodrigues *et al*. (2010) utilizam dados científicos. Por outro lado, Borgman (2015), Sales (2014), Sayão e Sales (2014) utilizam o termo dados de pesquisa (*data scholarship*). Da mesma forma, também não há um consenso quanto ao uso do termo *e-science* ou *cyberinfrastructure*. Tal situação já havia sido observada na indexação das bases de dados *Library and Information Science Abstracts* (LISA) e *Library and Information Science & Technology Abstracts* (LISTA), em estudo bibliométrico sobre a literatura referente ao tema, realizado por Costa e Cunha (2015). Nesse estudo**,** os autores observaram que o termo *e-science* foi mais utilizado como indexador, em detrimento do termo *cyberinfrastructure*.

Essa pluralidade de conceitos relacionados e muitas vezes interpretados como sinônimos reflete a ebulição pela qual a temática tem passado – é uma área em formação, tanto é que ainda não há um vocabulário controlado que tenha estruturado os termos referentes a *e-science*, à exceção de um embrião de cabeçalho de assunto relacionados ao tema, desenvolvido pela Library of Congress[5]. O surgimento de uma nova área pressupõe o advir de novos termos que serão amadurecidos conforme a temática vai se consolidando. Como exemplo, é pertinente recordar o aparecimento do conceito de biblioteca digital, por vez chamada de biblioteca virtual e até mesmo ciberteca.

## 1.1 PROBLEMA E JUSTIFICATIVA

A sociedade da informação traz em sua essência a necessidade de acesso rápido e preciso à informação, seja essa científica, tecnológica ou não. O amadurecimento dessa sociedade culminou com o desenvolvimento de novas tecnologias da informação e comunicação que geram e armazenam dados em quantidades jamais vistas.

No meio desse dilúvio de dados encontram-se pesquisadores, analistas de sistemas, bibliotecários, arquivistas, agentes formuladores de políticas públicas, políticos, dentre outros; todos preocupados em refletir sobre os aspectos inerentes à informação. Enquanto os bibliotecários e arquivistas discutem as formas de tratamento técnico para armazenar e recuperar a informação, analistas de sistemas procuram trabalhar na miniaturização de componentes de armazenamento dos dados, bem como em aplicativos para processamento de

[5] Cabeçalho de assunto disponível em - http://www.loc.gov/rr/scitech/tracer-bullets/esciencetb.html

dados em grande escala. Por outro lado, pesquisadores refletem como analisar a enorme quantidade de dados coletados.

Do ponto de vista da política pública, os técnicos tentam nortear e regulamentar o mercado da informação, direcionar as questões inerentes à segurança da informação, disciplinar o armazenamento da informação de valor primário etc. Enquanto isso, deputados e senadores procuram legislar sobre temas intrínsecos a uma política de informação, tais como o depósito legal, a lei de acesso à informação, a liberdade de imprensa, dentre outros. Ainda no âmbito governamental, não se pode esquecer dos programas e das iniciativas do governo federal em lidar com a sociedade da informação, a exemplo do antigo Programa Sociedade da Informação.

O cenário exposto evidencia a emergência do tema informação e sua complexidade. O Brasil carece de uma política explícita que norteie as ações do Estado em diversos seguimentos do tema informação, desde os mais tradicionais, como o depósito legal, até os mais recentes como, por exemplo, os dados oriundos do *big data* e da *e-science*.

Se o *big data*, ao viabilizar um *marketing* personalizado na venda de produtos ou no redirecionamento de campanhas eleitorais mostra-se promissor; sob outro ângulo, ele apresenta seu lado perverso, quando informações pessoais são acessadas e manipuladas por governos em prol da dita segurança nacional. O exemplo de maior repercussão, até o momento, foi o Edward Snowden que revelou o programa de vigilância da National Security Agency.

Aspectos éticos e políticos também vêm à tona quando se reflete sobre os dados científicos coletados em larga escala. Qual o limite[6] e as regras para a reutilização dos dados? Por quanto tempo eles poderão ser armazenados? Quais as regras para o compartilhamento de dados?

A respeito do tratamento técnico dos dados científicos primários, as dúvidas também são muitas. Como se dará a curadoria desses dados? Qual a infraestrutura tecnológica adequada para armazenar esse volume de dados de forma a garantir sua preservação? Quais serão as habilidades requeridas para o profissional que deseja trabalhar com organização de dados da *e-science* no Brasil? Como organizar esses dados de forma a disponibilizá-los em rede?

Em razão do exposto, é prudente refletir sobre a necessidade de um tratamento adequado que viabilize o processo de coleta, armazenamento, organização, busca, recuperação, difusão e preservação dos dados científicos primários coletados em grande escala. Caso contrário, os dados coletados podem se tornar inelegíveis ou, o que seria mais drástico, se perder em um grande volume de dados, por falta de tratamento técnico adequado.

---

[6] O limite de reutilização apresenta discussões densas quando se trata de dados sobre o ser humano, a vida.

A grande parte das indagações acima precisam ser refletidas em diversos níveis da esfera governamental de um país. Parece coerente afirmar que um conjunto de respostas às indagações precisa ser elaborado por meio de uma política pública para a gestão de dados científicos.

Assim surgem as perguntas que permeiam este projeto: quais os países que já dispõem de uma política de gestão de dados científicos oriundos da e-science? Quais os principais problemas e desafios inerentes à construção e implantação de uma política estruturada de gestão de científicos? Dentre os países que já possuem tal política, quais as similaridades? Por fim, como o Brasil está se preparando para esse novo contexto de tratamento da informação oriunda das pesquisas científicas?

A partir da conjuntura exposta, o problema de pesquisa que aglutina as questões acima é a *ausência de uma política governamental explícita que norteie as ações relacionadas com a gestão de dados científicos no contexto brasileiro*.

E possível afirmar que as discussões sobre a gestão dos dados científicos no âmbito mundial encontram-se em estágio inicial, sendo que no Brasil, especificamente, situa-se numa fase embrionária.

Sales (2014, p. 20) afirma que com "a chamada *e-Science*, ou quarto paradigma científico, fica patente que a adição de outros recursos ao texto, como imagens, sons e interatividade, agora se torna fundamental". Esta tese vai além dessas preocupações iniciais de Sales (2014) ao se propor a elaborar um conjunto de diretrizes que possam ser utilizadas pelo governo brasileiro, no que diz respeito à condução de normas e esclarecimentos para o dado coletado em grande escala.

Ao se considerar que as políticas governamentais têm procurado auxiliar o Brasil a sair da condição de emergente para integrar o privilegiado rol dos países desenvolvidos, bem como que o acesso à informação científica e tecnológica é essencial ao desenvolvimento econômico de um país, a pesquisa mostra-se oportuna para apontar diretrizes que norteiem a gestão dos dados científicos, de forma a contribuir com um dos aspectos de políticas de ciência e tecnologia, defesa nacional e informação científica e tecnológica.

Diante do exposto, acredita-se que essa pesquisa trouxe contribuições teóricas para a Ciência da Informação ao propor um *framework* com diretrizes para a elaboração de uma política para a gestão de dados científicos no Brasil.

## 1.2 OBJETIVOS

### a) Objetivo Geral

O objetivo geral desta pesquisa é elaborar um esboço de diretrizes governamentais para a gestão de dados científicos no Brasil.

### b) Objetivos Específicos:

Os objetivos específicos desta pesquisa são:

**OE 1** – Identificar os países desenvolvidos que possuem ações de governo para a gestão de dados científicos.

**OE 2** – Analisar as ações de governo de países desenvolvidos sobre a gestão de dados científicos nos países identificados.

**OE 3** – Identificar os principais problemas e as soluções inerentes à construção de uma política estruturada para a gestão de dados científicos.

**OE 4** – Identificar a postura das agências de fomento no Brasil com relação ao tema.

**OE 5** – Identificar o posicionamento dos pesquisadores brasileiros envolvidos com o tema.

# 2 FUNDAMENTAÇÃO TEÓRICA

Com o objetivo de contextualizar o tema desta pesquisa, neste capítulo é apresentada uma breve exposição sobre a organização social e o impacto da ciência contemporânea. Em seguida, é comentado o fenômeno do *big data*, a contextualização e o surgimento da *e-science*, seu cenário no mundo e no Brasil. Na sequência, discute-se sobre a política de ciência e tecnologia no Brasil e as políticas de informação que permearam essa área. Espera-se que, a partir da teoria enfocada, se encontre o subsídio teórico necessário para a discussão sobre a política de gestão de dados científicos no Brasil.

## 2.1 A ORGANIZAÇÃO SOCIAL DA CIÊNCIA CONTEMPORÂNEA

Discorrer sobre a ciência contemporânea não é tarefa trivial. Escolher os marcos históricos a serem abordados tornam a tarefa mais árdua. Entretanto, um recorte precisa ser feito para situar o leitor na discussão sobre o crescimento da produção científica que culminou com o dilúvio de dados. Face ao exposto, esse capítulo procurar abordar de forma sucinta quais as mudanças que ocorreram no processo de fazer pesquisa no mundo, bem como o panorama da pesquisa científica atual.

Assim, a partir do entendimento de que a ciência é socialmente construída em função da relação entre cientistas e a sociedade, e que há um processo de transmissão do conhecimento produzido pelo cientista para outros cientistas e demais membros da sociedade, esta tese optou por acomodar a teoria sob o fenômeno de produção da ciência, desde as associações científicas que impulsionaram a publicação periódica sobre o conhecimento científico então produzido, até os dias atuais. Em razão desse marco histórico, será delineada a evolução da ciência no que diz respeito ao aumento exponencial da publicação científica, o protagonismo da internet no processo de comunicação científica e, mais recentemente, ao fenômeno de dados coletados em grande escala.

Entende-se que o princípio básico do fazer ciência inclui *comunicar* os resultados da pesquisa realizada para a validação de seus pares. Ou seja, a palavra ‘comunicar’ em si já traz à tona os processos de *comunicação científica*, objeto de estudo da Ciência da Informação, trabalhados por Solla-Price (1976), Garvey (1979), Ziman (1984), Muller (1995), Meadows (1999), dentre outros autores.

As origens da ciência moderna se encontram na Inglaterra do Século XVII (ALFONSO-GOLDFARB, 1994; SOLLA-PRICE, 1976). Nesse período, a ciência não precisava de grandes

justificativas. Quando sofria ataques sua resposta estava sempre voltada para o futuro e não para o passado.

De acordo com Caribé e Mueller (2010), a literatura cientifica revela que as primeiras academias e a divulgação da ciência surgiram no Século XVII. Os cientistas que iniciaram essas academias comunicavam seus resultados de pesquisa por meio de cartas (no inglês, *Letters*). Isso porque essas cartas se confundiam com a correspondência comum e na época não eram abertas pelo governo, o que evitava serem alvo da Inquisição. Essas cartas enviadas às Academias deram origem aos primeiros periódicos científicos.

A restauração da monarquia[7] em Londres fez com que pequenos grupos se reunissem para debater questões filosóficas. Essas reuniões, em um momento posterior, tornaram-se mais regulares e oficiais, fato esse que levou à formação da Royal Society em 1622, que desde seu início interessou-se pela comunicação científica e, portanto, dedicou-se à coleta e análise de informações importantes sobre pesquisas em andamento. Assim, enquanto alguns membros viajavam para obter informações, outras pessoas foram selecionadas no exterior para fazer parte da nova sociedade. O volume de correspondência entre os membros da Royal Society logo passou a ser um problema de difusão da informação, o que culminou com a necessidade de se fazer uma publicação impressa com as cartas recebidas classificadas como mais importantes. Com isso, em 1665, foi criado o periódico científico *Philosophical Transactions* (MEADOWS, 1999, p. 5-6).

Movimento semelhante deu-se na França. Meadows (1999, p. 6) relata que em 1665 "Denis de Sallo começou um periódico dedicado a publicar notícias sobre o que acontecia na Europa" – o *Journal de Sçavans*, considerado o percursor do periódico moderno em humanidades.

De acordo com Meadows (1999), no início, a formação de sociedades ocorreu lentamente, mas, no Século XVIII, acelerou-se bastante. Esse fato permite afirmar-se que se aumentou o número de cientistas, também houve incremento no número de pesquisas realizadas e, por fim, o aumento do número de publicações produzidas.

A respeito do assunto, Solla-Price (1976) teoriza sobre a pequena ciência e a grande ciência (*little science* e *big science*), argumentando que a transição de uma para a outra foi gradual. O autor (SOLLA-PRICE, 1976, p. 3) defende a ideia de que "se um seguimento

[7] Em 1649, auge da Guerra Civil Inglesa, Carlos I foi executado no Palácio de Whitehall. Seu filho, Carlos II foi proclamado pelo parlamento escocês como Rei da Grã Bretanha e Irlanda. Porém, a proclamação foi dada como inválida em função de um estatuto aprovado no parlamento inglês. Nesse período, a Inglaterra tornou-se uma república liderada por Oliver Cromwell. A morte de Cromwell em 1658 resultou em uma crise política que culminou com a restauração da monarquia em Londres.

suficientemente amplo da ciência for medido de alguma forma razoável, o modo normal de crescimento é exponencial".

Há um consenso, na literatura, de que a Segunda Guerra Mundial acabou por institucionalizar a ciência, em particular, devido ao modo como os Estados Unidos (EUA) utilizaram-se da pesquisa científica durante o conflito. Sobre assunto, merece destaque o relatório de Vannevar Bush, publicado em 1945 – *Science the Endless Frontier*. No relatório, Bush defende que o Estado deve se responsabilizar pelo desenvolvimento científico do país, pois a pesquisa propicia avanços tecnológicos.

Assim, a ciência era vista como uma forma de propiciar o bem estar público, combater as enfermidades, promover a segurança nacional, gerar emprego e viabilizar o crescimento industrial. Em consonância com a tese de autonomia científica de Merton (2013), a ingerência do Estado não deveria ser tolerada, cabendo à comunidade científica a autonomia, o controle e o direcionamento de suas atividades. Na percepção de Schwartzman (2001), esse modelo ficou conhecido como Modelo Linear da Ciência.

A partir do Pós-Guerra e início da década de 1970 percebe-se uma ciência ainda autônoma, mas com maior regulação do Estado. Além disso, a visão positiva da ciência, bem como da tecnologia, passa a coexistir com os efeitos negativos que elas podem gerar no mundo. Como exemplo, citam-se a degradação do meio ambiente e o aumento das desigualdades sociais. Também são exemplos de aspectos negativos o enclausuramento e a fragmentação do saber (MORIN, 2010), ou a ainda a industrialização da vida (BORGMANN, 2006; MORIN, 2010). Nesse ínterim, os resultados das pesquisas começam a ser questionados sob a perspectiva de um retorno para a sociedade.

O período de guerras sempre foi marcado por grandes descobertas científicas. Após a Segunda Guerra Mundial o mundo se deparou com a Guerra Fria, um momento de grandes avanços na tecnologia espacial em função da disputa entre Estados Unidos e a antiga União Soviética na corrida para chegar à lua. Mas não foi só a área espacial que se beneficiou dos investimentos feitos em pesquisa nessa época. Os Estados Unidos, no auge da Guerra Fria, tinham interesse em desenvolver uma rede de computadores que ligasse pontos estratégicos para o país, tais como as bases militares e os centros de tecnologia. Essa rede de comunicação deveria ser desprovida de um controle centralizado para resistir a um possível ataque nuclear. Assim, se houvesse perda de uma parte da rede, o seu funcionamento não ficaria comprometido.

A Rand Corporation, organização situada em Santa Mônica, Califórnia, e ligada ao governo americano foi a responsável pelo conceito tecnológico que deu origem à Advance Research Project Agency Network, mundialmente conhecida como ARPANET. A rede teve

sua primeira versão disponibilizada em 1969 com apenas quatro pontos, mas cresceu como uma colônia bacteriana (ERCILIA, GRAEFF, 2008; RHEINGOLD, 1996).

O que acelerou a utilização da ARPANET pelo meio acadêmico foi a comunicação mediada pelo computador (CMC), por meio das mensagens eletrônicas (*e-mail*). O sucesso de utilização da rede levou à criação da TELNET em 1974. Apenas na década de 1990 a rede ganhou reconhecimento fora do meio acadêmico, momento em que foi lançado o primeiro provedor de acesso discado à Internet dos EUA. Em 1991 a criação do sistema de hipertexto – *World Wide Web* por Tim Berners-Lee, à época bolsista do CERN, acabou por facilitar a navegação pela *web*. O fato, atrelado ao desenvolvimento do primeiro programa navegador (*browser)* popularizou a Internet (CUNHA, 2004; ERCILIA, GRAEFF, 2008; RHEINGOLD, 1996).

A Internet modificou a forma de interação entre as pessoas e, na visão de Castells (2003, p. 7) tornou-se a espinha dorsal da sociedade contemporânea, "a base tecnológica para a forma organizacional da era da informação – a rede".

Hey e Trefethen (2005) argumentam que o advento da internet trouxe a possibilidade de os pesquisadores acessarem recursos armazenados em diferentes lugares, por meio de *sites*, provendo um ambiente de pesquisa *e-science* robusto e útil, onde diferentes grupos podem combinar suas atividades de pesquisa. Como exemplo, os autores citam o *e-Science Project on Integrative Biology*[8], com um orçamento de £ 2,3 milhões, para desenvolver um laboratório virtual para pesquisa de câncer e doenças do coração. O projeto foi liderado pela Universidade de Oxford e envolvia outras quatro universidades britânicas, além da Universidade de Auckland na Nova Zelândia.

Na Universidade de Oxford, o professor Denis Noble liderava as pesquisas no âmbito de modelos de comportamento elétricos das células cardíacas. Enquanto isso, na Nova Zelândia, Peter Hunter do Departamento de Bioengenharia da Universidade de Auckland, liderava estudos pioneiros sobre modelos mecânicos de batimentos cardíacos. Portanto, ambos os grupos desenvolviam pesquisas em nível mundial, cada um com sua especialidade. A relevância de ambos os projetos tornou emergente conectar os dois grupos de pesquisa em uma organização científica virtual. Essa organização virtual era de acesso restrito aos pesquisadores envolvidos no projeto e permitia que os mesmos tivessem acesso a recursos computacionais e aos supercomputadores britânicos (HEY; TREFETHEN, 2005).

[8] Disponível em: <www.integrativebiology.ox.ac.uk>. Acesso em: 2 out. 2016.

É nesse contexto que surge um novo tipo de ciência, colaborativa e dependente de uma infraestrutura tecnológica. Nas palavras de Hey e Trefethen (2005, p. 819), "a necessidade de auxílio para organização, registro e pesquisa dos dados está se tornando aguda*".*

No âmbito do *big data*, a ciência colaborativa tem seu marco inicial com o Projeto Genoma Humano. Tapscott e Williams (2007) consideram que o Projeto Genoma Humano representou um divisor de águas. Afinal, as indústrias farmacêuticas pararam com as suas tentativas isoladas de mapear o genoma e passaram a apoiar colaborações abertas (*open-science)*. A experiência deste projeto representa o resultado final de grande concentração de esforços públicos e privados em prol da informação genética do ser humano. Outros grandes exemplos de colaboração são: a) as iniciativas do CERN de descobrir a partícula da vida – uma partícula subatômica que poderia ser o bóson de Higgs; b) o Projeto Netuno do Observatório Oceânico EUA-Canadá; c) o Projeto de Celeste Digital Sloan dentre outros. Todos esses projetos têm em comum o enorme volume de dados coletados por sensores especializados.

Essas iniciativas que envolvem o compartilhamento de recursos e a infraestrutura tecnológica acabam por ser realizados em diferentes instituições, que, por sua vez, podem estar em distintos países. Esse fato, em conjunto com a facilidade de acesso à informação contribuiu para uma internacionalização e virtualização da ciência.

Cordeiro *et al*. (2013, p. 13) com fundamento em Kuhn (1962 apud KUHN, 2009) consideram que "o ato de fazer ciência passou por significativos aprimoramentos e refinamentos em sua metodologia de trabalho, incluindo [um] novo ferramental lógico-matemático, novos instrumentos de observação e novos paradigmas de estruturação do pensamento científico".

As infraestruturas tecnológicas necessárias para apoiar os projetos supracitados buscam apoio na tecnologia de *grid* computacional. De acordo com Góes *et al.* (2005), em meados da década de 1990, inspirados pelo sistema de energia elétrica e o desejo de alta performance computacional, como, por exemplo, um supercomputador, impulsionou

> [...] o desenvolvimento de uma nova infraestrutura computacional pelo acoplamento de recursos distribuídos geograficamente com bases de dados, servidores de armazenamento, redes de alta velocidade, supercomputadores e aglomerados para solucionar problemas de grande escala (GÓES *et al*., 2005).

Hey e Trefethen (2005, p. 819) argumentaram àquela época que "infelizmente as versões atuais de *grid* provem apenas uma pequena parte das funcionalidades requeridas para a colaborações *e-science*".

A partir do exposto, entende-se que são características da *e-science* a pesquisa colaborativa, produzida por uma equipe multidisciplinar, que coleta uma grande quantidade de dados, em diferentes lugares. Tais características impulsionaram para o chamado dilúvio de dados que precisam ser gerenciados para viabilizar sua preservação e posterior recuperação.

Essa nova mentalidade de compartilhamento do dado em prol de uma pesquisa colaborativa retoma o *ethos* científico proposto por Robert Merton. Sem adentrar na polêmica da interação ciência, universidade e indústria, intensificada após a década de 1950 e sem adentrar na teoria de ciência pós-acadêmica proposta por Ziman (2000), cabe ressaltar que o compartilhamento de dados abertos *online* está em sintonia com o conceito de *comunism*[9]*o* do *ethos* mertoniano.

É nesse cenário de comunismo dos dados científicos que merece ser destacada a Declaração de Berlin sobre acesso livre ao conhecimento nas ciências e humanidades que expõe:

> A internet transformou radicalmente as realidades práticas e econômicas da difusão do conhecimento científico e do patrimônio cultural. Pela primeira vez na história, a Internet oferece-nos a possibilidade de constituir uma representação global e interativa do conhecimento humano, incluindo o patrimônio cultural e a garantia de acesso mundial. [...] redigimos essa declaração para promover a Internet como o instrumento funcional ao serviço de uma base de conhecimento científico global e do pensamento humano, e para especificar medidas que os responsáveis políticos, os institutos de investigação, as entidades financiadoras, as bibliotecas, os arquivos e os museus devem considerar.

A respeito das formas de evolução da produção científica, ou mesmo da colaboração entre pesquisadores propiciada pela internet, Sales (2014, p. 20) argumenta que

> é preciso se valer de todos os artifícios da trazidos pelo advento da tecnologia para fazer com que a comunicação científica siga para além de um documento simples. O novo padrão de produção de conhecimento científico, baseado na geração intensiva de conjunto de dados, demanda tipos inéditos de publicações que consigam integrar dados de toda natureza e publicações tradicionais de formas digitais, criando um novo gênero de publicação web.

Se nos primórdios da ciência moderna, a sociedade se preocupou com o armazenamento dos dados científicos primários, registrados em cadernos pessoais, bem como com a preservação dos resultados das pesquisas publicados em artigos de periódicos e livros, o momento é propício para a sociedade se preocupar com a gestão dos dados científicos coletados *online*, de forma a garantir o acesso a futuras gerações de pesquisadores.

[9] A produção acadêmica deve ser pública em benefício do próprio desenvolvimento científico.

## 2.2 ASPECTOS GERAIS DO *BIG DATA*

*Big data* é um termo referente à grande quantidade de dados não estruturados, que atualmente são produzidos e disponibilizados em rede. No entendimento de Firestone (2010, p. vii), "a explosão das redes móveis, computação em nuvem e novas tecnologias deram origem a grandes e incompreensíveis mundos de informação, frequentemente denominados como big data".

Para Mayer-Schönberger e Cukier (2013, p. 4) o *big data* refere-se a "trabalhos [de processamento de dados] em grande escala que não podem ser feitos em escala menor, para extrair novas ideias e criar novas formas de valor de maneira que alterem os mercados, as organizações, a relação entre cidadãos e governos etc." Os autores entendem que as empresas de tecnologia da informação estão na linha de frente do uso do dilúvio de dados, afinal elas têm acesso à esses dados só pelo fato de estarem *online*.

Interessante observar que, na opinião de Mayer-Schönberger e Cukier (2013, p. 5), "a verdadeira revolução não está nas máquinas que calculam os dados, e sim nos dados em si e na maneira como os usamos" – o que pressupõe uma nova mentalidade de como os dados podem ser utilizados.

De acordo com Davenport (2014, p. 3-7), o conceito é revolucionário e começou a ganhar força no quarto trimestre de 2010. Para o autor, o *big data* é caracterizado por um grande volume de dados desestruturados, provenientes de diversas fontes e com a necessidade de análise de contínua dos mesmos (*streaming data*).

A literatura indica que a definição de *big data* pode apresentar variações conforme a área de aplicação, por exemplo, na ciência da computação, na análise de finanças e até mesmo no caso de um empresário que está lançando uma ideia para um empreendimento capitalista. Entretanto, há um consenso de que *big data* se refere à crescente capacidade tecnológica para captar, agregar e processar um volume cada vez maior de dados, que dificilmente seriam processados com as aplicações de tecnologia da informação tradicionais existentes. (BOLLIER, 2010; MAYER-SCHÖNBERGER; CUKIER, 2013; UNITED STATES, 2014).

São exemplos desses dados os *posts* da redes sociais, seja ela Facebook, Twitter ou algum outro aplicativo social, os dados coletados pela tecnologia de RFID, dados de localização geográfica de um usuário de aplicativo de mapas da Google – disponibilizados na rede por meio do seu telefone celular ou do aparelho GPS do automóvel, dados de compras *online* realizadas com cartão de crédito, dados dos programas de televisão e filmes assistidos na *smart* TV por meio do Netflix ou Youtube, dentre tantos outros exemplos. Esses dados podem ser utilizados

em benefício de políticas públicas na área de saúde e educação. Também têm aplicação no conceito de *smart cities* e têm sido frequentemente utilizados por empresas de comércio eletrônico para aprimorar suas estratégias de vendas.

Mayer-Schönberger e Cukier (2013, p. 8) argumentam que:

> *big data* relaciona-se com previsões. Apesar de ser descrito como ramo da ciência da computação chamado de inteligência artificial e, mais especificamente, uma área chamada 'aprendizado de máquina', esta ideia é enganosa. *Big data* não tem a ver com 'ensinar' um computador a 'pensar' como humano. Ao contrário, trata-se de aplicar a Matemática a enormes quantidades de dados a fim de prever probabilidades.

As análises geradas com enorme quantidade de dados modificam o de entendimento dos fenômenos, que deixam de ser estudados com base em dados amostrais para serem analisados *'no todo'*. Nas palavras de Mayer-Schönberger e Cukier (2013, p. 9), "big data tem a ver com o quê, e não com o porquê. Nem sempre precisamos saber a causa do fenômeno; em vez disso, podemos deixar que os dados falem por si".

Para Mayer-Schönberger e Cukier (2013, p. 28), compreender o fenômeno do *big data* e tirar proveito dele implica em mudar a mentalidade com que se vê o mundo, "[...] a obsessão pela exatidão remonta a era analógica e escassa de informações. Quando os dados são esparsos, todos os pontos de dados são essenciais e, assim, toma-se um cuidado maior para evitar que qualquer ponto influencie a análise".

Sobre o assunto, Davenport (2014) argumenta que há diferenças entre a análise aplicada ao *small* data não estruturado e a análise do *big data*. Uma delas é a noção de inferência estatística, afinal "generalizar os resultados de pequenas amostras para populações muito maiores – pode ser menos essencial" (DAVENPORT, 2014, p. 92). A tecnologia do *big data* permite que as organizações analisem toda a população de dados. Consequentemente, com esse modelo não há necessidade de se preocupar com conceitos tais como significância estatística e probabilidade de resultados.

Diante desse novo cenário, Firestone (2010, p. viii) reflete: "a habilidade de analisar essa enorme quantidade de dados muda a natureza da metodologia científica?".

Além do exemplo da análise de disseminação do H1N1 pela Google[10], tem-se o relatado na literatura do momento em que a Amazon demitiu seus editores e críticos literários que faziam as recomendações de livros. O engenheiro de *software* da empresa percebeu que não era necessário comparar as pessoas, mas sim encontrar associações entre os produtos comprados –

[10] Exemplo relatado no capítulo de introdução desta tese.

o que gerou a patente da *filtragem colaborativa item a item*, processo que revolucionou o *e-commerce* (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 35-36).

As tentativas de se produzir um bom *software* de tradução datam da década de 1940, mas a ideia ganhou impulso durante a Guerra Fria a partir da necessidade dos EUA de captar dados da antiga União Soviética. Em 1954, um computador da IBM traduziu 60 frases em russo para o inglês. Em 1966, estudiosos sobre o tema admitiram o fracasso do projeto, pois a tradução não envolve apenas regras, mas também exceções. Já no final da década de 1980, os pesquisadores da IBM desenvolveram um projeto em que alimentaram o computador com regras linguísticas explícitas, um dicionário e regras de probabilidade para o *software* calcular que palavra ou expressão do idioma seria mais apropriada. Em 1990 o projeto utilizou 10 anos de transcrições parlamentares. O processo desenvolvido pela IBM ficou conhecido como *tradução mecânica estatística* e transformou o problema de tradução em um problema matemático que parecia ter dado certo. Apesar dos dez anos de transcrições ser um banco de dados enorme para a época, o investimento era alto e a IBM acabou desistindo do projeto. Em 2006, a empresa Google entrou no ramo da tradução utilizando toda a Internet como banco de dados. A eficiência do sistema de tradução da empresa acabou por gerar uma piada interna – a qualidade das traduções melhora sempre que um linguista deixa a equipe, outro exemplo de aplicação do *big data* (MAYER-SCHÖNBERGER; CUKIER, 2013, p. 25-27).

Quanto ao volume de dados utilizados para propiciar traduções melhores, é interessante ressaltar a observação de Mayer-Schönberger e Cukier (2013, p. 28):

> [...] quando a quantidade de dados é enorme e de um tipo novo, em alguns casos a exatidão não é o objetivo, desde que possamos descobrir a tendência geral. Passar para a larga escala altera não apenas as expectativas de precisão, como também a habilidade prática de se alcançar a precisão. Apesar de parecer contraproducente a princípio, tratar os dados imperfeitos e imprecisos permite que façamos melhores previsões e entendamos melhor o mundo.

A discussão acima exposta corrobora que as previsões com fundamento nas correlações entre os dados são a essência do *big data*. Tal situação gera a necessidade de um novo perfil profissional – *o cientista de dados*. Sobre o assunto, Mayer-Schönberger e Cukier (2013, p. 88) comentam "em vez de se curvar diante de um microscópio para descobrir os mistérios do universo, o cientista de dados analisa banco de dados a fim de fazer uma descoberta".

Para Davenport (2014) são as pessoas que vão dar sentido ao *big data*. Para o autor, o cientista de dados representa o fator decisivo para o sucesso do *big data* nas organizações, afinal, os dados, de uma forma geral, são gratuitos ou baratos; da mesma forma, o *hardware* e o *software*. Em contrapartida, os analistas de dados se diferem dos analistas convencionais,

sendo que esses são caros e difíceis de contratar. Para o autor, "o cientista de dados clássico possui cinco atributos fundamentais: ele é um hacker, um cientista, um analista quantitativo, um conselheiro de confiança e um expert em negócios" (DAVENPORT, 2014, p. 85). O autor denomina o cientista de dados como pessoas multitalentosas, com necessidade de estímulo intelectual e de crescimento.

Sobre o atributo de ser um *hacker*[11], o autor refere-se à capacidade de codificar e ao domínio de arquiteturas tecnológicas de big data. Dentre elas, a experiência com linguagem de programação, especialmente as linguagens de *script* – Python, Hive, Pig e até mesmo o Java. Davenport (2014, p. 87) argumenta que essas linguagens "dispõem de recursos para dividir grandes problemas de processamento de dados em um *framework* distribuído Map Reduce".

Granville (2013) classifica os cientistas de dados em verticais e horizontais. O autor define os cientistas de dados vertical como aquele que têm conhecimento muito profundo em um campo específico. Eles podem ser cientistas da computação familiarizados com a complexidade computacional de todos os algoritmos de ordenação. Também, podem ser um estatístico com habilidades específicas. Ou, podem ser um engenheiro de *software*, com experiência em escrever código Python (incluindo bibliotecas gráficas) aplicada ao desenvolvimento de API e tecnologia *crawler*[12] *web*. Ou ainda, um especialista de banco de dados com especialização em bases de dados de gráficos, Hadoop e NoSQL. Já os cientistas de dados horizontais, na perspectiva de Granville (2013), são uma mistura de analistas de negócios, estatísticos, cientistas da computação e especialistas em domínio. Eles combinam visão com conhecimento técnico. Eles podem não ser especialista em modelos lineares e algumas técnicas estatísticas, mas conhecem técnicas modernas relacionadas ao *big data*.

Granville (2013) atribuiu características bem específicas aos profissionais que ele identifica como cientistas de dados verticais. Ao mesmo tempo, o autor utiliza um termo pejorativo para definir esses cientistas de dados como – *fake data scientists*. A partir do texto do autor, é possível inferir que, na sua percepção, o cientista de dados apto a lidar com as necessidades do *big data* é o analista horizontal, aproximando-se assim a visão de Davenport (2014).

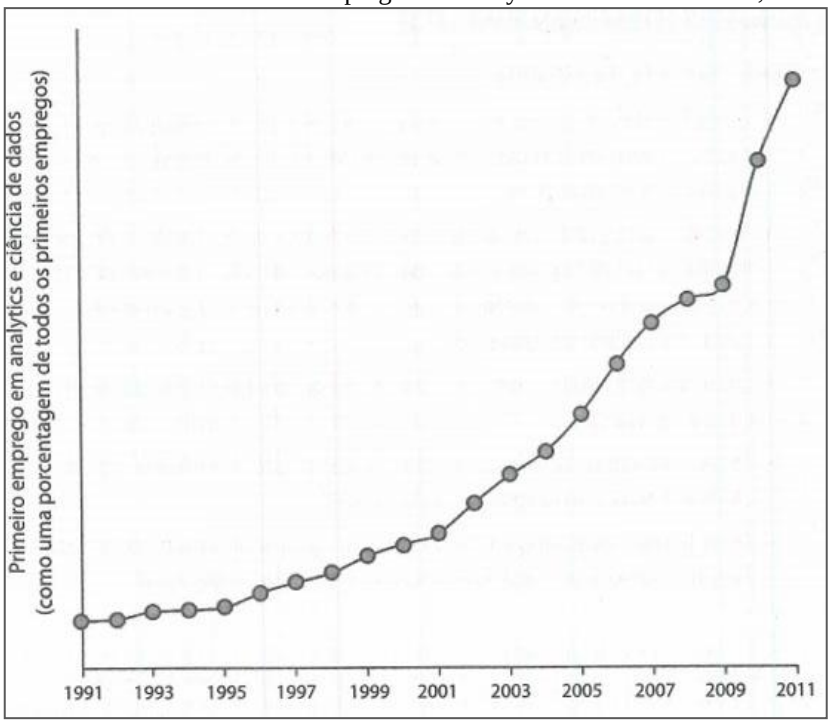
No que diz respeito a formação profissional do cientista de dados, Davenport (2014) acredita que ainda não há no mercado um curso superior em ciência de dados, mas cita

[11] No contexto do big data o *hacking* refere-se a um profissional com computação rápida e criativa. E não o conceito comum de "fora de lei" – aquele que contorna as normas legais do comportamento computacional (DAVENPORT, 2014, p. 88).
[12] Um *crawler* é um programa que visita sites e lê suas páginas e outras informações para criar entradas para um índice de mecanismo de busca (*search engine*) (CRAWLER, 2003).

programas de pós-graduação na área, dentre eles o mestrado oferecido pela School of Information, University of California, em Berkeley. Para o autor, outra opção de formação profissional são os programas especialização em *business inteligence* ou *business analytics*, que por sua vez, já incluem, ou planejam incluir a ciência de dados em seus currículos. Dentre esses cursos, são relacionados por Davenport (2014) os oferecidos pela North Carolina State University, a Northwestern University, New York University, o Stevens Institute of Technology dentre outros. Outra perspectiva de formação profissional, na visão do autor, são os programas de treinamento oferecidos por empresas tais como a Accenture, Deloittee IBM para seus respectivos funcionários. Também há menção ao programa de seis semanas desenvolvido por Jake Klamka, um cientista de dados com formação acadêmica em física de alta energia.

Sobre a expansão do mercado de trabalho para o cientista de dados, Davenport (2014, p. 110) acredita que o ritmo de procura por esses profissionais "se manterá na mesma velocidade alucinante", conforme demonstra o Gráfico 1.

Gráfico 1 – Crescimento do número de empregos em analytics e ciência de dados, de 1991 a 2011.




Fonte: Linkedin Analytics (DAVENPORT, 2014, p. 109).

Davenport (2014) comenta que inicialmente julgou o *big data* a um modismo, nas palavras do autor "um mero vinho velho analítico vertido em uma nova garrafa" (DAVENPORT, 2014, p. 3). Porém, após se dedicar ao estudo do conceito, concluiu que "os dados, a tecnologia e as pessoas são um pouco diferente dos empregados no *analytics*

tradicional" (DAVENPORT, 2014, p. 15). O autor procura clarificar a diferença entre os conceitos, conforme demonstra o Quadro 1.

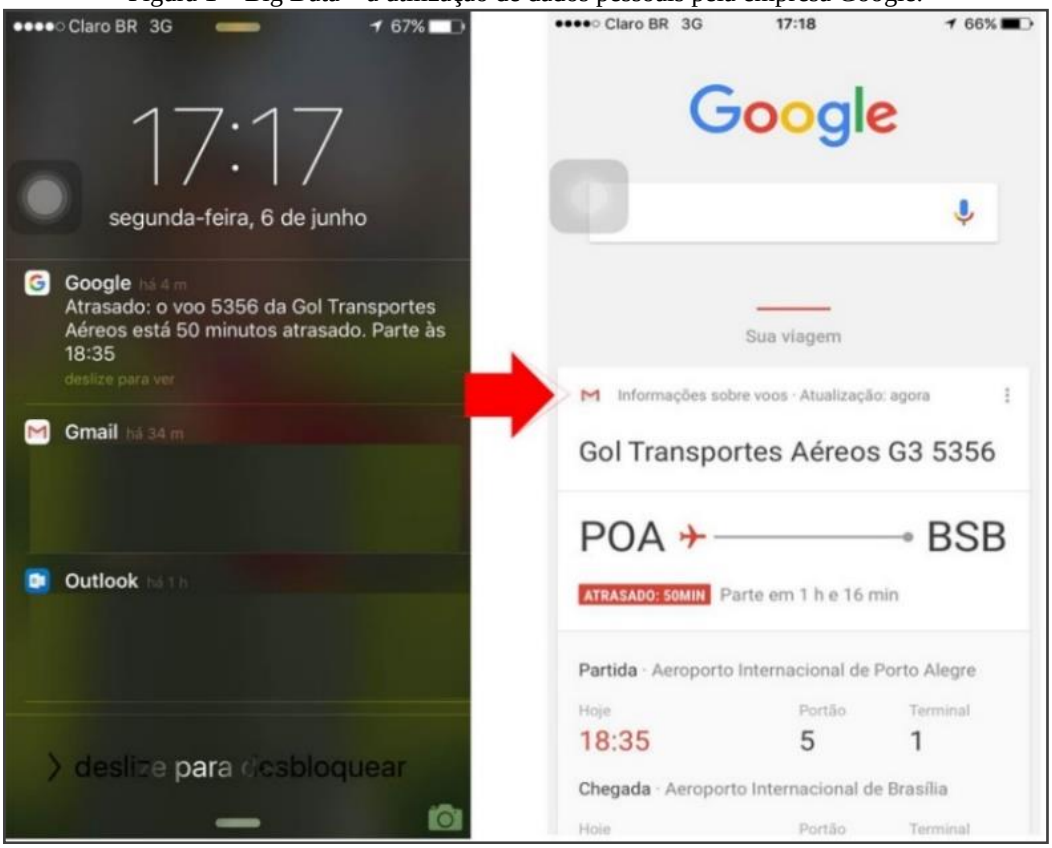
Quadro 1 – Diferença entre os conceitos do *big data* e o *analytics tradicional.*

| | Big Data | Analytics Tradicional |
|---|---|---|
| **Tipos de Dados** | Formatos não estruturados | Dados formatados em linhas e colunas |
| **Volume de Dados** | 100 terabytes a petabytes | Dezenas de terabytes ou menos |
| **Fluxo de Dados** | Fluxo constante de dados | Pool estático de dados |
| **Métodos de Análise** | Aprendizado de máquina | Baseados em hipóteses |
| **Objetivo Principal** | Produtos baseados em dados | Suporte ao processo decisório interno |

Fonte: Davenport (2014, p. 4).

Desde 2015, pelo menos, já se pode dizer que é uma realidade o fato de que os usuários da empresa Google são avisados por meio dos aplicativos em seu celular sobre atrasos no horário do voo que pegarão, o tempo estimado para chegar em casa devido ao trânsito, e, até mesmo recebem felicitações quando fazem aniversário. A Figura 1 demonstra a utilização de dados pela Google para avisar o atraso de um voo no Brasil.

Figura 1 – Big Data – a utilização de dados pessoais pela empresa Google.




Fonte: a autora.

Na área da saúde, Davenport (2014, p. 11) apresenta como exemplo o fato de que algoritmos conseguirão prever a possibilidade de que "*pessoas tenham um ataque cardíaco*" e, consequentemente, paguem mais por um plano de saúde. Outros exemplos apresentados referem-se a algoritmos para monitorar a condição financeira das pessoas, bem como, seu histórico de 'comportamento' e problemas com a polícia local.

Esse poder de uso de dados pessoais, disponíveis na *web* para empresas comerciais e até mesmo para o governo tem suscitado discussões sobre a privacidade individual que envolvem aspectos éticos tais como – quem permitiu a utilização dos 'meus dados pessoais'? Qual o limite para a utilização desses dados? Quais as regras para reutilização dos dados?

Outro ponto polêmico, onde o *big data* mostra seu lado perverso**,** refere-se à utilização de dados pessoais em grande escala em prol da segurança nacional de um País. O exemplo de maior repercussão, até 2016, foi o de Edward Snowden**,** que revelou o programa de vigilância da National Security Agency (NSA). Snowden revelou que o programa acessava vários tipos de dados[13] de usuários dos serviços de acesso à internet fornecidos pelas empresas AOL, Apple, Facebook, Google, Microsoft, Paltalk, Skype, Yahoo! e YouTube. As empresas negaram que tenham oferecido acesso aos dados para o governo americano.

No Brasil, as revelações de Snowden sobre espionagem da Petrobrás, bem como da presidente Dilma Roussef e seus assessores, levaram a reações contra o monitoramento exercido pelo governo americano. Dentre as retaliações do governo brasileiro, destaca-se o fato da presidente ter cancelado sua visita, que ocorreria em outubro de 2013, ao estado de Washington. Além disso, durante seu discurso realizado na abertura da 68ª Assembleia Geral da Organização das Nações Unidas (ONU), ocorrida no dia 24 de setembro de 2013, Dilma Roussef declarou que as ações de espionagem dos Estados Unidos no Brasil ferem o direito internacional, afrontam os princípios que regem a relação entre os países e soa como desrespeito às soberanias nacionais. A situação culminou com a elaboração, pelo Brasil e Alemanha, de uma resolução simbólica que procura garantir o direito à privacidade na era digital. O documento foi entregue à Organização das Nações Unidas, que o aprovou, em 18 de dezembro de 2013, com unanimidade entre os 193 Estados membros.

[13] São exemplos dos dados acessados: conteúdo de *e-mail*, conversas nos aplicativos de mensagens, vídeos e fotos 'baixados' na internet, conversa telefônica, dados de transações bancárias, dentre outros.

Se o mundo terá dados suficientes para estabelecer uma tendência geral para delinear o perfil de consumo, perfil de saúde e até mesmo o perfil de atitudes pessoais, o que será do ser humano quando as áreas de inteligência e a polícia de determinado país resolverem utilizar informações para evitar a criminalidade ou ações antiterrorismo? De repente nos vemos no cenário do filme de ficção científica *Minority Report*, dirigido por Steven Spielberg, lançado em 2002, que descreve a cidade de Washington de 2054. O filme aborda a redução da criminalidade a partir da possibilidade do crime em questão vir a ser executado. Pessoas são presas por pensarem em cometer um crime!

A ironia é que em apenas 12 anos da data de lançamento, a ficção se tornou realidade, ou seja, muito antes do cenário de 2054 relatado no filme. A única diferença é que a divisão pré-crime do filme determinava suas ações por meio de *um possível futuro* visualizado pelos paranormais e clarividentes *precogs*. Em contrapartida, a realidade de 2017 é a possibilidade de prever comportamentos a partir do dilúvio de dados disponibilizados *online* pelo próprio usuário em seu *post* ou *tweet*.

A respeito do tema, Mayer-Schönberger e Cukier (2013, p. 105) alertam sobre o risco da punição com fundamento nas probabilidades oferecidas pela análise do *big data*. Para os autores, "a possibilidade de usar previsões de big data sobre pessoas para julgá-las e puni-las antes mesmo que elas ajam, [...] renega a ideia de justiça e livre arbítrio".

O Escritório Executivo do Presidente dos EUA, por meio do relatório *Big data: seizing opportunities, preserving values*, expõe sua preocupação com o fenômeno e alerta:

> O que realmente importa sobre *big data* é o que ele faz. À parte de como definimos *big data* como um fenômeno tecnológico, a larga variedade de usos potenciais para a analítica do *big data* provocam questões cruciais sobre se nossas normas legais, éticas e sociais são suficientes para proteger a privacidade e outros valores em um mundo *big data*. Poderes computacionais e sofisticação sem precedentes tornam possíveis descobertas, inovações e progressos em nossa qualidade de vida. Porém, estas aptidões, muitas das quais não estão visíveis ou disponíveis ao consumidor médio, também criam uma assimetria de poder entre aqueles que intencionalmente os produzem (UNITED STATES, 2014, p. 3, grifo nosso).

O fato é que a Internet barateou o rastreamento de dados sobre pessoas comuns. Além disso, a Geração Y é conectada e utiliza todo o potencial das redes sociais, seja para informar o que está pensando em um *tweet*, ou para articular uma manifestação pelo Facebook. O que se percebe é que o volume de informações pessoais na *web* tem crescido assustadoramente, mas o

interessante é que, geralmente, essas informações são disponibilizadas pela própria pessoa. Da mesma forma que o volume de dados pessoais e dados comerciais cresceram exponencialmente, os dados científicos coletados *online* cresceram de forma semelhante, dando origem à necessidade de administrar e recuperar esses dados.

No que diz respeito ao grande volume de dados produzidos por instituições, no caso do Brasil, não podemos nos esquecer das grandes bases de dados do governo, como, por exemplo, as produzidas pelo IBGE, as coletadas pelo DATASUS, as recolhidas e geradas pelo IPEA, ou mesmo os dados financeiros do Governo Federal disponíveis no SIAFI[14]. Dentre esses dados, alguns classificam-se como dados abertos e serão abordados a seguir.

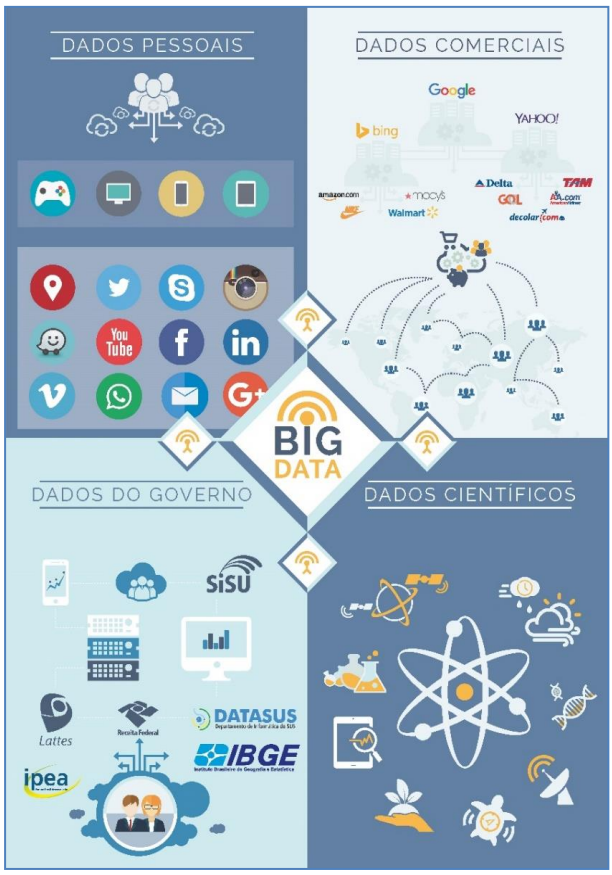
Analisando a questão dos dados produzidos pelo governo, Sayão e Sales (2015, p. 9) defendem que "embora estes dados não tenham sido originalmente coletados para fins de pesquisa, eles se tornam dados de pesquisa uma vez que tenham sido modificados ou expandidos". Os autores defendem que a partir do momento que os dados produzidos pelo Governo são utilizados por alguma área de pesquisa e sofrem alguma modificação, eles passam a ser dado de pesquisa.

Na realidade, o que se tem nesse caso é a utilização de dados governamentais abertos, que podem não ter sido produzidos para uma pesquisa acadêmica/científica, mas certamente foram gerados para a avaliação de programas de governo (saúde, educação, indústria e comércio, desenvolvimento tecnológico etc.), ou ainda, dados referentes ao orçamento do governo, como os disponíveis no SIAFI ou mesmo no Portal da Transparência[15]. A manipulação desses dados abertos por pesquisas científicas gera dados secundários.

Assim, entende-se que o *big data* é composto pelos diversos tipos de dados que muitas vezes são recombinados de forma a gerar novas análises e produtos. A Figura 2 ilustra o conceito de *big data*, *e-science*, dados de governo, dados abertos e a relação entre todos eles.

[14] O Sistema Integrado de Administração Financeira do Governo Federal consiste no principal instrumento utilizado para registro, acompanhamento e controle da execução orçamentária, financeira e patrimonial do Governo Federal. Disponível em: <http://www.tesouro.fazenda.gov.br/siafi>. Acesso em: 2 out. 2016.
[15] Disponível em: <http://www.portaltransparencia.gov.br/>. Acesso em: 2 out. 2016.

Figura 2 – Aspectos conceituais do Big Data.




Fonte: a autora.

### 2.2.1 Aspectos Tecnológicos do *big data*

Os sistemas de gerenciamento de bancos de dados (SGBD) adequados para o processamento de grandes quantidades de dados não são os tradicionais (MySQL, PostgreSQL, Oracle, SQLServer etc.), até mesmo em função do custo de armazenamento, como será demonstrado do Quadro 4. A respeito do assunto, Davenport (2014, p. 113) argumenta que "esses dados volumosos não podem ser bem manipulados por um *software* de banco de dados tradicional ou com servidores individuais [...] dessa forma uma nova geração de *software* de processamento de dados foi desenvolvida para resolver esse problema".

A Google lançou o *framework* MapReduce que distribui o processamento de dados por um grande nó de computadores interligados. Na sequência, a Yahoo lançou o Hadoop, uma plataforma de *software* em Java voltada para *clusters* e processamento de grandes massas de dados.

O Hadoop é um projeto de *software livre* desenvolvido pela Apache Software Foundation, por esse motivo as vezes é chamado de Apache Hadoop. A plataforma de

computação distribuída do *software* Hadoop é em Java, voltada para *clusters* e processamento de grandes massas de dados. Possui alta escalabilidade, grande confiabilidade e tolerância a falhas. Para Davenport (2014, p. 58) "O Hadoop é um ambiente de armazenamento e processamento de *big data* unificado em vários servidores". De acordo com o autor, "um *cluster* Hadoop com cinquenta nós e oitocentos núcleos de processamento é capaz de processar 1 *petabyte* de dados" (DAVENPORT, 2014, p. 59). No que diz respeito ao custo de armazenamento de dados, um *terabyte* armazenado em sistema de gerenciamento de banco de dados relacional tem um custo de US$37 milhões, enquanto o mesmo volume de armazenamento em um *cluster* Haddop é armazenado a um valor de apenas US$ 2 milhões. A respeito da plataforma, Chechia (2013) comenta que os maiores colaboradores para o seu aprimoramento são o Facebook, a Google, o Yahoo e a IBM.

Davenport (2014, p. 111) argumenta que o big data é "mais que apenas um grande volume de dados não estruturados. Ele também inclui as tecnologias que possibilitam seu processamento e análise". No intuito de expor as tecnologias utilizadas no Big Data, o autor elaborou uma síntese, conforme demonstra o Quadro 2.

Quadro 2 – Visão geral das tecnologias de big data.

| **Tecnologia** | **Definição** |
|---|---|
| Hadoop | *Software* de código aberto para o processamento de *big data* em uma série de servidores paralelos. |
| MapReduce | Um *framework* arquitetônico no qual o Hadoop se baseia |
| Linguagens de Script | Linguagens de programação adequadas ao *big data* (por exemplo, Python, Pig, Hive) |
| Aprendizado de Máquina | *Software* para identificar rapidamente o modelo mais adequado ao conjunto de dados |
| Visual Analytics | Apresentação dos resultados analíticos em formatos visuais ou gráficos. |
| Processamento de Linguagem Natural (PLN) | *Software* para análise de texto – frequências, sentido etc. |
| In-memory analytics | Processamento de *big data* na memória do computador para obter mais velocidade. |

Fonte: Davenport (2014, p. 112).

Ainda no que diz respeito às tecnologias de informação e o ambiente big data, merece ser destacada uma nova modalidade de ataque cibernético – o *ransomware*, um determinado tipo de *malware* que bloqueia um computador infectado e sequestra os dados de determinada organização. Os responsáveis pelo ataque devolvem os dados mediante pagamento (*ransom*). McDermott (2015) se apoia em Eric Geier da PC World para afirmar que existem três níveis de *ransomware*: *scareware*, *lock-screen viruses* e, o pior de todos, *encrypting malware*. A autora também comenta que uma das formas de se pagar o sequestrados dos dados é por meio da

moeda *bitcoin*[16]. Nesse caso, a empresa que sofreu o ataque pode utilizar os serviços da empresa de segurança KnowBe4 (knowbe4.com), de Clearwater – Flórida, que mantém uma carteira de *bitcoin* com a finalidade de ajudar indivíduos e empresas a comprarem dados sequestrado de volta.

Retomando as colocações de Hey e Hey (2006) de que os cientistas vão precisar de novos mecanismos de buscas, novas ferramentas de mineração de dados especializadas e que criarão repositórios digitais de dados científicos, faz-se necessário, analisar esse cenário no contexto brasileiro.

Por todo o exposto, deve-se refletir sobre o sucateamento pelo qual muitas unidades de pesquisa têm passado no Brasil. Se o pesquisador tem dificuldades para obter financiamento para a execução da pesquisa, a realidade da organização a qual o pesquisador está vinculado não é diferente. O orçamento necessário para que as unidades de Tecnologia da Informação se preparem para tratar o volume de dados imposto pela *e-science* requer um alto investimento. Além disso, como se discutiu acima, novas tecnologias estão sendo adotadas para facilitar o processamento desse grande volume de dados. Logo, há que se investir tanto na compra de equipamentos e, por vezes, em licenças de *software*, como na capacitação do profissional de tecnologia da informação. Por fim, não se pode negligenciar o fato dos dados serem um ativo institucional, portanto precisam passar pelo ciclo de preservação de dados de longo prazo, optando pelo mais adequado, isto é, o proposto pelo Modelo OAIS[17], ou pelo quinto estágio do ciclo de vida proposto pelo Projeto DataONE[18], ou mesmo, um modelo *customizado* adequado a realidade da instituição, o que novamente requer investimento em capacitação dos diversos profissionais envolvidos.

Sem dúvida nenhuma, o importante é dispor de um ambiente em que o dado seja preservado de forma a ser reutilizado em pesquisas futuras. Mas como oferecer esse ambiente sem a infraestrutura tecnológica adequada e sem a capacitação profissional necessária para viabilizar a gestão dos dados de pesquisa? Os próprios pesquisadores precisam entender a importância da preservação de longo prazo do dado produzido pela sua pesquisa, para a partir

---

[16] *Bitcoin* foi classificada pelo Tesouro dos Estados Unidos como primeira moeda digital descentralizada do mundo.

[17] Open Archival Information System (OAIS) foi traduzido para o português como Sistema Aberto de Arquivamento de Informação (SAAI). É um modelo de referência para arquivamento da informação digital. Fornece orientações para a preservação e manutenção no acesso às informações por longo prazo. Em 2003 tornou-se uma norma ISO 14721:2003. Ressalta-se que para Borgman (2015) o Modelo OAIS apresenta um consenso sobre a prática originada na comunidade de Ciências Espaciais para tratamento e arquivamento de dados por longo prazo.

[18] Projeto Data Observation Network for Earth (DataONE), tem envidado esforços para a preservação digital de dados de pesquisa. Desenvolveu o Modelo de Ciclo de Vida do dado na perspectiva do pesquisador, onde o seu quinto estágio refere-se a submissão dos dados para arquivamento de longo prazo.

daí sensibilizarem o alto nível estratégico das instituições de pesquisa e assim obterem apoio financeiro e institucional para os projetos em questão.

É possível afirmar que enquanto as tecnologias digitais permitem que os dados de científicos sejam criados, manipulados, disseminados, recuperados e armazenados com uma facilidade cada vez maior; a preservação de longo prazo dos conjuntos de dados produzidos pela *e-science* (*datasets*), apresentam desafios significativos. A menos que estratégias de preservação de dados sejam empregadas tempestivamente, esses dados tendem a se tornar inacessíveis muito rapidamente. O profissional que terá sob sua responsabilidade a gestão desse dado, seja ele o pesquisador, ou o profissional da informação, ou o cientista de dados, deverá estar atento para selecionar um método de tratamento e preservação que observe a natureza do material (*dados*) produzido, pois é a natureza desse material que revelará quais aspectos precisam ser conservados.

No presente, muitas das ações ligadas a biblioteca e/ou repositório digital envolvem a digitalização do material existente, como, por exemplo, livros e fotografias. Infelizmente, poucos projetos dessas bibliotecas digitais consideram a preservação além da digitalização inicial. A ação de copiar a informação sem alterá-la oferece uma solução de curto prazo para a preservação do acesso aos objetos digitais. Isto faz com que a informação seja armazenada em uma nova mídia antes que a mídia antiga se deteriore. Porém, em longo prazo, essa simples migração nem sempre funciona. Aqui entra, portanto, a necessidade de implantar uma política de preservação digital de longo prazo que leve em consideração todos os outros aspectos relacionados com a informação digital, bem como aspectos relacionados aos dados de pesquisa produzidos em grande escala.

Merece ser ressaltado que se as bibliotecas brasileiras ainda têm uma atuação tímida e isolada na preservação de documentos. A exceção parece ser a iniciativas da Rede Brasileira de Serviços de Preservação Digital (Cariniana), liderada pelo IBICT a partir de 2002, com o objetivo de preservar documentos eletrônicos. Dentre as atividades de destaque da Cariniana tem-se a adesão ao Programa LOCKSS da Universidade de Stanford. Em 2016, a Cariniana passou a oferecer para os membros da Rede o serviço de preservação de dados de pesquisa, utilizando como repositório de dados científicos o *software* Dataverse.

O repositório da Cariniana está direcionado para os pesquisadores do IBICT; de instituições parceiras da Rede Cariniana; da rede de pesquisa Dríade e para os periódicos na plataforma OJS/SEER. (REDE CARINIANA, 2017)

No que diz respeito à realidade brasileira sobre a preservação de dados científicos (dados brutos), faz-se necessário que os profissionais da informação estejam atentos às mudanças de

necessidades de informação do usuário de forma a preencher esse espaço profissional, caso contrário, corre-se o risco das atividades de curadoria, preservação e outros aspectos inerentes ao tratamento desse grande volume de dados, virem a ser realizadas pelos especialistas de tecnologia da informação, ou ainda por uma nova categoria de profissional que nasceu para atender as demandas do *big data* – o cientista de dados.

## 2.3 A *E-SCIENCE*

Dentre as mudanças que vem ocorrendo na condução da ciência contemporânea destacam-se: a colaboração *on-line* da ciência, a infraestrutura de tecnologia de informação necessária para suportar a colaboração *on-line* de pesquisadores, o grande volume de dados produzido, o benefício da computação em *grid* que apoia a análise de um grande volume de dados, dentre outros (JANKOWSKI, 2007).

A contemporaneidade do tema traz à tona questões conceituais que ainda não passaram pelo processo de reflexão necessário ao seu amadurecimento. Por exemplo, merece ser comentado que dentre as denominações utilizadas para *e-science*, também aparecem na literatura os termos relacionados: ciência orientada por dados (*data-driven science*), computação fortemente orientada a dados (*data-intensive computing*), ciberinfraestrutura (*cyberinfrastructure*), ciência com uso intensivo em dados, quarto paradigma da ciência (*fourth paradigm of science*), dilúvio de dados (*data deluge)*, E-infraestrutura (*E-infrastructure)*, dentre outros (ALVARO *et al*, 2011; CÉSAR JÚNIOR, 2011; GRAY, 2007; HEY; TREFETHEN, 2003; MARCUM; GEORGE, 2010).

No âmbito da Ciência da Informação, essa constatação já havia sido feita por Medeiros e Caregnato (2012). Os autores relataram que o termo *e-science* aparece na literatura com grafias diferenciadas e que por vezes recebem o nome de *cyberinfrastructure*, *cyberscience*, *eInfrastructure* e *eResearch*.

O importante é compreender que as modificações que veem ocorrendo no processo de desenvolvimento científico e tecnológico, bem como as questões conceituais que emergem nesse cenário trazem à tona um novo paradigma na forma de se fazer ciência – *o quarto paradigma*, baseado na computação intensiva de dados.

Dentre os termos acima mencionados, destacam-se *e-science* e *cyberinfrastructure* que surgem como termos relacionados e, por vezes, aparentemente sinônimos, contudo em países com iniciativas distintas, mas com objetivos comuns – obter financiamento para pesquisas que

envolvem o tratamento de um grande volume de dados. Por esse motivo, faz-se necessário contextualizar o surgimento de ambos respectivamente.

De acordo com Jankowski (2007) o termo *cyberinfrastructure* está enraizado com a iniciativa, dos cientistas americanos, no ano de 2003, em obter fontes de financiamento na National Science Foundation. Essa iniciativa culminou, em 2003, com a publicação do Atkins Report – intitulado *Revolutioninzing Science and Engineering Throught Cyberinfrastructure*. Nas palavras do relatório: "[...] se a infraestrutura é necessária para a economia industrial, então [...] *cyberinfrastructure* é necessária para a economia do conhecimento" (ATKINS *et al.*, 2003, p. 5).

Na percepção de Jankowski (2007), o relatório, por utilizar uma linguagem promocional e visionária em todo o documento, ficou associado à ideia de obter fonte de financiamento pela NSF. Dentre as expressões utilizadas no relatório têm-se – "uma nova era raiou" (ATKINS *et al*., 2003, p. 31); "o tempo está maduro" (ATKINS *et al*., 2003, p. 12) e "uma oportunidade única em uma geração a liderar a revolução" (ATKINS *et al*., 2003, p. 32). Merece ser comentado que Jankowski (2007) não aparenta surpresa ao constatar que as primeiras iniciativas de *Cyberinfrastructure* no Relatório Atkins tenham privilegiado as ciências naturais e biológicas. O autor (JANKOWSKI, 2007, p. 550), demonstra compreensão para com a escolha, afinal são áreas "onde grandes volumes de dados estão envolvidas nos esforços de investigação que necessitam de alta velocidade de processamento do computador: [tais como] a física de partículas, a astronomia, a meteorologia e a pesquisa de DNA". Para Jankowski (2007), estas iniciativas normalmente envolvem colaboração com as equipes de supercomputação dos centros de pesquisa.

Já o termo *e-science* surge de iniciativas europeias, especialmente no Reino Unido, onde John Taylor – diretor geral de Conselhos de Pesquisa do Escritório de Ciência e Tecnologia do Reino Unido cunha o termo em 1999[19], durante o lançamento de um programa de financiamento para pesquisas. A exemplo dos EUA, o Escritório de Ciência e Tecnologia do Reino Unido teve como foco foi as ciências naturais e biológicas, sendo projetado para processar grandes volumes de dados com a ajuda da computação em *grid*. Em 2001 foi criado o National e-Science Centre, tornando-se o principal órgão de coordenação para financiamento de projetos de *e-science* (JANKOWSKI, 2007).

Na percepção Jankowski (2007) uma das diferenças entre os projetos de *e-science* do Reino Unido para com os de trajetória americana, é o fato de o Reino Unido ter investido em

[19] Há referências de que o termo *e-science* foi criado no ano 2000, dentre elas Gray (2007).

um escritório patrocinado pelo governo para estimular e coordenar iniciativas de *e-science* nas ciências sociais. Em dezembro de 2004 foi lançado o National Centre for e-Social Science com uma estrutura descentralizada de nós que integram as universidades do Reino Unido. A maioria dos projetos financiados até agora [2007] seguem o paradigma da *e-science* com arquitetura de computação em *grid*. A única exceção é o nó Universidade de Oxford, que tem uma abordagem de conformação social. Inicialmente, os 11 projetos-piloto recebem apoio para explorar a aplicação de tecnologias de computação em *grid* nas ciências sociais.

Na perspectiva de Hey e Trefethen (2003), Gray (2007) e Marcum e George (2010) a *e-science* faz referência à coleção de instrumentos e tecnologias necessárias para apoiar a pesquisa científica do Século XXI, e amparar o grande volume de dados produzidos que precisam estar em rede, com a característica intrínseca da colaboração e da multidisciplinaridade.

Lippincott (2010, p. 63) argumenta que o termo *e-science* é utilizado pelo Reino Unido para "descrever a emergência da ciência orientada a dados (*data-drive science*)". Em contrapartida, os britânicos utilizam o termo *E-infrastructure* para se referir à infraestrutura de computação distribuída que fornece acesso compartilhado a grandes coleções de dados, as ferramentas de computação avançadas para análise desses dados, bem como, aos recursos de computação de grande escala e visualização de alta performance. O autor comenta que o conceito é similar ao de *Cyberinfrastructure* utilizado pela National Science Foundation (NSF). Face ao exposto, pode-se afirmar que Lippincott (2010) está em consonância com Hey e Trefethen (2005) que utilizaram o termo *E-infrastructure* para se referir as iniciativas no Reino Unido, esclarecendo que o termo *Cyberinfrastructure*, utilizado no Relatório Atkins (ATKINS *et al*., 2003), refere-se ao conjunto de ações da NSF.

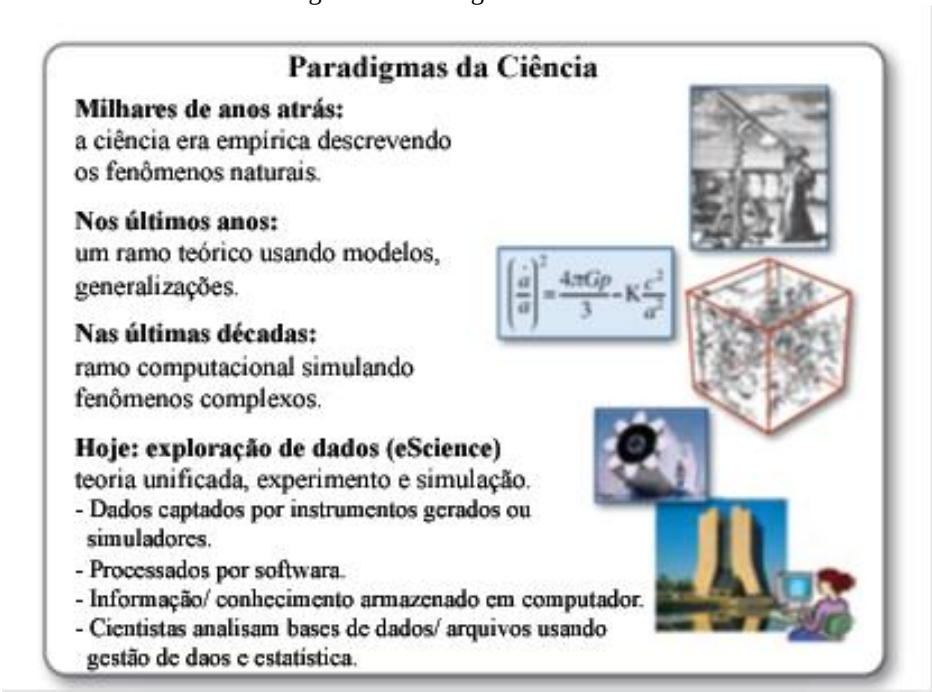
A análise da literatura consultada demonstra que, em ambos os casos, a utilização dos respectivos termos, afloram no contexto da obtenção de fontes de financiamento de pesquisas, sejam elas nos EUA ou no Reino Unido.

Nesse cenário, Gray[20] (2007) argumenta que a *e-science* é o ponto onde a tecnologia da informação encontra os cientistas. O autor explica sua perspectiva ao expor que a captação de dados de pesquisa é realizada por instrumentos (satélites, telescópios, sensores) ou é gerada por máquinas de simulação. Esses dados coletados, ou obtidos por meio de simulação, são processados por um *software* que por sua vez providenciará o armazenamento da informação

[20] Jim Gray foi vencedor do Prêmio Turing de 1998. É considerado um dos pioneiros em aplicações e técnicas computacionais para o tratamento de grandes quantidades de dados gerados por cientistas de outras áreas (CORDEIRO *et al*., 2013).

em bancos de dados. Gray (2007), ao comentar que um telescópio é operado por 20 a 50 pessoas e que há milhares de pessoas escrevendo códigos para lidar com a informação coletada pelo instrumento, utiliza o campo da astronomia para defender a sua tese. A Figura 3 ilustra os paradigmas da ciência na visão de Gray (2007), bem como retrata a evolução no processo de coleta de dados que culminou com a chamada *e-science* ou dilúvio de dados.

Figura 3 – Paradigmas da ciência.




Fonte: Gray (2007, tradução nossa).

A partir do exposto, Gray (2007) conclui que a tecnologia da informação evoluiu quanto ao desenvolvimento de sistema para coletar dados e realizar simulações. Por outro lado, não evoluiu tanto em ferramentas de análise de dados. Na visão do autor, [em 2007] o cientista e o pesquisador [tinham] dificuldades para codificar suas informações e compartilhá-las com seus pares. Além disso, ele expõe que os cientistas da *base científica da pirâmide*[21] [possuíam] no máximo o *software* MAGLAB e o Excel para apoiá-los na análise dos dados.

Na perspectiva de Jankowski (2007) *e-science* é um termo guarda-chuva que inclui muitas das características comumente associadas à forma como a pesquisa é conduzida em um

[21] Gray (2007) classifica os cientistas em três grupos. Os de nível 1 (topo da pirâmide) participam de projetos organizados e gerenciados de maneira sistemática. São grandes projetos que podem se dar ao luxo de ter um orçamento de *hardware* e *software*, bem como uma equipe de cientistas para escrever *softwares* sob medida para o experimento, dentre os exemplos dessa categoria estão os projetos do observatório oceânico EUA-Canadá, o projeto do grande colisor de *hádrons* realizado pelo CERN, normalmente custeado por agências de fomento como a National Science Foundation (NSF). Por outro lado, no caso dos cientistas da base da pirâmide, o nível 3, os recursos são obtidos pelos próprios pesquisadores que levam consigo a sua própria fonte de custeio, tendo acesso apenas a *softwares* de bancada para apoiar a análise de dados, dentre eles o MAGLAB e o Excel.

ambiente em rede – utilizando ferramentas baseadas na internet, envolvendo a colaboração entre os pesquisadores por vezes separados por grandes distâncias em uma escala global. Além disso, o termo é utilizado para se referir as iniciativas de computação em *grid*, colaboração global de pesquisadores e internet baseada em instrumentos. São esses dados, produzidos por esses instrumentos que precisam passar por um processo de curadoria, armazenamento, divulgação, reutilização e preservação digital.

## 2.4 DADOS CIENTÍFICOS / DADOS DE PESQUISA

Para Hey e Trefethen (2005, p. 819), "a necessidade de auxílio para organização, registro e pesquisa dos dados está se tornando aguda". Sobre o tema, Borgman (2015, p. 4, grifo do autor) argumenta que "a questão não declarada a fazer é: o que são dados?". Para a autora, "o único consenso sobre as diferentes definições é que nenhuma definição única será suficiente para definir o termo, uma vez que eles têm muitos tipos de valor, sendo que valor dos dados pode não ser aparente até muito tempo depois dos mesmos terem sido coletados, criados ou mesmo perdidos". Dando continuidade ao assunto, a autora defende que valor dos dados varia muito ao longo do tempo, lugar e contexto. Além disso, enfatiza que ter os dados corretos é geralmente melhor do que ter mais dados. Por outro lado, é importante destacar que os dados não têm nenhum valor ou significado quando estão isolados.

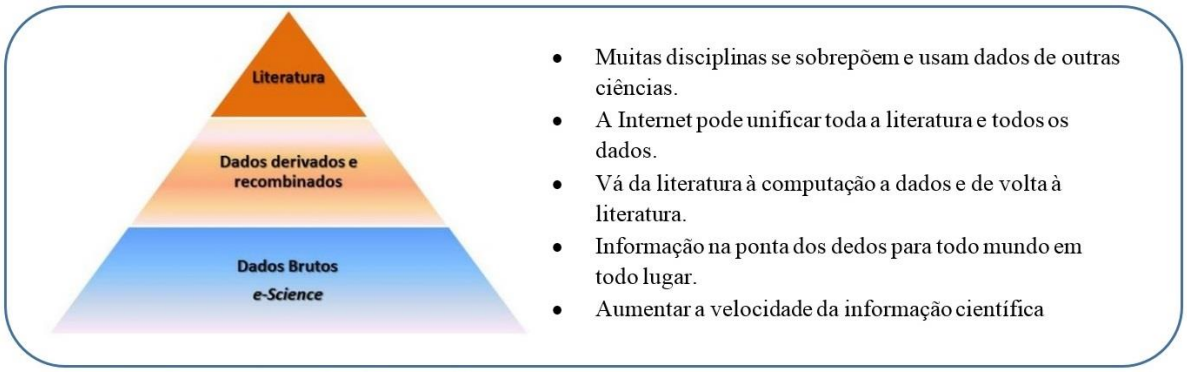
Nesse sentido, Borgman (2015, p. 17) defende que conceituar o termo dado não é algo trivial e aponta que a proposta de Machlup e Mansfield (1983) de dividir em três partes – dado, informação e conhecimento simplifica as relações complexas entre esses conceitos. A autora também recorda a colocação de Meadows (2001) de que - "o que nós consideramos ser dados básicos tem sempre um elemento de arbitrariedade nele".

Davenport, na área de Administração, em 2001, apresentou a diferença entre *dado, informação e conhecimento*. Esses conceitos foram exaustivamente trabalhados e discutidos na famosa pirâmide informacional, predominantemente nas áreas de administração, gestão do conhecimento, inteligência competitiva, dentre outras. Essa discussão em torno do termo *dado,* de certa forma, já demonstra sua complexidade.

Dentro do contexto empresarial, o autor defende que o dado é uma "simples observação sobre o estado do mundo" (DAVENPORT, 2001), além disso, apresenta como características do dado o fato dele ser facilmente estruturado, obtido por máquinas, ser frequentemente quantificado e facilmente transferível.

Aproximadamente seis anos depois, com a explosão de dados produzidos e transmitidos por máquinas nos ambientes de pesquisa, Gray (2007, p. 35) propõe uma nova pirâmide informacional, trazendo mais uma vez o *dado* na base da pirâmide e a literatura no topo. A questão é que Gray (2007) tem como projeto deixar todos os dados de pesquisa *online*, para dessa forma contribuir com o desenvolvimento da ciência de forma mais célere. A proposta de Gray (2007) é ilustrada na Figura 4.

Figura 4 – Todos os dados científicos *online*.




Fonte: Gray (2007, p. 25).

Na perspectiva de Borgman (2015, p. 17), pela sua complexidade, o termo dado, por si só, é digno de um livro. A autora (BORGMAN, 2015, grifo do autor) defende que "a questão o que é dado é melhor abordada como quando são dados". Para tanto, a autora se fundamenta em definições sobre dado do Dicionário de Oxford de 1646**,** que traz o uso do termo na teologia, bem como, no estudo de Rosenberg, sobre o termo dado, no Século XVIII. Além disso, a autora relembra que diversos autores (BLAIR, 2010; BROWN; DUGUID, 2009; BUCKLAND, 1991; BURKE, 2000; 2012; DAY, 2001; INGWERSEN; JAVELIN, 2005; LIU, 2004; MEADOWS, 2001) e outros autores da ciência da informação já discutiram sobre o fato do dado ser uma forma de informação.

Ao abordar os diferentes tipos de dados, Simberloff *et al.* (2005, p. 18) fazem uma metáfora com o universo financeiro ao argumentar que "assim como a moeda na esfera financeira assume diferentes formas, o dado digital também assume diferentes formas no universo de coleção de dados". Os autores (SIMBERLOFF *et al*., 2005, p. 18, grifo nosso) vão além, defendendo que as diferenças dos dados incluem a natureza do mesmo, sua reprodutibilidade, bem como, o nível de processamento ao qual o dado é submetido. Na percepção de Simberloff *et al.* (2005, p. 18) "cada uma dessas diferenças traz importantes implicações políticas".

Simberloff *et al.* (2005, p. 18-19) argumentam que o dado, quanto à sua natureza, em uma coleção, pode ser diverso. Dentre os exemplos citam: números, imagens, vídeo, arquivos de áudio, *software*, informações sobre a versão de um *software*, equações, animações, algoritmos, ou mesmo, modelos/ simulações. Os autores também argumentam que os dados podem ser diferenciados em função das suas origens. Nesse aspecto, eles podem ser "*observacionais, computacionais ou experimentais".* Além disso, eles enfatizam que esta distinção é fundamental para as escolhas a serem feitas sobre o arquivamento e preservação digital desses dados.

A questão que está em plena ebulição para os gestores de informação, cientistas de dados, assim como para os pesquisadores é: *quais dados devem ser armazenados e por quanto tempo*. Sem sombra de dúvida, a Arquivologia traz importantes contribuições nesse aspecto, pois já é tarefa rotineira para esses profissionais a elaboração de Tabelas de Temporalidade Documentais no âmbito de documentos orgânicos de origem primária e secundária.

Pois bem, no que diz respeito a essas questões no âmbito dos dados de pesquisa coletados em grande escala, na percepção de Simberloff *et al*. (2005, p. 19): "dados de observação, tais como observações diretas de temperatura do oceano em uma data específica, a atitude dos eleitores antes de uma eleição, ou fotografias [...] são registros históricos que não podem ser recoletados". Logo, para os autores, os dados observacionais são geralmente arquivados indefinidamente. Ou, utilizando a terminologia arquivista – *fariam parte do arquivo permanente*.

Dando continuidade ao assunto, Simberloff *et al*. (2005, p. 19): argumentam:

> [...] um diferente conjunto de considerações aplica-se aos dados computacionais, tais como os resultados da execução de um modelo pelo computador ou por uma simulação. Se a informação detalhada sobre o modelo (incluindo uma descrição completa do *hardware*, *software* e dados de entrada) está disponível, a preservação em um repositório [de dados] de longo prazo pode não ser necessária. Pois, os dados em questão podem ser reproduzidos. Assim, embora os resultados de um modelo possam não necessitar passar pelo processo de preservação, o arquivamento do próprio modelo e de um conjunto robusto de metadados pode ser essencial.

Já no que diz respeito aos dados experimentais, Simberloff *et al.* (2005, p. 19) defendem que "em princípio os dados de experimentos, que podem ser reproduzidos com precisão, não precisam ser armazenados por tempo indeterminado". Porém, os autores revelam que na prática, pode não ser possível reproduzir com precisão todas as condições experimentais, particularmente onde algumas condições e variáveis não podem ser conhecidas. Além disso, há situações, onde os custos de reprodução da experiência são proibitivos, nestes casos, em

específico, a preservação de longo prazo deve ser garantida para essa categoria de dados. Em síntese, Simberloff *et al.* (2005), ponderam que as questões de custo e capacidade de reprodutibilidade são a chave ao considerar políticas para a preservação de dados experimentais.

Fox e Harris (2013, p. 10) incluem em sua definição para dados os qualitativos e os estatísticos, conforme descrito a seguir:

> [...] inclui, no mínimo, observações digitais, acompanhamento científico, dados de sensores, metadados, cenários e modelos de saída, dados comportamentais observados ou qualitativos, visualizações e dados estatísticos coletados para fins administrativos e comerciais. Dado normalmente é visto como um *input* no processo de pesquisa.

A variedade na tipologia de dados exposta pelos autores acima mencionados (DAVENPORT, 2001; FOX; HARRIS, 2013; GRAY, 2007; MACHLUP; MANSFIELD, 1983; MEADOWS, 2001; SIMBERLOFF *et al*., 2005) corroboram a percepção de Borgman (2015) sobre a dificuldade de definir o que é dado.

A respeito do assunto, Sayão e Sales (2015, p. 7) argumentam que a "noção de dados pode variar consideravelmente entre pesquisadores e, ainda mais, entre áreas do conhecimento". Para explicar o ponto de vista, os autores teorizam "a constatação de que os dados são gerados para diferentes propósitos, por diferentes comunidades acadêmicas e científicas e por meio de diferentes processos intensifica ainda mais essa percepção de diversidade".

Em seu guia de pesquisa, Sayão e Sales (2015) propõem como formas de classificação de dados: a) quanto à sua origem (observacionais, computacionais e experimentais), b) quando a sua natureza e c) quanto à fase da sua pesquisa. Os autores se assemelham a proposta de Simberloff *et al*. (2005) nos itens a e b. Por outro lado, trazem uma proposta nova ao proporem uma classificação quanto à fase da pesquisa (dados brutos, crus ou preliminares; dados derivados; dados canônicos ou referenciais).

Do ponto de vista prático, pode-se dizer que o sensor que está implantado nas tartarugas do Projeto Tamar[22], ou mesmo o sensor que estava implantando no Leão Cecil[23], geram dados de biodiversidade. Quando se trata de informação georeferenciada, pode-se ter a latitude e a

[22] Um projeto iniciado em 1980 que representa uma das mais bem-sucedidas experiências de conservação marinha desenvolvidas no Brasil, serve de modelo para outros países.
[23] Leão africano que vivia no Parque Nacional de Hwange localizado no Zimbábue. Era monitorado por cientistas da Universidade de Oxford, no Reino Unido, que estudavam a longevidade e a conservação de leões no Zimbábue. O leão foi morto, aos 13 anos de idade, no ano de 2015, por turista americano em caçada de lazer, abrindo a discussão sobre esta prática e a sobrevivência de animais selvagens.

longitude indicando a posição de uma espécie de bromélia rara na Floresta Amazônica. Em ambos os casos os dados são armazenados em grandes bancos de dados.

Além dos sistemas acima citados, têm-se os dados geodésicos, os dados provenientes da área de energia nuclear, tais como os dados de monitoramento das simulações e operações de um reator nuclear, ou mesmo os dados sobre mudanças climáticas. Também são dados, aqueles produzidos por um laboratório e registrados manualmente em cadernos. Como, por exemplo, os dados produzidos pelo Laboratório de Membranas Poliméricas do Instituto de Energia Nuclear.

É preciso explorar semelhanças e diferenças na forma como os dados são criados, utilizados e compreendidos nas comunidades acadêmicas (BORGMAN, 2015). A partir da colocação da autora, é pertinente pensar no caso de um dado, como, por exemplo, da Tartaruga do Projeto Tamar coletado pelo sensor. *Quem é o autor desse dado? É o líder do projeto de pesquisa? É o pesquisador responsável pelo monitoramento daquela tartaruga específica? Como a comunidade acadêmica entende esses dados?* Identificar o autor[24] do dado traz à tona a resposta de como citar[25] o dado. Identificar a forma como o dado foi gerado permite classificá-lo de acordo com as propostas já existentes, como, por exemplo, a de Simberloff *et al*. (2005). Ao se classificar esse dado, por consequência, sabe-se o período de temporalidade dele no repositório de dados. Essa cadeia de atividades alimenta o ciclo de gestão de dados de pesquisa que será abordado a seguir, no tópico sobre preservação dos dados. Outras reflexões que impactam na gestão de dados de pesquisa são ilustradas na Figura 5.

---

[24] A discussão sobre autoria de dado é complexa. Borgman (2015, p. 253) pondera que autoria e outras formas de responsabilidade são convenções sociais. Essas convenções variam por pessoa, equipe, comunidade ou foro de publicação ao longo do tempo. Borgman se apoia em Wuchty, Jones e Uzzi (2007) para afirmar que até meados da década de 1950, a maioria das publicações acadêmicas possuíam autoria individual. Porém, conforme o número de autores por publicação cresceu, a responsabilidade pelas publicações se tornou mais difusa. Também se apoia em Davenport e Cronin (2001) e King (2014) ao argumentar que no final da década de 1990, artigos frequentemente possuíam vários autores e algumas vezes centenas de autores indicando que o percentual de autoria individual está declinando. Para a autora (BORGMAN, 2015, 251/257) a discussão sobre autoria retoma velhos e antigos debates sobre a responsabilidade de ideias ou documentos. As noções de autoria individual e coletiva variaram ao longo dos séculos, de acordo com a cultura e o contexto. Os colaboradores negociam quem será denominado autor em cada publicação e em qual ordem. A autoria pode ou não, ser atribuída a quem descreveu a narrativa, a quem coletou os dados, quem compilou a bibliografia, quem analisou os dados, ou quem contruiu os instrumentos de coleta dos dados. Em campos como a física de alta energia, a autoria pode ser coletiva. Por exemplo, o primeiro artigo sobre o bóson de Higgs, a instituição CERN identificou os autores em um *Atlas de Colaboração* e os listou em uma relação onde constaram 2.932 nomes. Essa atitude permaneceu por um determinado período de tempo, de forma a garartir que àqueles que contribuíram com os estágios iniciais da pesquisa, pudessem receber os devidos créditos pelas descobertas, em consequência disso, há casos de autoria póstuma. Na percepção de Borgamn (2015, p´. 257) a dissertação de Jillian Wallis (2012) é o mais completo estudo para datar as questões de autoria e responsabilidade pela gestão de dados, embora focado em um centro de pesquisa – o Center for Embedded Network Sensing (CENS).

[25] De acordo com Borgman (2015, p. 251/252) os dados raramente são citados. A autora argumenta que determinar o que constitui uma citação de dados ou o uso em si dos dados, é complexo. Estudos sobre citação de dados indicam que apenas uma pequena porcentagem de artigos incluem a citação de dados nas referencias bibliográficas ou notas de rodapé, apesar disso há indícios de que o número de citações vem aumentando em anos mais recentes. Borgman (2015) comenta que algumas áreas do conhecimento publicam *papers* de dados e sobre os instrumentos para dar o devido crédito para as contribuições específicas. Em outros casos, autores citam os próprios *papers* para se remeter aos dados contidos nele.

Figura 5 – Reflexões sobre a gestão de dados científicos.


Fonte: A autora.

Dentre essas reflexões, merece ser analisado com cuidado, do ponto de vista da tecnologia da informação as tecnologias necessárias para armazenar dados oriundos da *e-science*. É preciso analisar se a área de tecnologia da informação da instituição possui a infraestrutura tecnológica necessária, se há profissionais capacitados implementar as rotinas de *back-up* de dados e até mesmo restauração, caso seja necessário. Esses aspectos serão discutidos em mais profundidade no tópico sobre a preservação de dados.

No que diz respeito as expressões "*dados científicos*" e "*dados de pesquisa*" cabe ressaltar que ainda não há um consenso na literatura quanto ao uso das mesmas. Os autores Hey e Hey (2006), Rodrigues *et al* (2010), Bell (2011) e Príncipe *et al.* (2014) utilizam *dados científicos*. Por outro lado, Sales (2014), Sayão e Sales (2014) e Borgman (2015) utilizam o termo *dados de pesquisa* (*data scholarship*).

No que diz respeito aos *dados de pesquisa*, Borgman (2015) entende que é um termo cunhado para situar o conjunto complexo de relações entre dados e pesquisa. Os dados assumiram uma vida própria, ou pelo menos assim parece quando vistos a partir da imprensa popular, independente do contexto acadêmico em que eles são usados como evidência de alguns fenômenos. A autora considera que acadêmicos, estudantes e analistas de negócios parecem agora reconhecer que têm dados suficientes e as técnicas corretas para explorá-los, o que permite que novas perguntas sejam feitas, assim como novas formas de se obter evidências. Na percepção de Borgman (2015), algumas das coisas que podem ser feitas com os dados são

extremamente valiosas. No entanto, pode ser muito difícil determinar o quão valioso qualquer conjunto de dados pode ser, ou as maneiras em que eles assumem determinado valor.

Em síntese, Borgman (2015) quis dizer que o dado pode não ter valor aparente até ser perdido; ou ainda, pode ganhar valor quando associado a outro conjunto de dados, ou mesmo, no decorrer do tempo.

### 2.4.1 Ciclo de vida e preservação dos dados científicos

Na perspectiva da organização Blue Ribbon Task Force (2010, p. 6), a preservação digital possui quatro grandes contextos: o educacional (*scholarly discourse*), os dados de pesquisa (*research data*), os conteúdos de internet (*collectively produced web content*) e o conteúdo comercial e cultural (*commercially owned cultural content*).

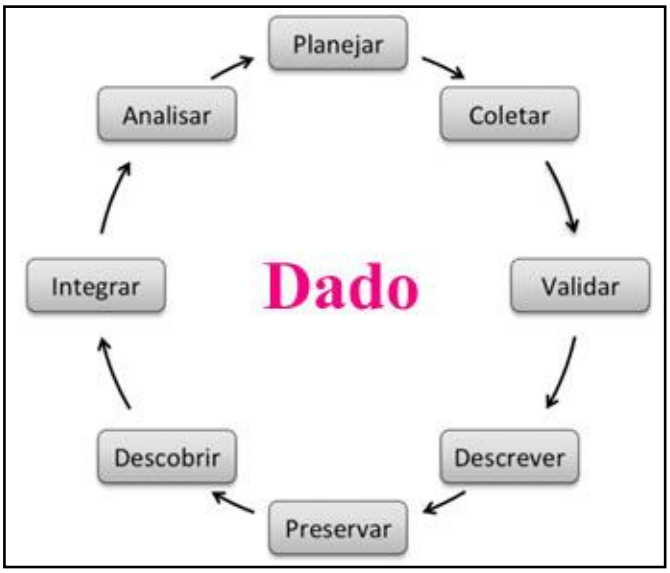
A preservação digital pode estar voltada para a digitalização de documentos em formatos não digitais, ou mesmo voltada para a recuperação de objetos digitais que já se tornaram obsoletos, como, por exemplo, o disquete. O importante é ter a compreensão de que a preservação envolve o uso de técnicas (por exemplo, migração, emulação, espelhamento) e a aplicação de políticas e de gestão de um projeto que tenha como objetivo dar acesso àqueles objetos de modo que eles permaneçam confiáveis, acessíveis e disponíveis para uso ao longo do tempo para quem deles precisar.

No que diz respeito à preservação de dados científicos brutos oriundos da *e-science*, esses dados já nascem digitais, são produzidos por equipamentos específicos (satélites, sensores etc.) e, em grande escala, trazendo peculiaridades específicas para o seu tratamento e, consequentemente, para a preservação. A respeito do assunto, Hey e Hey (2006, p. 515) comentam que:

> A fim de explorar os muitos *petabytes* de dados científicos que surgirão a partir dos experimentos científicos de última geração, tais como as simulações em supercomputadores, as redes de sensores e os levantamentos feitos por satélite; os cientistas necessitarão do auxílio de motores de busca especializados e de poderosas ferramentas de mineração de dados. Para criar essas ferramentas, os dados primários deverão ser registrados com os seus metadados relevantes de forma a ter algumas informações quanto à proveniência, o conteúdo e as condições em que os dados foram produzidos. Ao longo dos próximos anos, os cientistas criarão vastos repositórios digitais de dados científicos, o que exigirá serviços de gestão semelhantes aos das bibliotecas digitais mais convencionais, bem como outros serviços específicos de dados.

Nos Estados Unidos, o Projeto Data Observation Network for Earth (DataONE), da National Science Foundation, com sede em Albuquerque, Novo México, tem envidado esforços para a preservação digital de dados de pesquisa. O projeto DataONE tem uma missão ambiciosa: “Fornecer o acesso universal aos dados sobre a vida na Terra e o ambiente que o sustenta, bem como as ferramentas que os pesquisadores necessitam para tanto” (DataONE, 2016). Assim, o DataONE tem desenvolvido um *framework* distribuído e uma ciberinfraestrutura sustentável que atenda às necessidades da ciência aberta. A iniciativa vai ao encontro do movimento de ciência aberta e acata a diretriz do Governo Americano de aumentar o acesso aos resultados da investigação científica financiada pelo governo federal, conforme ilustra a Figura 6.

Figura 6 – Ciclo de Vida do Dado na perspectiva do pesquisador.




Fonte: DataONE (2016).

Na visão do DataONE o dado tem vida própria. A Figura 6 ilustra as etapas de sua criação e utilização. A gestão do dado começa quando o pesquisador ainda está planejando sua etapa de coleta. Os próximos três estágios (coletar, validar, descrever) são a base para o acesso a longo prazo do dado. Enquanto isso, os três últimos representam a descoberta e o uso dos dados.

Nesse sentido, o Projeto DataONE divulgou uma cartilha – *Primer on Data Management: What you Always wanted to Know*. A cartilha descreve algumas práticas de gestão de dados fundamentais, trazendo contribuições para desenvolver um plano de gestão de dados, bem como contribuições para criar o dado de forma eficaz, organiza-lo, gerenciá-lo, descrevê-lo, preservá-lo e compartilhá-lo, conforme retratado no Quadro 3.

Quadro 3 – Uma visão geral do Ciclo de Vida DataONE.

| Atividade | Descrição |
|---|---|
| **Planejar** | Fase de descrição dos dados que serão compilados, e como os dados serão administrados e tornados acessíveis ao longo da sua vida útil. |
| **Coletar** | Observações são feitas à mão, ou com sensores, ou outros instrumentos e os dados são colocados em um formato digital. |
| **Assegurar** | A qualidade dos dados é assegurada por meio de controles e inspeções. |
| **Descrever** | Os dados são descritos com precisão e usando os padrões de metadados apropriados. |
| **Preservar** | Os dados são submetidos a um arquivamento de longo prazo adequado. |
| **Descobrir** | Dados potencialmente úteis estão localizados e são obtidos, junto com as informações relevantes sobre os dados. |
| **Integrar** | Dados de fontes diferentes são combinados para formar um conjunto homogêneo de dados que podem ser facilmente analisados |
| **Analisar** | Os dados são analisados. |

Fonte: Strasser *et al*. (2012).

A filosofia do modelo DataONE parte da pergunta – "se você compartilhar seus dados com um cientista ou colega que não está envolvido com seu projeto de pesquisa, eles estarão aptos a ver sentido no dado? Será que eles vão ser capazes de usá-lo de forma eficaz e adequadamente?".

Além do Projeto DataONE, merece ser comentado que na visão de Borgman (2015, p. 20) entre os princípios mais conhecidos para arquivamento de dados tem-se o documento *Reference Model for a Open Archivel Information System*[26] (OAIS). A autora comenta que este documento apresenta um consenso sobre a prática originada na comunidade de Ciências Espaciais para tratamento e arquivamento de dados. A autora observa que essas orientações também têm sido amplamente adotadas nas ciências e ciências sociais como diretrizes para o arquivamento de dados.

A respeito do Modelo OAIS, de acordo com Borgman (2015, p. 22), "ao definir dados, em termos gerais, o modelo usa o termo dados de forma transformadora – conjunto de dados, unidade de dados, formato de dados, banco de dados, objeto de dados, entidade de dados, e assim por diante". Dentre os exemplos, para a definição de dado tem-se:

[26] Consultative Committee for Space Data Systems.

> uma representação de múltiplas interpretações de informações de um modo organizado, adequado à comunicação, compilação ou processamento. Exemplos de dados incluem uma sequência de *bits*, uma tabela de números, os caracteres em uma página, a gravação dos sons feitos por uma pessoa ao falar, ou uma amostra de rocha coletada durante uma expedição na Lua.

Borgman (2015, p. 21) defende que "entre as categorias mais discretas dos dados, estão os níveis de processamento definidos pelo Sistema de Informação de Dados sobre a Terra da NASA". Nesse sistema, dados com uma origem comum se distinguem pela forma como eles são tratados, conforme demonstra o Quadro 4.

De acordo com a NASA (2016):

> Produtos de dados da EOSDIS[27] são processados em diversos níveis, variando do Nível 0 ao Nível 4. Os produtos de Nível 0 são dados brutos na maior resolução do instrumento. Em níveis mais elevados, os dados são convertidos em parâmetros e formatos mais úteis. Todos os instrumentos da EOS devem gerar produtos de Nível 1. A maior parte gera produtos de Nível 2 e 3, e muitos geram produtos de Nível 4.

Quadro 4 – Níveis de processamento de dados.[28]

| Nível do Dado | Descrição |
|---|---|
| **Nível 0** | Dados de instrumentos e de carga em resolução total, reconstruídos e não-processados, com qualquer e todos os artefatos de comunicação removidos (por exemplo, quadros de sincronização, cabeçalhos de comunicação, dados duplicados). (Na maioria dos casos, o Sistema de Operação de Dados EOS (EDOS) fornece esses dados para os Data Centers como conjuntos de dados de produção para processamento pelo Departamento de Ciência de Processamento de Dados ou por um SIPS (*Science Investigator-led Processing Systems* – Sistema de Processamento liderado por Investigador Científico) para produzir resultados de níveis superiores). |
| **Nível 1A** | Dados de instrumentos em resolução total, reconstruídos e não-processados, com referência ao tempo e com informações auxiliares anotadas, incluindo coeficientes de calibração geométricos e radiométricos e parâmetros de georeferenciamento (por exemplo, Plataforma Ephemeris), computados e anexados, mas não aplicados ao Nível 0 de dado[29] |
| **Nível 1B** | Dados no Nível 1A que foram processados por unidade do sensor (nem todos os instrumentos possuem dados de origem para o Nível 1B[30] |
| **Nível 2** | Dados derivados de variáveis geofísicas na mesma resolução e posição que os dados de origem para o Nível 1. |
| **Nível 3** | Variáveis mapeadas em grades de escala uniforme do espaço-tempo, geralmente com alguma integridade e consistência. |
| **Nível 4** | Modelos derivados ou resultados da análise de dados de níveis inferiores (por exemplo, variáveis derivadas de múltiplas medições). |

Fonte: Borgman (2015, p. 22); Feldman (2016), NASA (2016).
Nota: Tradução livre com fundamento nas fontes citadas.

[27] Earth Observing System Data and Information System – em português - *Sistema de Informação de Dados sobre a Terra*.
[28] O Quadro original pode ser visualizado no *site* da NASA. Disponível em: <http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/>.
[29] De acordo com Feldman (2016) o Nível 1A o nível de dados de arquivo é preferido pelos cientistas da NASA, pois se houver mudanças de calibração do sensor, os dados não precisam ser recoletados.
[30] De acordo com Feldman (2016) os dados de nível 1B são dados 1A Nível que tiveram calibrações de instrumentos / radiométricos aplicada.

Ao se analisar o ciclo de vida dos dados proposto pelo Projeto DataONE, bem como os níveis de processamento de dados Sistema de Informação de Dados sobre a Terra da NASA, percebe-se a complexidade do tratamento dos dados coletados pela *e-science* e, consequentemente, sua preservação. Nesse cenário, não é exagero afirmar que a formação do profissional da informação precisa ter pontos revistos à luz das novas necessidades de tratamento da informação pelos usuários. Além disso, a equipe multidisciplinar vai de forma gradativa se impondo para as bibliotecas que desejam enfrentar esse novo desafio.

No âmbito da preservação de dados de pesquisa, para Simberloff *et al.* (2005, p. 19) *"*inicialmente, os dados podem ser recolhidos na forma bruta, por exemplo, como um sinal digital gerado por um instrumento ou sensor. Estes dados não processados são frequentemente sujeitos a subsequentes etapas de refinamento e análise, dependendo dos objetivos da investigação". Logo, para o autor, o dado pode apresentar uma série de versões. Nesse sentido, Simberloff *et al.* (2005, p. 19), argumentam que "[...] embora os dados não processados possam não representar a forma mais completa, os dados derivados podem ser mais facilmente utilizáveis por outros [pesquisadores]". Assim, na visão do autor, a preservação dos dados em múltiplas formas pode ocorrer em muitas circunstâncias. (SIMBERLOFF *et al*., 2005, p. 19, grifo nosso).

Strasser (2015) defende que enfrentar os desafios inerentes da pesquisa do Século 21 exige uma boa gestão de dados de pesquisa (GDP). Para o autor, ao se planejar com cuidado a documentação e preservação dos dados, os objetivos de ter dados de pesquisa reprodutíveis e transparentes são muito mais fáceis de alcançar. Além disso, dados bem geridos são mais fáceis de utilizar e viabilizar sua reutilização, o que se traduz em uma maior colaboração para pesquisadores e um máximo de retorno do investimento para as agências de fomento.

Bell (2011, p. 13), simplificando o processo de gestão dos dados de pesquisa argumenta "a ciência com uso intensivo de dados consiste em três atividades básicas: captura, curadoria e análise*"*.

A curadoria de dados pode ser entendida como a gestão e a preservação de dados a longo prazo, incluindo-se nesse contexto o fato de agregar valor aos dados digitais, bem como viabilizar a criação de novos dados, de forma colaborativa, a partir dos já existentes. Além disso, a atividade de curadoria também pode propiciar a redução dos riscos de obsolescência digital (ABOTT, 2008; DIGITAL CURATION CENTRE, 2016; GIARETTA, 2004; HEY; TANSLEY; TOLLE, 2011).

Para o Digital Curation Centre (2016), a curadoria digital "envolve a manutenção, a preservação e a agregação de valor aos dados da pesquisa digital em toda sua vida útil. A gestão ativa dos dados de pesquisa, por sua vez, reduz as ameaças ao seu valor de pesquisa de longo prazo e reduz o risco de obsolescência digital". A instituição vai além ao comentar sobre o compartilhamento e reutilização de dados – os dados curados disponíveis em repositórios digitais de confiança podem ser compartilhados entre a comunidade mais ampla de pesquisa do Reino Unido.

Conway (2011), Sayão e Sales (2012), bem como Sales (2014) salientam que a teoria da curadoria digital traz, no contexto da preservação digital, o diferencial de que a informação não "apenas" deve ser preservada digitalmente, mas sim de passar pelo processo de curadoria digital, o que envolveria o tratamento da informação desde a coleta dos dados de pesquisa até o reuso da informação por outros integrantes do fluxo informacional.

Considerando o exposto acima a respeito da curadoria, merece ser comentado que essa atividade não aparece de forma explícita no Modelo de Ciclo de Vida do dado do Projeto DataONE, mas pode-se inferir que está implícita nas atividades de descrição e preservação.

### 2.4.2 *Workflow* Científico

As características da ciência colaborativa permitem que o cientista tenha acesso a uma ampla disponibilidade de sistemas computacionais de alto desempenho, *grids* e *clouds*. Consequentemente, cientistas e engenheiros implementam aplicativos cada vez mais complexos para acessar e processar grandes repositórios de dados e executar experimentos científicos em plataformas de computação distribuída. A maioria desses aplicativos é projetada como sistema de fluxo de trabalho (*workflow*) que incluem análise de dados, métodos de computação científica e técnicas de simulação complexas.

De forma genérica, a Workflow Management Coalition[31] define *workflow* (fluxo de trabalho) como a "automação de um processo de negócios, no todo ou em parte, durante o qual documentos, informações ou tarefas são passadas de um participante para outro". O termo processo, por sua vez, indica um conjunto de tarefas associadas com o objetivo de criar um produto, calcular um resultado ou fornecer um serviço. Portanto, cada tarefa (ou atividade) representa um trabalho que forma um passo lógico do processo global.

[31] Workflow Management Coalition (WfMC). Disponível em: <http://www.wfmc.org/>. Acesso em: 2 out. 2016.

De acordo com Talia (2013) os sistemas de *workflow* científico "são capazes de integrar rotinas de *software* existentes, conjuntos de dados e serviços em composições complexas que implementam processos de descoberta científica".

Talia (2013) defende que os sistemas fluxo de trabalho científicos mais utilizados são: o Taverna, o Pegasus, o Triana, o Askalon, o Kepler, o GWES e o Karajan, que fornecem ferramentas e *frameworks* inovadores para a implementação de aplicações em Ciência e Engenharia.

O *software* Taverna[32] é um *workflow* científico desenvolvido pela Universidade de Manchester, em código aberto baseado em Java. Tem como principal objetivo apoiar a comunidade de ciências da vida (biologia, química e medicina) na projeção e execução de fluxo de trabalho científico. Apesar de o *software* ter sido desenvolvido no âmbito da bioinformática, ele pode ser utilizado em outras áreas do conhecimento.

O *workflow* Pegasus[33] foi desenvolvido na Universidade do Sul da Califórnia. Inclui um conjunto de tecnologias para executar aplicativos baseados em *workflow* em vários ambientes diferentes (*desktops, clusters, Grids* e *Clouds*). Ele tem sido utilizado em várias áreas científicas, incluindo bioinformática, astronomia, física gravitacional, oceanografia, dentre outras.

O Kepler é um *software* de *workflow* científico desenvolvido a partir de Ptolomeu II34, uma ferramenta de modelagem orientada a atores, destinada principalmente ao projeto de sistemas embutidos e em tempo real. O *software* ajuda os usuários a compartilhar e reutilizar dados, fluxos de trabalho e componentes desenvolvidos pela comunidade científica para atender às necessidades comuns. O foco da ferramenta está na análise e modelagem de dados. Ao invés de tentar fornecer uma semântica genérica para todos os tipos possíveis de processos encontrados, o Kepler separa o mecanismo de execução do modelo de fluxo de trabalho e atribui um modelo de computação para cada fluxo. O software trabalha com diversas áreas, desde física, até ecossistema e bioinformática.

[32] Disponível em: <http://www.taverna.org.uk/>. Acesso em: 2 out. 2016.
[33] Disponível em: <https://pegasus.isi.edu/>. Acesso em: 2 out. 2016.
[34] Ptolomeu II é uma estrutura de *software* de código aberto que suporta a experimentação com *design* ator-orientado. Está em desenvolvimento desde 1996, é um sucessor de Ptolomeu Clássico, que foi desenvolvido desde 1990. Fonte: http://ptolemy.eecs.berkeley.edu/ptolemyII/

### 2.4.3 Repositórios de Dados Científicos

O grande volume de dados produzidos na *e-science* traz à tona pelo menos duas grandes vertentes a serem trabalhadas, a primeira diz respeito a infraestrutura tecnológica necessária para permitir o compartilhamento de dados. Já a segunda trata da preservação desses dados para a posteridade, que por sua vez, traz consigo questões de armazenamento, regras de acesso e até mesmo regras de reutilização dos dados. No ano de 2016, foram identificadas duas iniciativas de grande porte no âmbito de repositório de dados: a Research Data Aliance e o Datacite a seguir comentados.

#### 2.4.3.1 Research Data Alliance (RDA[35])

A Research Data Alliance (RDA) foi lançada, em 2013, como uma organização orientada para a comunidade, pela Comissão Europeia, pela NSF e pelo Instituto Nacional de Normas e Tecnologia dos Estados Unidos, e também pelo Departamento de Inovação do Governo australiano com o objetivo de desenvolver uma infraestrutura social e técnica para permitir o compartilhamento aberto de dados de pesquisa.

Com mais de 4.600 membros de 115 países[36], a RDA fornece um espaço neutro onde seus membros podem reunir-se através de Grupos Globais de Trabalho e de Interesse focados para desenvolver e adotar infraestrutura que promova o compartilhamento de dados e a pesquisa orientada por dados, bem como acelerar o crescimento de uma comunidade coesa de dados, que integra contribuintes em domínios, pesquisas e fronteiras nacionais, geográficas e geracionais.

A estrutura organizacional da RDA é composta por quatro núcleos: o Conselho, o Comitê Consultivo Técnico, o Comitê Consultivo Organizacional e o Secretariado. O Conselho é responsável pela supervisão, sustentabilidade e sucesso global da RDA, incluindo a aprovação dos Grupos de Trabalho e de Interesse para garantir o alinhamento com as metas da RDA. O Conselho tem nove membros, incluindo dois co-presidentes do Conselho. Cumpre destacar que um destes membros é vinculado à Rede Nacional de Pesquisa (RNP), órgão do Ministério da Ciência, Tecnologia, Inovação e Comunicações do Brasil (MCTIC). As operações e a administração geral da RDA são conduzidas pelo Secretariado, que é distribuído internacionalmente.

---

[35] Todas as informações contidas nesse tópico foram retiradas do site da organização Research Data Alliance durante a realização da pesquisa.

[36] Dados contabilizados em dezembro de 2016. Disponível em: <www.rd-alliance.org>. Acesso em: 2 out. 2016.

O Comitê Consultivo Técnico é responsável pela direção técnica da RDA e fornece aconselhamento ao Conselho RDA, bem como ajudar a desenvolver e rever Grupos de Trabalho e de Interesse da RDA para promover o seu impacto e eficácia. Este Comitê também é responsável pelo desenvolvimento, manutenção e evolução do Roteiro Técnico RDA. Um co-presidente deste Comitê serve como membro observador do Conselho RDA.

O Comitê Consultivo Organizacional é composto por representantes dos Membros Organizacionais também conhecidos como Assembleia Organizacional. O Comitê Organizacional aconselha o Conselho RDA sobre as direções, processos e mecanismos da RDA. Com o apoio do Secretariado, este Comitê é responsável pelo desenvolvimento e manutenção do documento do Plano Organizacional e de Processo da RDA. Um co-presidente deste Comitê serve como membro observador do Conselho. Toda a estrutura supracitada pode ser observada na Figura 7.

Figura 7 – Estrutura Organizacional da RDA.


Fonte: Research Data Alliance (2016).

A participação na RDA está aberta a qualquer pessoa que aceite os seus princípios orientadores de abertura, consenso, equilíbrio, harmonização, com uma abordagem dirigida à comunidade e sem fins lucrativos. No que diz respeito aos membros individuais, que por sua vez, fazem parte dos Grupos de Trabalho e Grupos de Interesse, estes representam a principal parte da organização RDA. Os profissionais que normalmente integram o corpo de membros individuais possuem os seguintes perfis: profissionais de dados, especialistas e gerentes em tecnologia da informação, pesquisadores e cientistas, professores universitários, bibliotecários

e arquivistas, gerentes de projeto, membros da alta administração organizacional, desenvolvedores de políticas e outros perfis que são afetados pelo dilúvio de dados de pesquisa.

A composição de membros dos Comitês Consultivos (Technical Advisory Board, Secretariat, Organisational Advisory Board) revela uma participação predominante de países como Estados Unidos, Reino Unido e Austrália. Além disso, percebe-se a predominância das seguintes disciplinas ou domínios: oceanografia, ciências da vida e da saúde, agricultura, espaço e mudanças climáticas.

### 2.4.3.2 Datacite[37]

DataCite se apresenta como "uma importante organização global sem fins lucrativos que fornece identificadores persistentes (DOI)[38] para dados de pesquisa". Sendo que o seu objetivo é "ajudar a comunidade de pesquisa localizar, identificar e citar dados de pesquisa com confiança". Trabalham em várias frentes para alcançar esse objetivo, dentre elas, a) apoiam a criação e atribuição de DOI e metadados que o acompanham; b) oferecem o serviço que apoia a pesquisa avançada e a descoberta de conteúdo de pesquisa e; c) promovem a citação de dados através de esforços de construção de comunidade e comunicação responsiva e materiais de divulgação.

Ainda segundo o *site*, ela ressalta que oferece uma série de serviços para atender as diversas necessidades da comunidade global de pesquisa, que esta comunidade foi reunida para enfrentar os desafios de tornar os dados da pesquisa visíveis e acessíveis, e ainda colaboram como uma rede global de pesquisadores para fornecer o seguinte apoio aos pesquisadores, data centers, editores de revistas e agências de financiamento. Este apoio se dá da seguinte forma:

- aos pesquisadores em seus esforços para encontrar, identificar e citar dados de pesquisa e outros objetos de pesquisa;
- aos Centros de dados fornecendo DOI para conjuntos de dados, fluxos de trabalho e padrões;
- aos Editores de periódicos, permitindo que artigos de pesquisa sejam vinculados a dados / objetos subjacentes;
- aos organismos de financiamento, ajudando-os a compreender o alcance e o impacto do seu financiamento.

[37] Todas as informações contidas nesse tópico foram retiradas do site da organização DataCite durante a realização da pesquisa.
[38] Digital Object Identifier.

## 2.5 AÇÕES GOVERNAMENTAIS PARA DADOS CIENTÍFICOS EM PAÍSES DE PRIMEIRO MUNDO

A literatura internacional revela que as iniciativas mais maduras tanto em termos de infraestrutura tecnológica, como de diretrizes para a gestão de projetos de dados científicos, concentram-se nos Estados Unidos e no Reino Unido, sendo que o Reino Unido aparece na vanguarda, tendo lançado o primeiro programa em 2001. Em contrapartida, nos EUA, o início é marcado pela publicação do *Relatório Atkins* em 2003. Além dessas duas grandes iniciativas, há registros de ações no Canadá e até mesmo no Brasil.

Inúmeros governos e agências de fomento, segundo Shearer (2015, p. 4) começam a elaborar políticas públicas relacionadas com a gestão de dados científicos (por vezes nomeada de gestão de dados de pesquisa). Geralmente essas políticas visam ampliar a eficiência da pesquisa, motivar a reutilização de dados, acelerar as ações cooperativas entre pesquisadores e suas entidades. Para a autora:

> As jurisdições com os ambientes de políticas mais abrangentes são o Reino Unido, os Estados Unidos, a Austrália e a União Europeia. Detalhes de políticas variam entre regiões, agências e domínios, mas eles também têm uma série de coisas em comum. Os componentes políticos mais frequentes são os requisitos em torno de padrões e metadados, o compartilhamento de dados e a retenção de dados e/ou preservação a longo prazo. Planos de gestão de dados (GDP) são geralmente necessários no contexto dessas políticas, já que obrigam os investigadores a pensarem sobre como eles irão gerenciar seus dados antes do projeto ter se iniciado, um requisito chave para as boas práticas de gestão de dados. As políticas também contêm consistentemente disposições para a proteção da confidencialidade, propriedade intelectual e dados sensíveis (SHEARER, 2015, p. 4).

Shearer (2015) defende que os objetivos de uma política para a gestão de dados científicos são: acelerar o processo de investigação, apoiar novos *insights* e descobertas, fomentar a colaboração entre pesquisadores, melhorar a eficiência da investigação e facilitar a prestação de contas. Para a autora, uma determinada política de Research Data Management (RDM) refletirá os objetivos e princípios em que se baseia. Portanto, embora muitas políticas contenham elementos semelhantes, pode haver maior ênfase em alguns requisitos sobre os outros. Por exemplo, uma política baseada no princípio do compartilhamento de dados provavelmente se concentrará nas práticas-chave necessárias para fornecer acesso aos dados, ao passo que uma política baseada na administração de dados se concentrará nas funções e responsabilidades envolvidas no gerenciamento de dados. Com fundamento no exposto,

Shearer (2015, p. 8) apresenta os elementos comuns em uma política de RDM, conforme descrito no Quadro 5.

Quadro 5 – Elementos comuns em uma política de RDM.

| **Requisitos da Política** | |
|---|---|
| Qualidade e padrões de dados | Os investigadores são obrigados a aderir aos padrões internacionais para permitir o acesso e reutilização.<br>A documentação de dados e os metadados devem acompanhar dados para que os dados sejam compreensíveis por outros. |
| Acesso e compartilhamento de dados | Os investigadores são obrigados a disponibilizar os dados para serem partilhados (normalmente após a publicação dos resultados ou pouco depois, embora algumas agências autorizem períodos de embargo).<br>Requisitos para o depósito de metadados em um catálogo local ou nacional. |
| Retenção e preservação de dados | Os dados devem ser mantidos por um período de tempo mínimo. Sempre que possível, os investigadores devem depositar os seus dados num arquivo de longo prazo para garantir a preservação dos seus dados. |
| Planos de gestão de dados | As propostas de pesquisa devem incluir um plano de gestão de dados. |
| **Disposições comuns às políticas** | |
| Privacidade | Os direitos e a privacidade dos indivíduos que participam na pesquisa devem ser protegidos em todos os momentos. Assim, os dados disponibilizados para uso mais amplo devem estar livres de identificadores que permitam ligações a participantes individuais da pesquisa e variáveis que podem levar à divulgação dedutiva da identidade de sujeitos individuais. |
| Conhecimento tradicional | No que se refere aos conhecimentos locais e tradicionais, os direitos dos detentores de conhecimentos não devem ser comprometidos. |
| Dados de natureza sensível | Quando a liberação de dados pode causar danos, aspectos específicos dos dados podem precisar ser protegidos (por exemplo, localização de ninhos de aves ameaçadas de extinção e localização de santuários ecológicos - locais sagrados) |
| Propriedade intelectual / Propriedade dos dados | Poderá ser necessário, por vezes, atrasar a publicação por um curto período de tempo para permitir a elaboração do pedido. |
| **Outros aspectos** | |
| Princípios | As políticas de dados aderem a um conjunto de princípios gerais que articulam seu valor. |
| Âmbito / Cobertura da Política | Descreva o escopo dos dados cobertos pela política. |
| Funções e Responsabilidades | A política identifica as várias partes responsáveis pela gestão dos dados nas diferentes fases do ciclo de vida. |
| Acompanhamento e execução | Os meios pelos quais as políticas serão monitoradas ou aplicadas são descritos na política. |

Fonte: Shearer (2015, p. 8).

De acordo com o Digital Curation Center (DCC), o *checklist* para planejar uma gestão de dados é composto por sete itens, conforme descrito no Quadro 6.

Quadro 6 – *Checklist* para a gestão de dados de pesquisa.

| | |
|---|---|
| Coleção de dados | Quais dados serão coletados ou criados?<br>Como os dados serão coletados ou criados? |
| Documentação e Metadados | Que padrões, documentação e metadados irão acompanhar os dados? |
| Ética e Compliance Legal | Como serão tratadas as questões éticas?<br>Como serão gerenciados os direitos autorais e direitos de propriedade intelectual (DPI)? |
| Armazenamento e Backup | Como os dados serão armazenados e apoiados durante a pesquisa?<br>Como o acesso e a segurança serão gerenciados? |
| Retenção e Preservação | Que dados devem ser conservados e / ou preservados?<br>Qual é o plano de preservação de longo prazo para os dados? |
| Compartilhamento de dados | Como os dados serão compartilhados?<br>São necessárias restrições na partilha de dados? |
| Responsabilidades e recursos | Quem será responsável pelo gerenciamento de dados?<br>Que recursos serão necessários para entregar o plano de gerenciamento de dados? |

Fonte: Digital Curation Center (2017).

Esta tese tem como um de seus objetivos delinear diretrizes para uma política de gestão de dados científicos no Brasil. Em função disso, a revisão de literatura priorizou as iniciativas mais maduras em programas de *e-science* / *cyberinfrastructure* – a exemplo dos Estados Unidos e Reino Unido. Porém, algumas iniciativas consideradas relevantes em Portugal e Espanha também foram comentadas ao longo da revisão de literatura.

O Reino Unido lançou, em 2001, um programa pioneiro – *e-Science Core Programme* – que recebeu um aporte financeiro de £250 milhões com o objetivo de estimular a *e-science* em todos os campos de pesquisa. O programa tinha como objetivo:

> prover a infraestrutura e facilidades necessárias para a pesquisa colaborativa, acelerar a emergência da próxima geração de padrões de plataforma aberta para serviços globais de informação, resolver os principais desafios em processamento, comunicação, e armazenamento de grandes volumes de dados (VAZ, 2011, p. 10, grifo nosso).

O *e-Science Core Programme* é um programa gerido pelo Conselho de Pesquisa em Ciências da Engenharia e Física, em nome das comunidades de todos os Conselhos de Pesquisa. O programa tem apoiado o desenvolvimento de tecnologias genéricas, como o *software* conhecido como *middleware* – necessário para permitir que diferentes recursos trabalhem de forma integrada por meio de redes, bem como criem *grids* computacionais.

Hey e Trefethen (2002) relatam que o objetivo do *e-Science Core Programm* é identificar os requisitos genéricos de *middleware* decorrentes dos projetos-piloto *e-science*, em colaboração com cientistas, cientistas da computação e da indústria. Para tanto, será

desenvolvido um *grid middleware* robusto e com força industrial que não só irá apoiar áreas de aplicação individuais, mas também ser de relevância para a indústria e comércio. De acordo com os autores, o programa foi estruturado em torno de seis elementos-chave:

1. Implementação de um banco de dados nacional *e-Science Grid* baseado em uma rede de Centros de *e-Science*.
2. Promoção do desenvolvimento de *grid middleware* genérico.
3. Projetos de Grid da Colaboração de Pesquisa Interdisciplinar (IRC).
4. Estabelecimento de uma estrutura de apoio para projetos-piloto *e-Science*.
5. Apoio ao envolvimento em atividades internacionais.
6. Suporte para requisitos de rede de *e-Science*.

Hey e Trefethen (2002) também alertam para o fato de que o sucesso para os projetos de *e-Science* não envolvem apenas questões técnicas de infraestrutura tecnológica tais como escalabilidade, confiabilidade, interoperabilidade, tolerância a falhas, gerenciamento de recursos, desempenho e segurança. É preciso atenção para questões inerentes às pessoas envolvidas nos projetos tais como a vontade de trabalhar de forma colaborativa, aceitando o compartilhamento de recursos e dados. A respeito do assunto, os autores argumentam que "para que a ciência faça o melhor uso de seus recursos limitados, a partilha desses dados científicos reunidos de forma dispendiosa é claramente de suma importância. No entanto, a motivação para qualquer cientista individual não é tão clara. Talvez as agências de financiamento precisem acrescentar algum incentivo para encorajar essa abordagem de compartilhamento de dados científicos".

Em 2002, ano de publicação do artigo, Hey e Trefethen (2002) já alertavam para questões que não seriam resolvidas nos três anos iniciais do *Core Programm*. Dentre elas, os autores destacam a segurança e a coleta de dados científicos e a conservação a longo prazo dos dados científicos, juntamente com as suas anotações de metadados associadas.

De acordo com Shearer (2015, p. 9) o Reino Unido emitiu em 2011 um conjunto de princípios comuns sobre política de dados – "Common Principles on Data Policy". Esse conjunto de princípios exige que os "dados sejam disponibilizados abertamente com o menor número possível de restrições". A partir dessa publicação, foram implementadas várias políticas de acesso à dados de pesquisa, a exemplo do Wellcome Trust, que financia pesquisa biomédica. Porém, apesar de sofrerem variações em termos de detalhes, de uma forma geral estão alinhadas aos princípios comuns. Shearer (2015) argumenta que de acordo com uma visão geral publicada pela Universidade de Bath, as políticas de gestão de dados científicos geralmente cobrem os seguintes elementos: a) tipos de dados abrangidos pela política, b) expectativas de partilha de

dados, incluindo acesso e prazos, c) períodos mínimos de retenção de dados, d) utilização de metadados e normas de documentação, e) exceções justificadas à partilha de dados, f) custos associados à gestão de dados que podem ser pagos através de subvenções e, g) reconhecimento de criadores de dados.

Já no contexto dos Estados Unidos destacam-se os trabalhos vinculados às universidades de Purdue e de Washington. Além disso, percebe-se um interesse de grandes corporações, como a Microsoft pelo tema, destacando-se o fato de Tony Hey, atual vice-presidente da área de pesquisa da Microsoft, ter sido o diretor do *e-Science Core Programme* no Reino Unido. Hey e Trefethen (2002) relacionam os projetos que envolvem *grid* computacional nos EUA, conforme descrito no Quadro 7.

Quadro 7 - Projetos em *Grid* financiados nos Estados Unidos.

| Projeto | Agência de Fomento | URL |
|---|---|---|
| IPG | NASA | http://www.nas.nasa.org/About/IPG/ipg.htm |
| Science Grid | DOE | http://www-itg.lbl.gov/Grid/ |
| GrilPhyN Grid | NSF | http://www.griphyn.org/ |
| PPDataGrid | DOE | http://www.ppdg.net/ |
| NVO | NSF | http://www.srl.caltech.edu/nvo/ |
| NESSGrid | NSF | http://www.nessgrid.org/html/np.html |
| Distributed Terascale Facility (TeraGrid) | NSF | http://www.teragrid.org/ |
| DISCOM (ASCI) | DOE | http://www.cs.sandia.gov/discom/ |
| Earth Systems Grid | DOE | http://public.lanl.gov/radiant/research/grid.html |
| FusionGrid | DOE | http://www.fusoingrid.org/ |
| BIRN | NIH | http://birn.ncrr.nuh.gov/ |
| iVDGL | NSF | http://www.ivdgl.org/ |
| GridCenter | NSF | http://www.grids-center.org |
| GrADS | NSF | http://nhse2.cs.rice.edu/grads/ |

Fonte: Hey e Trefethen (2002).

Os autores, em seu estudo, também identificaram projetos que envolvem grid computacional realizados na União Europeia, conforme descrito no Quadro 8.

Quadro 8 - Projetos em *Grid* financiados na União Europeia.

| **Projeto** | **Agência de Fomento** | **URL** |
|---|---|---|
| DataGrid (CERN) | European Commission | http://www.datagrid.cnr.it, http://www.cern.ch/grid/ |
| EuroGrid (Unicore) | European Commission | http://www.eurogrid.org/ |
| Damien (Metacomputing) | European Commission | http://www.hlrs.de/organization/pds/projects/damien/ |

| Projeto | Agência de Fomento | URL |
|---|---|---|
| AVO (Virtual Observatory) | European Commission | http://www.astro-opticon.org/archives.html |
| GRIP (Unicore/Globus) | National Center for Research Resources | http://www.unicore.org/links.htm |
| GridLab | European Commission | http://www.gridlab.org |
| CrossGrid | European Commission | http://www.crossgrid.org/crossgrid/crossgrid.html |
| Grid-Ireland | | http://www.cs.tcd.ie/coghlan/, http://www.cuc.ucc.ie/ |
| Grid for remote computing | | http://sara.unile.it/grb/grb.html |

Fonte: Hey e Trefethen (2002).

## 2.6 DADOS CIENTÍFICOS NO BRASIL

No Brasil, a problemática dos dados oriundos da *e-science* ainda é pouco trabalhada. A busca bibliográfica, realizada em bases de dados nacionais e internacionais, revela uma incipiência de estudos que contemplem as contribuições da Biblioteconomia e da Ciência da Informação para a *e-science* no contexto brasileiro. Em pesquisa que utilizou a Internet como fontes de dados secundários foram identificadas duas iniciativas mais estruturadas – o Portal da Biodiversidade e o Programa eScience da FAPESP, que serão trabalhados em um tópico específico dada a relevância das inciativas.

No contexto da ciência da informação brasileiro, Cunha (2010) foi vanguardista ao afirmar que acervo de dados oriundo da *e-science* tende a crescer e que o tratamento desses dados pela biblioteca universitária exigirá a capacitação de recursos humanos para tal atividade. Moura (2011, p. 165) tem estudado a *e-science* a partir de "uma amostra de *blogs* científicos mantidos por pesquisadores como estratégia para o registro e a divulgação dos resultados parciais de sua pesquisa, *sites* colaborativos internacionais e de centros internacionais que apoiam as práticas." Dentre os resultados da pesquisa da autora, merece destaque a criação do *Online Dictionary of E-Science, Cyberculture and Scientific Narratives*[39]. Por outro lado, analisando a *e-science* no campo da comunicação científica, destaca-se o pioneirismo do trabalho de Medeiros e Caregnato (2012) ao abordar o compartilhamento de dados científicos.

Vaz (2011) faz uma comparação entre o cenário de tratamento da *e-science* no Reino Unido e no Brasil. O autor conclui que no Brasil há poucos cientistas que têm conhecimento ou

[39] Disponível em: http://mamoura.eci.ufmg.br/dictionary/. Acesso em: 2 out. 2016.

interesse sobre o tema, evidenciando o atraso do país nesse cenário. Tal situação reforça a necessidade de se fomentar pesquisas nesse tema. Em especial, a necessidade de pesquisas que apontem a contribuição da ciência da informação no tratamento desses dados.

Sobre o assunto, a percepção de Sales (2014; p. 49) é de que: "os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a entender que estes dados, se preservados e bem gerenciados, constituem uma excelente fonte de recursos informacionais que podem ser compartilhados e reutilizados como insumo para novas pesquisas".

Corroborando a percepção de Sales (2014), durante o desenvolvimento desta tese foi constatado que o Brasil possui os seguintes repositórios: a) Repositório de Dados do Programa de Pesquisa de Biodiversidade da Amazônia Ocidental (PPBIO[40]), b) Repositório de Dados do Programa de Pesquisas Ecológicas de Longa Duração (PELD[41]), Portal GEOINFO de infraestrutura de dados espaciais da EMBRAPA (*com 1.081 itens catalogados*) dentre outros que serão comentados.

Além desses, merece ser comentado o Portal da Biodiversidade[42] (SISBio) lançado pelo Ministério do Meio Ambiente e pelo Instituto Chico Mendes (ICMBio) em 26 de novembro de 2015, esse conteúdo será apresentado no Capítulo 2.6.2 em função da relevância da iniciativa.

No que diz respeito a eventos relacionados à temática, uma das evidências da emergência do tema no Brasil foi a realização do *VI Workshop de e-Science*, no período de 16 a 19 de julho de 2012, na cidade de Curitiba. Em 2013, no período de 23 a 26 de julho, na cidade de Maceió, foi realizada a 33º edição do Congresso da Sociedade Brasileira de Computação, tendo como evento paralelo o BreSci – *VII Brazilian e-Science workshop* que propôs a realização de um fórum de discussão sobre temas relevantes na *e-science.* Dentre os trabalhos apresentados, destacam-se o de Sales e Sayão (2013) sobre *ciberinfraestrutura para integração, acesso, compartilhamento e reuso de dados de pesquisa da área nuclear*, bem como a palestra *e-Science in Brazil: The SINAPAD perspective* ministrada por Antônio Tadeu Azevedo Gomes.

[40] O PPBio foi criado em 2004 com o objetivo de desenhar uma estratégia de investimento em ciência, tecnologia e inovação (CT&I) que aponte prioridades, integre competências em diversos campos do conhecimento e dissemine informações sobre biodiversidade que possam ser utilizadas para diferentes finalidades.

[41] O PELD, criado em 1999, é uma iniciativa pioneira no sentido de obter informações relevantes para a conservação da biodiversidade e uso sustentável dos recursos naturais dos ecossistemas brasileiros. As informações coletadas no PELD, que incluem longas séries temporais de dados sobre os ecossistemas e sua biota associada.

[42] O Portal da Biodiversidade representa a interface *web* do Sistema de Autorização e Informação em Biodiversidade (SISBIO). Disponível em: <https://portaldabiodiversidade.icmbio.gov.br/portal/>. Acesso em: 2 out. 2016.

Ainda em 2013, o mês de maio, a Fundação de Amparo à Pesquisa do Estado de São Paulo e a Microsoft promoveram o Latin American Workshop. O evento teve como objetivo discutir o avanço do conhecimento científico, a partir do aumento da capacidade de análise de grandes volumes de dados, em todas as áreas do conhecimento. Estiveram presentes pesquisadores da Europa, América do Sul, América do Norte e Oceania. A abertura do evento contou com a presença de Tony Hey – vice-presidente da Microsoft Research. No Brasil, destacaram-se as participações de Marcos Buckeridge, professor do Instituto de Biociências da USP e Carlos Joy, coordenador do Programa Biota da FAPESP. Durante o evento, foram apresentados onze projetos sobre o tema conduzidos pela Microsoft Research e colaboradores (FAPESP, 2013; UNIVERSIDADE ESTADUAL DE SÃO PAULO, 2013).

O ano de 2013 mostra-se marcante para eventos na área, pois em maio de 2013, foi realizado, na cidade de Marília, o *Encontro Internacional de Dados, Tecnologia e Informação*, evento promovido Programa de Pós-Graduação em Ciência da Informação da UNESP, ocasião em que foi apresentado o trabalho de Costa *et al.* (2013) que discute a relação da *e-science* com a biblioteconomia e a ciência da informação.

Em 2014, no mês de maio, o IBICT promoveu o I Seminário Internacional de Preservação Digital (SINPRED). Na ocasião, palestrantes internacionais apresentaram e discutiram sobre o sistema LOCKSS[43]. Dentre os palestrantes, Cunha (2014) trouxe para a discussão o aspecto de preservação dos dados digitais oriundos da *e-science*.

Em outubro de 2014, foi realizado na cidade de São Paulo, o *IEEE 10th International Conference on e-Science*. No âmbito da Ciência da Informação, o Seminário Nacional de Bibliotecas Universitárias (SNBU), que foi realizado, em Belo Horizonte, no período de 16 a 21 de novembro de 2014, na cidade de Belo Horizonte, incluiu como eixo temático a preservação de acervos digitais.

Outro evento de destaque foi realizado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), no dia 27 de março de 2014, com o objetivo de apresentar o *Programa FAPESP de Pesquisa em e-Science*.

Em termos de infraestrutura de pesquisa, no Brasil, diversas infraestruturas para e-Ciência (também chamada de e-Infrastructures ou Cyberinfrastructures) estão sendo desenvolvidas. Entre as mais conhecidas, pode-se citar o SINAPAD – Sistema Nacional de processamento de alto desempenho – uma rede de centros de computação de alto desempenho,

[43] *Lots Of Copies Keeps Stuff Save.* É um sistema de código aberto que cria uma rede de replicação de dados (cópias compartilhadas de periódicos eletrônicos), permitindo que os participantes acessem dados preservados confiáveis através de uma conexão restrita a um grupo.

geograficamente distribuídos, instituída pelo Ministério da Ciência, Tecnologia, Inovações e Comunicações. São oito unidades, denominadas "Centros Nacionais de Processamento de Alto Desempenho[44]" (CENAPADs), operadas respectivamente pela UFRGS, UFMG, UFC, Unicamp, UFRJ, UFPE, INPE e LNCC. Este último coordena o sistema por delegação do MCT. Outra iniciativa de destaque no Brasil é o eScience Research Network[45] da Universidade de São Paulo.

### 2.6.1 O Programa FAPESP de Pesquisa em e-Science

O Programa e-Science da FAPESP[46] foi criado em 2013, seu principal objetivo à época foi descrito como "encorajar abordagens novas, ousadas e não convencionais para pesquisa de ponta, multidisciplinar, integrando grupos de pesquisa em Computação e outras áreas. " (FAPESP, 2014). Ainda de acordo com o documento sobre o programa, a colaboração da ciência da computação com outras áreas "visa investigar como os avanços da pesquisa em computação podem ajudar a vencer desafios científicos e tecnológicos em outros domínios e vice-versa".

Em 27 de março de 2014, a FAPESP realizou um evento de apresentação e discussão do o *Programa FAPESP de Pesquisa em e-Science*. O objetivo do evento foi apresentar a comunidade cientifica as diversas características do programa e tirar dúvidas dos participantes que pretendiam submeter propostas. Os pesquisadores envolvidos na apresentação do Programa foram Medeiros e Cesar Junior (FAPESP, 2014).

Durante apresentação sobre o Programa *e-Science* foi contextualizado que seu nascimento se deu em função da constatação recorrente da Coordenação de Ciência e Engenharia da Computação da Fundação que nos últimos anos, os pedidos de apoio a projetos de pesquisa em diferentes áreas tinham em comum a necessidade de pesquisa computacional (dados, modelagem, visualização, algoritmos). Assim, a partir dessa constatação, foram realizados em 2009 dois *workshops* com a participação de cientistas da computação e de outras áreas com o objetivo de "identificar melhor as demandas de pesquisa computacional dos cientistas no Estado de São Paulo" (ALISSON, 2014). Na ocasião, chegou-se à conclusão sobre

[44] Disponível em: <https://www.lncc.br/sinapad/>. Acesso em: 2 out. 2016.
[45] Disponível em: <http://escience.ime.usp.br/index.php>. Acesso em: 2 out. 2016.
[46] Folder eletrônico disponível em: <http://www.fapesp.br/publicacoes/2015/folder_escience.pdf>. Acesso em: 2 out. 2016.

a importância do avanço da computação em outros domínios do conhecimento e o fato da a *e-science* ser um tema multidisciplinar transversal de pesquisa.

A primeira chamada de propostas de pesquisa[47] do Programa e-Science foi aberta para "pesquisadores associados a instituições públicas de ensino superior ou de pesquisa no Estado de São Paulo" (FAPESP, 2014). O edital disponibilizou quatro milhões de reais para apoiar projetos relevantes que envolvessem modelos matemáticos, repositórios digitais e gerenciamento de dados, novos *hardwares*, *softwares*, protocolos, ferramentas e serviços, voltados para atender demandas de pesquisas nas áreas de ciências agrárias, artes, humanidades e ciências sociais; engenharia e física, clima e ciências da terra, e à prática e educação em *e-science*. O objetivo principal do edital era "identificar, selecionar e expandir pesquisa de classe mundial, básica e aplicada, associada aos tópicos relevantes descritos na chamada". (FAPESP, 2014). O Programa da FAPESP busca integrar modelagem computacional e infraestrutura de dados e pesquisas em diversas áreas do conhecimento. O resultado da primeira chamada de proposta foi divulgado no dia 29 de outubro de 2014, tendo sido selecionadas quatro propostas, conforme descrição do Quadro 8.

Em 2015, a FAPESP divulgou novo edital de pesquisa[48], novamente com orçamento de quatro milhões de reais, cujo resultado foi divulgado em agosto de 2016[49], conforme resultados apresentados no Quadro 9. A chamada também foi aberta para "pesquisadores associados a instituições públicas de ensino superior ou de pesquisa no Estado de São Paulo" (FAPESP, 2015).

Quadro 9 – Projetos aprovados nos editais de 2014 e 2015 da FAPESP.

| Pesquisador | Instituição | Título do Projeto | Edital | Vigência |
|---|---|---|---|---|
| Gilberto Câmara Neto | Instituto Nacional de Pesquisas Espaciais / MCTI | e-Sensing: Análise de grandes volumes de dados de observação da terra para informação de mudanças de uso e cobertura da terra<br><br>Nº do processo FAPESP: 2014/08398-6 | 1º Edital | 01/01/2015 até 31/12/2018 |

[47] A chamada para propostas de pesquisas está disponível em: <http://www.fapesp.br/en/8354>. Acesso em: 2 out. 2016.
[48] A chamada para propostas de pesquisas está disponível em: <http://www.fapesp.br/en/9888>.
[49] O resultado dos projetos aprovados no segundo edital estão disponíveis em: <http://www.fapesp.br/en/10488>.

| Pesquisador | Instituição | Título do Projeto | Edital | Vigência |
|---|---|---|---|---|
| Luciana Alvim Santos Romani | Embrapa Informática Agropecuária / Embrapa | AgroComputing.net - Infraestrutura digital e novos métodos computacionais para análise e mineração de grandes bases de dados climáticos e de sensoriamento remoto para aperfeiçoar o monitoramento e previsão agrícola<br>Nº do processo FAPESP: 2014/08293-0 | 1º Edital | 01/01/2015 até 31/12/2016 |
| Marcelo Knörich Zuffo | Escola Politécnica / USP | CiberArqueologia - Realidade virtual e eScience ao encontro da Arqueologia<br>Nº do processo FAPESP: 2014/08418-7 | 1º Edital | 01/01/2015 até 31/12/2016 |
| Marco Henrique Terra | Escola de Eng. de São Carlos / USP | Sistema de referência de atitude, orientação e posição baseado em filtro de Kalman robusto implementado em FPGA<br>Nº do processo FAPESP: 2014/08432-0 | 1º Edital | 01/01/2015 até 30/06/2017 |
| Luís Antonio Coelho Ferla | Escola de Filosofia, Letras e Ciências Humanas / EFLCH / Unifesp | Pauliceia 2.0: uma plataforma espaço-temporal para Humanidades Digitais<br>Nº do processo FAPESP: 2016/04846-0 | 2º Edital | 01/02/2017 até 31/01/2019 |
| Sonia Maria Dozzi Brucki | Hospital das Clínicas de São Paulo / HC / SSSP | Avaliação da orientação topográfica em um ambiente de realidade virtual em pacientes com comprometimento cognitivo leve<br>Nº do processo FAPESP: 2016/04984-3 | 2º Edital | 01/10/2016 até 30/09/2018 |
| Cairo Lúcio Nascimento Júnior | Divisão de Engenharia Eletrônica / IEE / ITA | Aplicação de técnicas de inteligência computacional e de análise de big data em um experimento com sistemas multi-agentes na área de finanças<br>Nº do processo FAPESP: 2016/04992-6 | 2º Edital | 01/10/2016 até 30/09/2018 |
| Eloisa Dezen-Kempter | Faculdade de Tecnologia / FT / Unicamp | Uma estrutura para integrar dados de multissensores com Modelagem da Informação da Construção em apoio à Conservação e Gestão de Patrimônio Histórico<br>Nº do processo FAPESP: 2016/04991-0 | 2º Edital | 01/03/2017 até 28/02/2019 |

Fonte: A autora com fundamento dos Editais da FAPESP e informações da Biblioteca Virtual.

Do ponto de vista da política para a gestão de dados científicos, merece ser ressaltado que o Programa da FAPESP exige que o projeto submetido por meio do edital de Chamada de

Propostas de Pesquisa apresente "a explicação de como será sua política de gestão de dados, incluindo a especificação do tipo de dado gerado, a forma e o tempo que serão disponibilizados, o modo de preservação e os tipos de cuidados tomados em relação a questões de privacidade e ética". Conforme apresentação do Programa realizada em 2014, o edital apresentou novidades tais como a solicitação de *Data Management Policy*. Além disso, o projeto deveria envolver um cientista da computação e um de outro domínio, mostrar que a pesquisa seria desenvolvida em ambos os domínios, evidenciar treinamento em e-Science, apresentar política de disseminação e reuso de resultados dentre outros aspectos (FAPESP, 2014).

### 2.6.2 O Portal da Biodiversidade

O Portal de Biodiversidade teve em seu desenvolvimento o auxílio de pesquisadores da Escola Politécnica da USP, que conseguiram reunir em uma única interface de busca as informações de bancos de dados mantidos pelo ICMBio e pelo Jardim Botânico do Rio de Janeiro. O Portal oferece buscas textuais e geoespaciais, visualização e *download* de registro de ocorrências de espécies. Além disso, "já conta com mais de um milhão de registros (coordenadas geográficas) de espécies, resultantes da integração de nove bases de dados mantidas pelo ICMBio" (BRASIL. ICMBio, 2015).

No que diz respeito às preocupações do ICMBio com uma política de gestão dos dados de pesquisa, merece ser ressaltado a publicação da Instrução Normativa nº 03, de 01 de setembro de 2014 que, dentre outros, "[...] regulamenta a disponibilização, o acesso e o uso de dados e informações recebidos pelo Instituto de Informações Chico Mendes de Conservação e Biodiversidade por meio do SISBio". Além dessa, o ICMBio ainda publicou a Instrução Normativa nº 2 de 25 de novembro de 2015 que *"*Institui a política de dados e informações sobre biodiversidade do Instituto Chico Mendes de Conservação da Biodiversidade e dispõe sobre sua disponibilização, acesso e uso*"*.

### 2.6.3 Infraestrutura Nacional de Dados Espaciais no Brasil (INDE)

Uma infraestrutura de dados espaciais é descrita como um conjunto de tecnologias, políticas e arranjos institucionais que facilitam a disponibilidade e o acesso aos dados espaciais. Ela contribui para o estabelecimento de padrões, bem como a integração de informações geoespaciais.

A informação geoespacial (IG), por sua vez, é definida pela INDE (2017) como aquela que

> associa a cada entidade ou fenômeno uma localização na Terra, traduzida por sistema geodésico de referência, podendo ser derivado das tecnologias de levantamento, associadas a sistemas globais de posicionamento apoiados por satélites, bem como de mapeamento ou de sensoriamento remoto.

Augusto (2010) descreve a informação geoespacial como sensível e de extrema importância para áreas estratégicas tais como, o monitoramento ambiental, serviços de aviso de fenômenos da natureza (tornados, enchentes), proteção de florestas, previsão do tempo, monitoramento de mudanças climáticas dentre outros. Além disso, a autora enfatiza que esse tipo de informação na maioria dos casos é produzida, mantida e adquirida por organizações públicas em todas as esferas do governo.

Em âmbito mundial, no ano de 1994, Bill Clinton, por meio da Ordem Executiva 12906, criou a *National Spatial Data Infrastructure*, ato legal que reconheceu a importância da informação geoespacial ao afirmar que ela é "crítica para promover o desenvolvimento econômico, melhorar a nossa gestão de recursos naturais e proteger o meio ambiente".

Ainda em âmbito mundial, em 2002 tem início a preparação da Directiva INSPIRE[50] pela Comunidade Europeia. Na ocasião foram criados Grupos Temáticos para a) dados de referência e metadados, b) arquitetura e normas, c) políticas de dados e assuntos legais, d) estratégias de implementação e financiamento e e) análises de impacto.

A Diretiva entrou em vigor cinco anos depois, em 15 de maio de 2007 sob a designação Diretiva 2007/2/EC do Parlamento Europeu e do Conselho e 14 de março de 2007, publicada no Jornal Oficial das Comunidades, em 25 de abril de 2007, que estabelece a criação da Infraestrutura Europeia de Informação Geográfica. Esta Diretiva pretende promover a disponibilização de informação de natureza espacial, utilizável na formulação, implementação e avaliação das políticas ambientais da União Europeia. Percebe-se assim, a criação de um marco legal para tratar a informação geoespacial em âmbito mundial.

No Brasil, a INDE é criada por meio do Decreto nº 6.666 de 27 de novembro de 2008 que "institui, no âmbito do Poder Executivo federal, a Infraestrutura Nacional de Dados Espaciais - INDE, e dá outras providências".

Dentre os objetivos da INDE, o decreto estabelece:

[50] Trata-se de uma diretiva enquadradora que define as condições globais para a criação da Infraestrutura Europeia de Informação Geográfica e dá a possibilidade aos cidadãos europeus de facilmente encontrarem, através da Internet, informação útil em termos de Ambiente e outras temáticas, permitindo também que as autoridades públicas beneficiem mais facilmente de informação produzida por outras autoridades públicas.

> I - promover o adequado ordenamento na geração, no armazenamento, no acesso, no compartilhamento, na disseminação e no uso dos dados geoespaciais de origem federal, estadual, distrital e municipal, em proveito do desenvolvimento do País;
> II - promover a utilização, na produção dos dados geoespaciais pelos órgãos públicos das esferas federal, estadual, distrital e municipal, dos padrões e normas homologados pela Comissão Nacional de Cartografia (CONCAR); e
> III - evitar a duplicidade de ações e o desperdício de recursos na obtenção de dados geoespaciais pelos órgãos da administração pública, por meio da divulgação dos metadados relativos a esses dados disponíveis nas entidades e nos órgãos públicos das esferas federal, estadual, distrital e municipal.

Para atingir os objetivos dispostos, o próprio decreto já estabeleceu no §1º que seria "implantado o Diretório Brasileiro de Dados Geoespaciais – DBDG, que deverá ter no Portal Brasileiro de Dados Geoespaciais, denominado "Sistema de Informações Geográficas do Brasil – SIG Brasil", o portal principal para o acesso aos dados, seus metadados e serviços relacionados".

No Brasil, a INDE nasce com o propósito de:

> catalogar, integrar e harmonizar dados geoespaciais existentes nas instituições do governo brasileiro, produtoras e mantenedoras desse tipo de dado, de maneira que possam ser facilmente localizados, explorados e acessados para os mais diversos usos, por qualquer cliente que tenha acesso à Internet. Os dados geoespaciais serão catalogados através dos seus respectivos metadados, publicados pelos produtores/mantenedores desses dados (INDE, 2017).

As instituições brasileiras envolvidas na INDE são o Instituto Brasileiro de Geografia e Estatística (IBGE) e a Secretaria de Planejamento e Investimentos Estratégicos do Ministério do Planejamento, Orçamento e Gestão. A Gestão da INDE, por sua vez, é realizada por meio de um Conselho Superior, um Conselho Consultivo, um Comitê Técnico e Grupos de Trabalho.

Atualmente o Portal Brasileiro de Dados Espaciais já disponibiliza para consulta um conjunto de normas referentes a: a) padronização de marcos geodésicos, b) caracterização do Sistema Geodésico Brasileiro, c) parâmetros para transformação de Sistemas Geodésicos, d) recomendações para levantamentos relativos estáticos, e) normas técnicas da cartografia nacional, f) perfil de metadados Geoespaciais do e Brasil – (Perfil MGB) e, g) e-PING padrões de interoperabilidade de governo eletrônico.

A consulta à literatura internacional revela que as iniciativas para tratar a informação geoespacial iniciam na década de 1990. No Brasil, essas iniciativas ganham corpo na década de 2000, e a criação de um marco legal para o país exigiu a articulação de diferentes atores e instituições, em diferentes níveis governamentais. A partir de 2008, por meio da publicação do Decreto da INDE, percebe-se a criação de uma legislação específica para a área, que culminou

com a padronização de marcos geodésicos, dentre outras atividades, que enfim culminaram com a criação de padrões de interoperabilidade para esse tipo de informação.

No âmbito da informação sobre Ciências Espaciais em nível mundial, merece ser comentada a observação de Borgman (2015) de que a área segue o modelo OAIS, fato que certamente contribuiu para o amadurecimento da área no Brasil, culminando com a publicação da INDE.

### 2.6.4 Outras iniciativas de gestão de dados científicos no Brasil

Em termos de movimentação nas instituições de pesquisa em prol de impulsionar um movimento de apoio à ciência aberta, pode-se dizer que o IBICT vem liderando no País o movimento de acesso aberto à informação científica desde o início dos anos 2000. Hoje a instituição é considerada referência em projetos voltados ao movimento do acesso livre à informação científica e tecnológica. Exemplo desse compromisso é a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), lançada em 2002, que utiliza a tecnologia de arquivos abertos e integra sistemas de informação de teses e dissertações de instituições de ensino e pesquisa brasileiros. Além da BDTD, o Instituto apoia, desde 2009, em conjunto com a FINEP, a criação de repositórios institucionais abertos em universidades públicas e centros de pesquisas financiados com recursos públicos.

Outra iniciativa de destaque do Instituto é a formação da Rede de Serviços de Preservação Digital Cariniana, responsável pela preservação dos periódicos eletrônicos na plataforma OJS/SEER no Brasil e que tem planos de ampliar o projeto, abrangendo documentos de outros tipos e em variadas mídias.

A respeito das iniciativas do IBICT de acesso aberto à informação científica, merece ser ressaltado que o Instituto, apenas em novembro de 2015, se articulou e organizou o *Workshop Desafios no Contexto Contemporâneo para promover a nova ciência baseada em dados de pesquisa*. Participaram com destaque no evento o Projeto DataONE, a RNP, a EMBRAPA Satélites, o INPA, o ICMBio, a USP dentre outros. Durante o evento houve uma cobrança dos participantes para que o IBICT liderasse**,** perante o MCTI**,** a elaboração de um conjunto de diretrizes sobre a gestão de dados de pesquisa no Brasil. Desde então, o Instituto tem, de maneira ainda tímida, tentado se articular perante os demais *stakeholders* para mapear às necessidades dos pesquisadores quando a gestão de dados, bem como os principais pontos de uma política que norteie a gestão dos dados de pesquisa.

Em março de 2017, a RNP e o IBICT anunciaram uma chamada pública no valor de cento e dez mil reais com o objetivo de apoiar um Grupo de Trabalho, em Ciência da Informação e Ciência da Computação, sobre acesso aberto a dados de pesquisa, tendo como objetivo principal contribuir para a criação de um Programa Nacional de Acesso Aberto a Dados de Pesquisa que incentive o compartilhamento de dados entre pesquisadores. A expectativa das instituições é de que o resultado do Grupo de Trabalho identificará iniciativas de pesquisadores que já trabalham com dados abertos no Brasil.

Outros órgãos têm iniciado o desenvolvimento de seu repositório de dados, dentre eles merece ser citado como exemplo o Instituto de Energia Nuclear que já criou a plataforma CarpeDIEN[51] (Dados e Informações em Engenharia Nuclear) e aos poucos vem alimentado dados de pesquisas de energia nuclear do Instituto. Em sua página inicial, já se observa que o sistema oferece a busca pelo autor do dado, por assunto e data de publicação.

A respeito das iniciativas do IEN no âmbito da gestão de dados de pesquisa, merece destaque a publicação do *Guia de Gestão de Dados de Pesquisa*, em novembro de 2015, por Sayão e Sales (2015).

No que diz respeito à curadoria das informações (altamente técnicas) a serem inseridas nesses repositórios é importante ressaltar que no Brasil a profissão bibliotecário é de graduação. Portanto, o profissional que tiver como objetivo trabalhar como bibliotecário de dados, ou cientista de dados, terá que se aprofundar no tema que escolher trabalhar, seja ele energia nuclear, infraestrutura para os dados espaciais, ou biodiversidade, por exemplo. Nesse sentido, parece prudente que as instituições reflitam sobre o modelo de organização do Centro de Informação Nuclear (CIN) da Comissão Nacional de Energia Nuclear (CNEN) que já na década de 1970 trabalhava com uma equipe multidisciplinar na biblioteca.

## 2.7 CONTEXTO DA *E-SCIENCE* NA CIÊNCIA DA INFORMAÇÃO

No contexto da pesquisa cientifica, o surgimento do "quarto paradigma[52]" traz impactos profundos sobre a ciência e, portanto, exige um profundo exame das funções das instituições empenhadas no avanço da ciência e no apoio aos cientistas, dentre elas as bibliotecas.

---

[51] Disponível em: <http://carpedien.ien.gov.br/>. Acesso em: 2 out. 2016.

[52] Para Gray (2007) os outros paradigmas classificam-se três, sendo o primeiro representado pela ciência há mil anos atrás, marcada pela descrição de fenômenos naturais. O segundo paradigma representa a ciência realizada há poucos séculos, onde está presente o uso de modelos e generalizações, e o terceiro paradigma é representado pela ciência das últimas décadas como o auxílio computacional para a simulação de fenômenos complexos. Todos esses fenômenos têm em comum o aumento do volume de dados coletados. Assim, para Gray (2007), o quarto paradigma coloca a computação como um elemento de sustentação da ciência realizada no Século XXI.

No Google Acadêmico, em junho de 2014, ao se pesquisar sobre *e-science* e bibliotecas, os artigos recuperados em destaque são de Tony Hey, o primeiro, *The data deluge: an e-science perspective*, citado por 405 trabalhos e; o segundo: *E-science and its implications for the library community*, citado por 70 trabalhos. Em função do envolvimento de seu atual diretor de pesquisa com a *e-science*, parece natural que a Microsoft Corporation tenha despertado seu interesse para essa nova e crescente área.

O aspecto transversal da Ciência da Informação faz com que, em algum momento, os dados oriundos da *e-science* convirjam para as preocupações da comunicação da informação e, em outros momentos, para questões inerentes à organização da informação. No que se refere à comunicação da informação, Sayão e Sales (2012) argumentam que:

> [...] dados e informações digitais gerados pelas atividades de pesquisa necessitam de cuidados específicos, tornando-se necessário a criação de novos modelos de custódia e gestão de conteúdos científicos digitais que incluam ações de arquivamento seguro, preservação, formas de acrescentar valor a esses conteúdos e de otimização da sua capacidade de reuso [...]. É nesse ambiente que surge o conceito de curadoria digital de dados científicos.

A literatura norte-americana já revela uma preocupação dos bibliotecários com esse novo cenário. Luce (2010, p. 3) argumenta que para as "bibliotecas universitárias a evolução gradual da *e-science* provoca desafios profundos e, ao mesmo tempo, proporciona às bibliotecas uma oportunidade de redefinir seus papéis e agregar valor ao seu portfólio de serviços".

Consciente do impacto e das oportunidades para as bibliotecas universitárias, a Association of Research Libraries (ARL) criou uma Força Tarefa *e-Science* (*e-Science Task Force*), em 2006, que definiu o domínio da *e-science*. Essa força tarefa foi seguida por um grupo de trabalho contínuo que teve como missão desenvolver a compreensão dos membros para as mudanças de habilidades profissionais e infraestruturas necessárias para o tratamento de um novo tipo de dado – o oriundo da *e-science* (SOEHNER; STEEVES; WARD, 2010).

A identificação de diferentes abordagens sendo empreendidas por instituições isoladas (EUA, Canadá, Reino Unido) para a compreensão do fenômeno da *e-science* incitou a ARL a desenvolver um levantamento, em 2009, com o objetivo de identificar o envolvimento das bibliotecas com a questão do tratamento dos dados oriundos da *e-science*. O instrumento de coleta de dados foi enviado para 123 bibliotecas membros da ARL nos EUA e Canadá. Dentre as indagações do questionário constava a seguinte pergunta: *Serão os bibliotecários aqueles que intervirão e enfrentarão o desafio*? Foram obtidas respostas de 57 bibliotecas membros da ARL, destacando-se que: 21 bibliotecas afirmaram fornecer infraestrutura ou serviço para *e-*

*science*, 23 bibliotecas afirmaram que estão planejando oferecer esse tipo de serviço, e 13 bibliotecas afirmaram não oferecer suporte para *e-science*. Além disso, o levantamento demonstrou que, entre as bibliotecas respondentes, 42% contrataram e 39% planejam contratar membros de equipe com habilidades em *e-science* (SOEHNER; STEEVES; WARD, 2010).

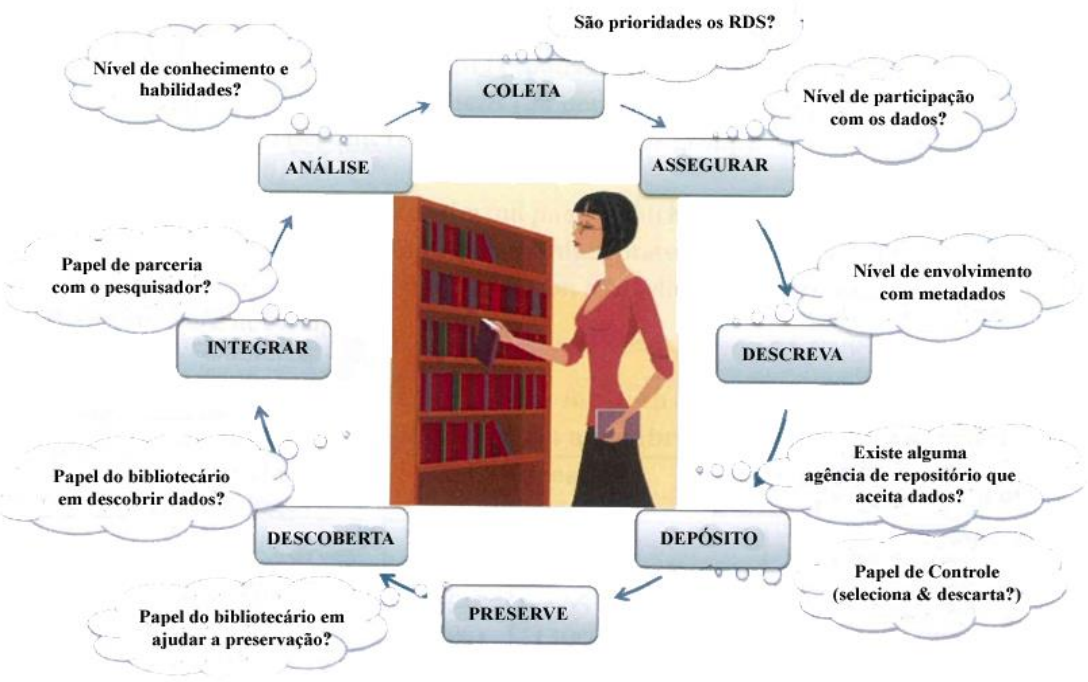
A ARL tem realmente se dedicado a compreender o fenômeno da *e-science*. Em pesquisa nas base de dados *Library and Information Science Abstract* (LISA) e *Library and Information Science & Technology Abstracts* (LISTA) sobre *e-science* foi constatado que a instituição publicou, pelo menos, sete trabalhos, sendo eles: 1) *ARL launches e-science agenda*; 2) *E-science and data support services: a study of ARL members institutions*; 3) *E-science in research libraries: new ARL reports to research libraries*; 4) *Reinventing librarianship: themes from ARL-CNI Forum*; 5) *ARL activities*; 6) *ARL e-science survey* e 7) *Libraries and changing research practice: a report of the ARL-CNI Forum on e-research and cyberinfrastructure*.

Para Álvaro *et al* (2011), a *e-science* pode fornecer um campo potencial para bibliotecários ramificarem-se para além dos limites das práticas tradicionais de biblioteca. Na visão dos autores, a *e-science* não é prática comum e, em função disso, a Biblioteconomia deverá prosseguir neste novo território com cautela.

Luce (2010) argumenta que com visão, investimentos estratégicos e com uma "alavancagem" de sua *expertise* em gestão da informação, as bibliotecas digitais podem se tornar um recurso essencial para o tratamento digital da informação oriunda da *e-science*, que deve estar disponível para a próxima geração de comunidades de pesquisa. Da mesma forma, Soehner, Steeves e Ward (2010, p. 7) comentam que "curadoria de dados, preservação, acesso e metadados são áreas da *e-science* onde as bibliotecas encontram uma afinidade natural".

O momento mostra-se oportuno tanto para profissionais da informação, como para profissionais da tecnologia de informação, dentre outros. Os limites dessa nova área ainda não estão definidos, por isso faz-se necessário que o bibliotecário posicione-se como um profissional que apresenta capacidades para lidar com o tratamento de dados oriundos da *e-science*. Como exemplo dessa capacidade, cita-se o ciclo do bibliotecário, proposto por Tenopir, Birch e Allard (2012, p. 12), ilustrado na Figura 8.

Figura 8 – Bibliotecário ponderado *se* possui as habilidades para oferecer pesquisa de dados.




Fonte: Tenopir, Birch, Allard (2012, p. 12, tradução nossa).

## 2.8 POLÍTICA BRASILEIRA EM CIÊNCIA & TECNOLOGIA

A política de ciência e tecnologia no Brasil é trabalhada, entre outros, por Morel (1979), Meis e Leta (1996), Lopes, (1998), Baumgarten (2001), Schwartzman (2001), Correa (2003), Muniz (2008), Rezende (2010), Videira (2010), dentre outros. A relação entre soberania nacional e ciência e tecnologia é analisada por Amaral (2001; 2004). A partir dos autores acima citados será apresentado um breve histórico sobre a política brasileira em ciência e tecnologia.

O Brasil, a partir de 1930, inicia um forte processo de industrialização com a participação do governo, por meio da estatização de serviços de infraestrutura e a participação em áreas estratégicas como a siderúrgica, petrolífera e de extração de minérios e incentivo financeiro público (PELAEZ; SZMRECSÁNYI, 2006). Para Schwartzman (2001, p. 30) o ano de 1930 é considerado pela historiografia brasileira como a "data em que o Brasil ingressou no mundo moderno".

O período rompe com o modelo de um país agrário exportador e é marcado por importantes modificações sociais, políticas e econômicas que repercutiram nas medidas da política educacional ou científica. Importante ressaltar que essas modificações deixaram latente a necessidade de recursos humanos especializados para atuar na burocracia pública, bem como nos setores industriais (CORREA, 2003; MOREL, 1979).

Entre 1930 e 1949, com a criação de várias universidades, dentre elas a Universidade de São Paulo em 1934, é observada uma expansão quantitativa da oferta do ensino superior. Porém, para Morel (1979, p. 40) a "ação estatal nessa época se restringiu à criação de estabelecimento de ensino superior, não há ainda uma definição em relação à ciência propriamente dita". A autora (MOREL, 1979, p. 41) argumenta que as transformações no sistema produtivo acabaram por evidenciar a necessidade "de se desenvolver o sistema científico, tecnológico nacional".

No final da década de 1940, duas importantes iniciativas partem da comunidade científica brasileira – a criação, em 1948, da Sociedade Brasileira para o Progresso da Ciência (SBPC) e a criação do Centro Brasileiro de Pesquisas Físicas (CBPF), em 1949 (CORREA, 2003; MOREL, 1979). Na visão de Correa (2003), no final da década de 1940 e no início da década de 1950 foram criadas importantes instituições que vieram a fazer parte, em um momento posterior, do chamado Sistema Nacional de Desenvolvimento Científico e Tecnológico.

A percepção de Videira (2010) é de que, no Brasil, há um consenso entre os autores de considerar o ano de 1951 como o marco do desenvolvimento da ciência e da tecnologia no país, pois, nesse ano foram criados o Conselho Nacional de Pesquisas (CNPq) e a Campanha Nacional de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

A Lei nº 1.310, de 15 de janeiro de 1951, cria o Conselho Nacional de Pesquisas, hoje Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), com o objetivo de promover e estimular o desenvolvimento da investigação científica e tecnológica em qualquer domínio do conhecimento. Para Morel (1979, p. 45), a criação do CNPq "foi orientada pela necessidade [de o] Brasil se equiparar às outras nações na pesquisa de energia nuclear[53], elemento que a Segunda Guerra Mundial [demonstrou] ser de vital importância para a segurança nacional".

No mesmo ano, por meio do Decreto nº 29.741, de 11 de julho de 1951, também é criada a Campanha Nacional de Aperfeiçoamento de Pessoal de Nível Superior, hoje Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES). Ela foi criada com o objetivo de "assegurar a existência de pessoal especializado em quantidade e qualidade suficientes para atender às necessidades dos empreendimentos públicos e privados que visam o

[53] Esta tese não abordará as polêmicas questões que envolvem a exploração e venda de urânio no período de 1940 a 1975. Entende-se que o período foi marcado pela exportação de urânio, tório e outros materiais físseis, de forma negligente para os EUA, por meio de acordos que prejudicaram o desenvolvimento de uma política nuclear brasileira.

desenvolvimento do país." Essa fundação do Ministério da Educação desempenhou, e ainda desempenha, papel fundamental na expansão e consolidação da pós-graduação *stricto sensu* em todos os estados da Federação (CAPES, 2013; MOREL, 1979).

Ainda que essas iniciativas tenham sido tardias, elas foram fundamentais para que, a partir da década de 1950, ocorresse a instalação de indústrias automobilísticas[54], bem como da indústria naval e da indústria pesada de máquinas e equipamentos elétricos. A partir de então, pode-se formar a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) e o Instituto Tecnológico da Aeronáutica (ITA). Além disso, a Petrobrás pode usufruir de engenheiros de várias especialidades, que ajudaram a estabelecer, internacionalmente a sua liderança tecnológica na exploração de petróleo (PELAEZ; SZMRECSÁNYI, 2006; DAVIDOVICH, 2011).

Para Morel (1979) e Correa (2003), o ano de 1951 (pós-guerra) marca a institucionalização da política científica no Brasil. As autoras comentam que a década de 1950 foi marcada pelo investimento na formação de recursos humanos qualificados. Em contrapartida, ambas argumentam que durante a década de 1960 a política de C&T foi caracterizada por medidas descontínuas, com quedas de investimentos no período de 1956 a 1961. Além disso, nesse período houve uma fuga de cérebros brasileiros para o exterior, em função dos baixos salários e falta de condições de trabalho.

O golpe militar de 1964 acabou por alterar o modelo político e econômico então adotado pelo Brasil. Morel (1979, p. 51) observa duas tendências que norteiam a política científica nesta época "do lado da 'segurança', o cerceamento de manifestações de crítica ao governo; do lado do 'desenvolvimento', a ênfase na pesquisa científica e na formação de cientistas e profissionais especializados, como elementos indispensáveis ao crescimento econômico, e a criação de um Brasil grande potência".

Vale a pena comentar que, apesar do volume de investimentos feitos na ciência durante o governo militar, professores universitários e pesquisadores de institutos de pesquisa governamentais foram perseguidos no período da ditadura no Brasil, alguns chegando a ter seus direitos políticos cassados, sendo obrigados a deixar o país e, outros sendo exonerados de cargos públicos (SCHWARTZMAN, 2001, p. 15).

Durante o governo do presidente Castelo Branco[55], percebe-se que o desenvolvimento do país depende de dois sistemas institucionais: o das universidades e o dos institutos de

[54] Em 1956 é criado o primeiro carro nacional, Romi-Isetta com 236cc e três lugares. Em 1957 o Chevrolet 3100 e o Ford F-60 passam a ser fabricados no Brasil (ANFEA, 2006).
[55] Castelo Branco governou o Brasil de 1964 a 1967.

pesquisa, o que coloca o sistema científico e a formação de recursos humanos qualificados em posição estratégica (MOREL, 1979).

Morel (1979) relembra que o artigo 179 da Constituição de 1967 mantém o dever do poder público de incentivar a pesquisa e o ensino científico e tecnológico. Para a autora, o ano de 1967 é marcado pela vinculação da ciência e tecnologia à política externa nacional como elemento essencial à soberania do país.

De acordo com Morel (1979), durante o governo do presidente Costa e Silva[56] é possível observar-se um incremento nas medidas relacionadas à política científica do país. Esse governo considerou o estímulo à pesquisa científica e tecnológica como instrumento de aceleração do desenvolvimento, fato que resultou na incorporação do termo "ciência e tecnologia" ao discurso governamental de uma forma até então não vista na História do Brasil. A literatura revela que com o investimento em C&T, realizado durante a ditadura militar, o governo procurou formar um parque industrial no Brasil com ênfase na produção de urânio e de energia nuclear.

O Decreto Lei nº 719, de 31 de julho de 1969, cria o Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT) com o objetivo de dar apoio financeiro aos programas e projetos considerados prioritários pelo governo brasileiro. De acordo com Correa (2003) o FNDCT substituiu o Fundo de Desenvolvimento Técnico Científico (FUNTEC).

Schwartzman (2001) descreve a década de 1970 como "anos do milagre[57]" e considera que o governo do presidente Ernesto Geisel[58] foi marcado pela abundância de recursos para a ciência e tecnologia, bem como pela facilidade de aprovação de projetos. Porém, na percepção de Videira (2010), a influência do Ministro João Paulo dos Reis Veloso junto ao presidente Ernesto Geisel foi a responsável pelo aumento dos recursos para financiamento de pesquisas e formação de recursos humanos. Assim, Videira (2010, p. 110) expõe que a "situação da ciência no Brasil entre 1974 e 1979 foi apenas razoável".

Morel (1979) observa que, entre 1967 a 1970, a política científica é integrada ao planejamento global do Estado – uma vez que o avanço tecnológico passa a ser uma condição *sine qua non* para projetar o Brasil como grande potência mundial.

[56] Costa e Silva governou o Brasil de 1967 a 1969.
[57] O termo utilizado entre aspas por Schawartzman (2001, p. 14) representa o grande paradoxo do desenvolvimento econômico, bem como da euforia nacional devido à conquista da Copa de 1970 em relação a persistência da pobreza, desigualdade social e a repressão política vivida no país.
[58] Ernesto Geisel foi presidente do Brasil de 1974 a 1979.

Correa (2003) considera que o objetivo central da política, no período de 1967 a 1973, foi manter as elevadas taxas de crescimento econômico, que, por sua vez, exigiam avanços significativos no grau de capacitação do país para viabilizar a criação e adaptação de tecnologias. Nesse sentido, a autora entende que o período foi significativo em termos de montagem da infraestrutura de C&T no país. Por outro lado, o setor manteve-se isolado da iniciativa privada.

Dentre as iniciativas de destaque da década de 1970, Morel (1979) e Correa (2003) ressaltam que a criação do Sistema Nacional de Desenvolvimento Científico e Tecnológico (SNDCT), por meio do Decreto nº 225, de 15 de janeiro de 1975, possibilitou a organização das diversas fontes de recursos alocados pelo governo brasileiro para as atividades de ciência e tecnologia.

Na percepção de Correa (2003, p. 129), a redução do investimento público na área de C&T, durante a década de 1980, acabou por inibir o setor científico e tecnológico, que ainda não estava consolidado e já iniciava o "movimento inverso, rumo ao sucateamento e à desagregação".

Videira (2010) corrobora a percepção de Correa (2003) ao considerar que houve um desprezo para com a ciência e a tecnologia durante o governo do presidente João Batista de Oliveira Figueiredo[59]. Nas palavras do autor: "os recursos começaram a diminuir progressivamente, dando início a um processo de sucateamento dos institutos de pesquisas e [das] universidades" (VIDEIRA, 2010, p. 112).

A percepção de Schwartzman (2001) é de que a pesquisa militar gerou gigantescos investimentos públicos. Porém, o fim da Guerra Fria pareceu anunciar o fim dos investimentos públicos na pesquisa militar. Surgiu nessa época a necessidade de a academia estabelecer parcerias com o setor privado, o que deu origem a um novo modo de produção científica "muito mais pragmático, interdisciplinar, *ad hoc* e contaminado por interesses comerciais e empresarias do que antes" (SCHWARTZMAN, 2001, p. 6).

É possível afirmar que a crise do petróleo, a incapacidade de financiamento interno do país e o consequente crescimento da dívida externa, aliados à política norte americana de elevação das taxas de juros, comprometeram a política de desenvolvimento científico e tecnológico no Brasil durante as décadas de 1970 e 1980. Nas palavras de Schwartzman (2001, p. 4), a explicação típica para o país não ter alcançado o grande salto para frente é que "o Brasil

[59] João Batista de Oliveira Figueiredo foi presidente do Brasil no período de 1979 a 1985.

teve azar uma vez, vítima que foi do aumento no preço do petróleo, das elevadas taxas de juros e da queda das cotações internacionais de seus produtos básicos de exportação".

De acordo com Correa (2003, p. 162), as décadas de 1970 e 1980 foram marcadas pelo "distanciamento entre a pesquisa básica e a pesquisa tecnológica, mantendo-se a coletividade acadêmica distanciada das demandas sociais e do setor produtivo". Tal situação resultou em um estímulo à importação de tecnologia (*know how*).

Para Rezende (2010, p. 175-176), "a política industrial dos anos 1970 e 1980 não foi articulada com programas de C&T". Ela foi baseada em um "modelo de substituição de importações, imposto por barreiras comerciais e tarifárias, que dificultavam ou impediam a importação de produtos similares". Para o autor, esse modelo culminou com a reserva de mercado de informática por meio da Lei nº 7.232, de 29 de outubro de 1984. Dentre as críticas, Rezende (2010, p. 176) argumenta que o modelo "estimulou a criação de empresas nacionais no setor [de informática], superprotegidas da concorrência externa, porém, sem programas consistentes de capacitação tecnológica".

Na percepção de Morel (1979), Baumgarten (2001) e Correa (2003), o conjunto dessas crises abriu espaço para a implantação da política neoliberal no Brasil. Nesse contexto, surge a denominada agenda para a competitividade do país, que traz consigo o discurso de que "as necessidades do setor privado da economia requerem um novo papel da investigação acadêmica e das universidades" (CORREA, 2003, p. 253).

Em 1985, no primeiro dia da Nova República, é criado o Ministério da Ciência e Tecnologia (MCT). Há que se ressaltar que o MCT já havia sido criado por meio do Decreto Lei nº 2000 de 1967, mas o Ministério só veio a ganhar vida institucional com a eleição o presidente Tancredo Neves (VIDEIRA, 2010).

Para Videira (2010), o MCT nasceu com duas vulnerabilidades: a primeira refere-se ao fato de o Ministério ter sido criado por um presidente que não chegou a tomar posse, mas apesar disso, o então presidente José Sarney manteve as escolhas feitas por Tancredo Neves para o Ministério, inclusive mantendo a indicação de Renato Archer para ocupar a pasta, apesar da rivalidade[60] existentes entre ambos na política regional. A segunda está relacionada à falta de planejamento das atividades da organização.

[60] De acordo com Videira (2010, p. 35), "as disputas políticas progressistas entre ambos [Sarney e Archer], oriundos do Maranhão, poderiam gerar problemas". O autor relata que Archer tentou nomear Fabio Celso Guimarães para a presidência da FINEP. Porém, apesar de Sarney ter aprovado a nomeação, ela não foi publicada sob o pretexto de que Tancredo Neves tinha outro nome para o cargo. Diante da situação, Renato Archer ameaça solicitar sua exoneração, mas uma comitiva de deputados maranhenses, em conjunto com José Sarney, sela um acordo de paz entre ambos.

Schwartzman (2001), por sua vez, considera que a criação do MCT em 1985 não foi suficiente para assegurar à comunidade científica brasileira todo o espaço, reconhecimento e apoio que ela esperava receber. Na visão do autor, "um regime político aberto não conduz necessariamente a um enfoque progressista em matéria de ciência, tecnologia e educação" (SCHWARTZMAN, 2001, p. 5).

Sobre o período, Muniz (2008) e Videira (2010) consideram que Sarney teve decisões desastrosas no âmbito da ciência e tecnologia, dentre elas a extinção do ministério, por meio da Medida Provisória nº 029 de 15 de janeiro de 1989. Apesar da turbulência política no âmbito do ministério, para Videira (2010, p. 129), o governo de Sarney apresentou uma "expansão exponencial do número de bolsas de pós-graduação".

A comunidade científica reagiu à extinção do MCT e com a promulgação da Constituição de 1988, o Congresso Nacional passou ter poderes para legislar sobre a ciência e tecnologia. Videira (2010, p. 126) argumenta que "com a pressão exercida pela comunidade científica, em alianças com políticos, a MP caducou, e o MCT foi transformado em uma secretaria especial, ligada diretamente à Presidência da República".

Na visão de Videira (2010, p. 127) o governo do presidente Fernando Collor de Mello "mostrou profunda desconsideração política pelo setor [de ciência e tecnologia], acompanhada de redução brutal no orçamento para a área de C&T". Dentre as medidas notoriamente criticadas, ressalta-se a extinção[61] e, passado um mês, a recriação da CAPES. O autor argumenta que a política neoliberal do presidente desmantelou a política de C&T do país. De acordo Correa (2003), o governo de Collor foi marcado pela busca de um ambiente favorável à entrada de capital estrangeiro no país.

Itamar Franco assume a presidência do país em 1992, após o *impeachment* de Fernando Collor de Mello. De acordo com Videira (2010), o governo de Itamar apresentou ligeira, porém não suficiente, recuperação orçamentária para a ciência e tecnologia. De acordo com Lima (2011), Itamar Franco priorizou a economia do país, alcançando uma queda da inflação ao final de seu governo. Para o autor, esse período revela uma ausência de conectividade entre a política econômica e a política de C&T.

Para Videira (2010) o primeiro mandato do presidente Fernando Henrique Cardoso foi marcado pela estagnação na área de ciência e tecnologia. Em contrapartida, durante o segundo mandato de FHC ocorreu a criação dos Fundos Setoriais de Desenvolvimento Científico e Tecnológico que viabilizou a recuperação da capacidade de investimento na área de C&T e

[61] Extinta pela Medida Provisória nº 150 de 15 de março de 1990.

áreas de pesquisa consideradas estratégicas pelo governo, dentre elas a biotecnologia, a nanotecnologia, os recursos naturais e as tecnologias de informação e comunicação.

A literatura revela que a conjuntura política e econômica da época foi favorável à criação dos Fundos Setoriais. Dentre os fatores que contribuíram para a sua criação, merecem destaque: a irregularidade do financiamento do Estado para o desenvolvimento científico e tecnológico, o enxugamento do Estado por meio das privatizações ocorridas à época (por exemplo, o setor de telecomunicações), bem como a experiência proporcionada pela Lei nº 9.478, de 06 de agosto de 1997, que criou o CT-Petro e regulamentou que uma parcela dos *royalties* da produção do petróleo e gás fossem destinados ao MCT para financiar pesquisas na área.

Na percepção de Araújo *et al.* (2012, p. 7) os fundos setoriais foram instituídos com "o propósito de criar condições mais estáveis de financiamento público às atividades de ciência, tecnologia e inovação no Brasil". Para os autores, a legislação brasileira procurou "evitar que restrições de natureza fiscal causassem descontinuidades nas políticas de CT&I adotadas pelo governo federal", a exemplo dos cortes orçamentários para C&T que ocorreram nas décadas de 1970, 1980 e 1990.

Em 2014, o Brasil contava com 16 Fundos Setoriais, sendo 14[62] relativos a setores específicos e dois transversais. Dentre os transversais, um é voltado à interação universidade-empresa (FVA – Fundo Verde-Amarelo), enquanto o outro é destinado a apoiar a melhoria da infraestrutura de instituições científico e tecnológicas (CT-Infra).

Apesar da criação dos fundos setoriais, durante o governo FHC houve uma queda do número de bolsas de pesquisa oferecidas pelo CNPq. A respeito do assunto, Rezende (2010, p. 178) comenta que "com o recente corte determinado pelo pacote econômico do governo, o orçamento do CNPq cai de R$490 milhões [de reais] para cerca de R$ 400 milhões e, em consequência, o número de bolsas em 1998 deverá ser 10% menor". A percepção de Rezende (2010) é que durante os governos Collor, Itamar Franco e FHC houve uma desnacionalização do parque industrial brasileiro, resultando na importação de tecnologia. O autor também critica os fundos setoriais e a própria gestão do MCT durante o governo de Fernando Henrique Cardoso. Nas palavras de Rezende (2010, p. 212):

> [...] se por um lado os grupos que atuam nas áreas contempladas pelos fundos setoriais foram fortemente apoiados, a grande maioria dos pesquisadores que trabalham em outras áreas e não estão no Estado de São Paulo ficou sem oportunidade de conseguir financiamento para as suas atividades de pesquisa.

[62] CT - Aeronáutico, CT - Agronegócio, CT - Amazônia, CT - Aquaviário, CT - Biotecnologia, CT - Energia, CT - Espacial, CT - Hidro, CT - Info/Cati, CT - Mineral, CT - Petro, CT - Saúde, CT - Transporte e Funtell.

Há um consenso na literatura de que ao final do segundo mandato do presidente FHC foi realizado um importante evento de C&T – a 2ª Conferência Nacional de C,T&I que culminou com a publicação de dois documentos de extrema relevância para a política na área– o *Livro Verde* e o *Livro Branco*, publicados em 2001 e 2002 respectivamente. O Livro Verde contém as "metas de implementação do Programa Sociedade da Informação" (LIVRO VERDE, 2001, p. *v*), ou seja, um delineamento das ações a serem implementadas para viabilizar o ingresso do Brasil na Sociedade do Conhecimento. O Livro Branco, por sua vez, estabelece de forma sistematizada as diretrizes estratégicas para C, T & I até o ano de 2010.

Vieira (2010) considera que o governo do presidente Lula teve notável expansão dos recursos financeiros de ciência e tecnologia. Dentre os pontos positivos do governo para C&T, o autor destaca a criação dos Institutos Nacionais de Ciência e Tecnologia (em substituição aos Institutos do Milênio), a aprovação da Lei de Inovação[63] em 2004 e da Lei do Bem[64] em 2005, bem como a elaboração, em 2007, do Plano de Ação da Ciência, Tecnologia e Inovação (PACTI) para o período de 2007 a 2010, conhecido como PAC da Ciência. Além disso, a sociedade tem interagido com o MCT por meio da Semana Nacional de Ciência e Tecnologia, das Olimpíadas de Matemática e das Conferências Nacionais de Ciência e Tecnologia[65], realizadas respectivamente em 1985, 2001, 2005 e 2010.

O governo de Lula contou com três Ministros de Ciência e Tecnologia. O primeiro, Roberto Amaral (01/01/2013 – 21/01/2004), importante político de refundação do Partido Socialista Brasileiro (PSB) em 1984, teve seus ideais formados durante a Guerra Fria. Não à toa, suas iniciativas à frente do MCT voltaram-se para questões de soberania nacional e energia nuclear. Porém, introduziu no ministério uma política voltada para a redistribuição dos recursos na área de ciência e tecnologia, querendo assim fomentar novos polos de desenvolvimento científico e tecnológico.

O segundo Ministro, Eduardo Campos (23/01/2004 – 18/07/2005), assumiu o Ministério após Roberto Amaral, o que garantiu ao PSB a permanência da Pasta. Foi o mais novo ministro

[63] Lei nº 10.973, de 2 de dezembro de 2004, dispõe sobre incentivos à inovação e à pesquisa científica e tecnológica no ambiente produtivo e dá outras providências.

[64] Lei nº 11.196, de 21 de novembro de 2005 que institui o Regime Especial de Tributação para a Plataforma de Exportação de Serviços de Tecnologia da Informação, o Regime Especial de Aquisição de Bens de Capital para Empresas Exportadoras, e o Programa de Inclusão Digital; dispõe sobre incentivos fiscais para a inovação tecnológica e dá outras providências. É conhecida como Lei do Bem por criar a concessão de incentivos fiscais às pessoas jurídicas que realizarem pesquisa e desenvolvimento de inovação tecnológica.

[65] De acordo com Rezende (2010a, p. 17) "as conferências nacionais de ciência e tecnologia têm historicamente oferecido à sociedade um espaço democrático para se manifestar sobre suas propostas e aspirações para o setor. Não por acaso, o próprio Ministério da Ciência e Tecnologia nasceu sob a égide da Primeira Conferência, convocada em 1985 pelo primeiro titular da Pasta, o saudoso ministro Renato Archer, preocupado em ouvir a sociedade sobre os rumos que o novo ministério deveria tomar".

da casa, mas ao mesmo tempo é visto como um dos ministros mais fortes que já passaram pelo MCT. Isso se deve ao fato de Eduardo Campos, à época ser presidente do PSB, político respeitado e com grande expectativa de poder, tanto que em 2014 lançou sua candidatura à Presidência da República, tendo como vice Marina Silva. Até sua morte, durante a campanha eleitoral, Campos era considerado um candidato viável, se não para 2014, pelo menos para a eleição seguinte de 2018.

Um político com essa envergadura deu mais visibilidade para o MCT. Eduardo Campos trabalhou no âmbito do MCT questões de inclusão social e tecnologias assistivas. A exemplo cita-se o Projeto Chapéu de Couro. Para driblar as restrições orçamentárias, valeu-se de sua condição de ex-congressista para obter mais recursos financeiros para o Ministério por meio de emendas parlamentares. Além disso, conseguiu importantes aprovações no Congresso Nacional para área de ciência e tecnologia. A exemplo cita-se a aprovação da Lei de Inovação Tecnológica.

O sucessor de Eduardo Campos foi Sérgio Rezende (19/07/2005 – 31/12/2010), físico, com doutorado pelo MIT, respeitado no meio acadêmico, mas sem a força política de seu antecessor. Com a sua nomeação Lula manteve o MCT com o PSB, garantido assim a composição de sua base aliada no governo. Dentre os programas de destaque do ministro Sérgio Rezende tem-se a criação do Sistema Brasileiro de Tecnologia (SIBRATEC), instituído por meio do Decreto nº 6.259/07 com a finalidade de apoiar o desenvolvimento tecnológico do setor empresarial nacional. As atividades apoiadas pelo programa estão consonância com as prioridades estabelecidas nas políticas industrial, tecnológica e de comércio exterior do Brasil.

O governo da Presidente Dilma Rousseff dá continuidade à política de desenvolvimento científico e tecnológico do ex-presidente Lula. Durante a entrega do prêmio Jovem Cientista, ocorrida no dia 18 de dezembro de 2012, Dilma, em seu discurso (ABEL, 2012), afirmou que "[...] sem ciência, tecnologia e inovação, nós não seremos essa nação desenvolvida e não seremos esse país que sepultou, em definitivo, a pobreza extrema e a pobreza".

Apesar da ênfase em seu discurso, durante a vigência de seu governo, foram nomeados cinco Ministros de Estado de Ciência, Tecnologia e Inovação, resultando na média de um ministro por ano, o que consequentemente contribuiu para as dificuldades de obtenção de recursos para a área, bem como a descontinuidade de projetos da Pasta. Foram eles: Aloísio Mercadante (01/01/2011 a 24/01/2012), Marco Antônio Raupp (24/01/2012 a 17/03/2014), Clélio Campolina Diniz (17/03/2014 a 01/01/2015), Celso Pansera (02/10/2015 a 14/04/2016), tendo como Ministro Interino, Emília Curi no período de 14/04/2016 a 12/05/2016.

Fato é que nenhum dos Ministros do MCTI, durante o governo Dilma Roussef, foi escolhido por ter envergadura na área de C, T & I, mas sim para compor a base de apoio político parlamentar do governo.

Dentre as muitas críticas aos nomeados de Dilma para o Ministério da Ciência Tecnologia e Inovação destaca-se que o único Ministro envolvido com ciência, tecnologia e inovação foi o Dr. Marco Antonio Raupp. Por outro lado, inúmeras foram as críticas a Aloísio Mercadante, retirado do Ministério da Educação para atender conveniências políticas partidárias do governo. Dentre as críticas ao Ministro Mercadante, destaca-se o fato de ter publicado uma "Estratégia Nacional de Ciência, Tecnologia e Inovação", rompendo assim a tradição de publicação do "Plano de Ação da Ciência, Tecnologia e Inovação (PACTI)", vista pelos funcionários da casa como uma medida que enfraqueceu a posição do Ministério na disputa por recursos públicos. O Ministro Celso Pansera, também colocado à frente do Ministério em função da necessidade de composição política do governo, foi duramente criticado por ser um aliado do deputado Eduardo Cunha, controvertido ex-presidente da Câmara que teve seu mandato cassado.

Após o *impeachment* da Presidente Dilma Rousseff, assume o Presidente Michel Temer, que em decorrência da profunda crise política e econômica enfrentada pelo país durante os anos de 2015 e 2016, conduziu uma ampla reforma administrativa no Poder Executivo no intuito de enxugar ministérios e cargos comissionados com o objetivo de reduzir gastos do governo. Durante a reforma, o número total de Ministério passou de 32 para 24. Ressalta-se que alguns ministérios tiveram sua estrutura integrada a outro ministério. Como exemplo cita-se a fusão do Ministério do Desenvolvimento Agrário com o Ministério do Desenvolvimento Social e Combate à Fome, que deu origem ao Ministério do Desenvolvimento Social. Além dele, têm-se o caso da fusão do Ministério da Ciência, Tecnologia e Inovação com o Ministério das Comunicações, dando origem no dia 1 de maio de 2016 ao Ministério da Ciência, Tecnologia, Inovações e Comunicações tendo a sua frente o Ministro Gilberto Kassab.

Kassab desde o momento de sua posse enfrentou duras críticas à fusão dos dois ministérios, bem como a falta de recursos para a Pasta. As reivindicações tiveram origem em diversas entidades de classe, dentre elas a SBPC, a Academia Brasileira de Ciências (ABC), o Conselho Nacional das Fundações Estaduais de Amparo à Pesquisa (CONFAP), o Conselho Nacional de Secretários Estaduais para assuntos de Ciência, Tecnologia e Inovação (CONSECTI) etc.

A respeito do posicionamento e luta dessas entidades contra a fusão do MCTI com o Ministério das Comunicações merecem ser destacados pelo menos dois fatos, primeiro o

Manifesto assinado pela Sociedade Brasileira de Progresso da Ciência (SBPC) contra a fusão dos ministérios. Já o segundo fato foi a audiência pública realizada na Câmara dos Deputados, no dia 15 de junho de 2016, também com o objetivo de discutir a fusão das Pastas.

O Manifesto contra a fusão do MCTI com o Ministério das Comunicações foi assinado por 14[66] entidades e enviado ao Presidente Michel Temer no dia 11 de maio de 2016. O manifesto concentra-se na importância da pesquisa e desenvolvimento científico e tecnológico como propulsor de crescimento do país, a exemplo dos países de primeiro mundo. Além disso, procura enfatizar a diferença de procedimentos, objetivos e missões entre os ministérios objetos da fusão.

A audiência pública foi marcada pelo discurso veemente de Helena Nader, presidente da SBPC, posicionando-se contra a fusão e afirmando "*não vamos desistir do MCTI*". Na visão de Nader, os 30 anos de existência do Ministério trouxeram inestimável impacto para a ciência, tecnologia e inovação no País. Ela argumenta que "Concordo que é preciso enxugar a máquina, mas há outras maneiras de fazer isso que não seja sobre o Ministério da Ciência, Tecnologia e Inovação. Em todos os lugares do mundo, a ciência e a tecnologia são consideradas o motor da economia. Quando o Governo Federal faz isso, os Estados podem começar a fazer também" (SBPC, 2016).

No que diz respeito ao orçamento para a C,T&I, o Ministro Kassab, durante 2016 também enfrentou a preocupação das entidades com a falta de recursos para a Rede Nacional de Pesquisa (RNP) o que levaria a descontinuidade dos serviços de internet para 20 universidades, para a Empresa Brasileira de Pesquisa e Inovação (EMBRAPII) e para o Centro de Gestão e Estudos Estratégicos (CGEE), esse último, notadamente enfraquecido durante a gestão do Ministro Celso Pansera.

Apesar do apelo das entidades de classe, o Presidente Michel Temer deu continuidade à fusão desses ministérios, conforme inicialmente anunciado, e o ano de 2016 encerrou com o Ministro Kassab permanecendo à frente da Pasta.

---

[66] Academia Brasileira de Ciências (ABC), Academia de Ciências do Estado de São Paulo (ACIESP), Academia Nacional de Medicina (ANM), Associação Brasileira de Universidade Estaduais e Municipais (ABRUEM), Associação Nacional de Entidades Promotoras de Empreendimentos Inovadores (ANPROTEC), Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (ANDIFES), Associação Nacional de Pesquisa e Desenvolvimento das Empresas Inovadoras (ANPEI), Conselho Nacional das Fundações de Apoio às Instituições de Ensino Superior e de Pesquisa Científica e Tecnológica (CONFIES), Conselho de Reitores das Universidades Brasileiras (CRUB), Conselho Nacional das Fundações Estaduais de Amparo à Pesquisa (CONFAP), Conselho Nacional de Secretários Estaduais para Assuntos de C, T & I (CONSECTI), Fórum de Pró-Reitores de Pesquisa e Pós-Graduação ( FOPROP), Fórum Nacional de Gestores de Inovação e Transferência de Tecnologia (FORTEC), Sociedade Brasileira para o Progresso da Ciência (SBPC).

## 2.9 CONSIDERAÇÕES SOBRE CIÊNCIA E TECNOLOGIA A PARTIR DA DÉCADA DE 1980

Em síntese, Correa (2003) classifica em três etapas a política de desenvolvimento científico e tecnológico no Brasil. A primeira etapa compreende o período de 1946 a 1964 – marcado pela institucionalização da ciência e tecnologia. A segunda etapa tem início em 1964 e vai até 1985 – foi idealizada pelo Governo Militar, período em que a ciência e a tecnologia procuram assegurar a soberania nacional. Essa visão, que relaciona C&T com a soberania nacional é defendida por Roberto Amaral, ex Ministro de Ciência e Tecnologia entre os anos de 2003 e 2004. Amaral (2011) defende que "o desenvolvimento em ciência e tecnologia está no centro do desenvolvimento econômico, social e militar, e condiciona os conceitos de soberania e defesa". Na sua opinião, soberania nacional e dependência científico tecnológica são incompatíveis.

A terceira etapa inicia em 1985 e seu término se dá em 1989, o período é descrito como uma fase de transição e instabilidade. Por fim, o quarto período tem início em 1990 e vai até 2002 (período compreendido pela obra), e na percepção da autora (CORREA, 2003) ele é marcado por um discurso liberal, em que o Estado assume um papel de gestor da política de C&T, em substituição ao papel, anterior, de interventor. Nas palavras de Corrêa (2003), "até o final dos anos 1970, o Brasil teve (mal ou bem) uma política para a ciência, durante a década de 1980 houve um certo vazio em termos dessa política, os anos 1990 se caracterizam por uma política para a inovação".

Salerno e Kubota (2008) corroboram a percepção de Correa (2003) sobre políticas de inovação. Os autores argumentam que, a partir dos anos 1980 e 1990, o modelo linear de ciência foi sendo substituído pela abordagem sistêmica de inovação, fato observado com maior ênfase no Brasil a partir do lançamento, em 31 de março de 2004, da Política Industrial Tecnológica e de Comércio Exterior.

A ênfase na política de fomento à inovação nas empresas brasileiras é explícita no "Plano de Ação de Ciência, Tecnologia e Inovação (PACTI)", publicado pelo MCT em 2007. O PACTI previa um orçamento de 41,2 bilhões em investimentos até 2010. Ressalta-se que suas prioridades estratégias, em consonância com a Política Nacional de CT&I foram: a) expansão e consolidação do Sistema Nacional de Ciência, Tecnologia e Inovação; b) promoção da inovação tecnológica nas empresas; c) pesquisa, desenvolvimento e inovação em áreas estratégicas; e d) ciência, tecnologia e inovação para o desenvolvimento social.

Durante a *4º Conferência Nacional de Ciência e Tecnologia*, realizada em Brasília, no período de 26 a 28 de maio de 2010, foi colocado que nenhum país que tenha como objetivo promover o desenvolvimento para se tornar uma potência de inovação tecnológica pode deixar de investir em tecnologias da informação e comunicação. Novamente, percebe-se a preocupação do governo com a inovação nas empresas brasileiras.

Na era da sociedade da informação, o conhecimento oriundo da pesquisa científica é entendido como uma oportunidade para a inovação tecnológica. Em tempos de economia globalizada, as atividades de P&D possuem estreita relação com o setor produtivo. De forma simplista, é possível afirmar que desse modelo econômico surgem os sistemas nacionais de ciência, tecnologia e inovação, constituído por diferentes atores em prol de alavancar o país na área da produção de inovação tecnológica e, assim, garantir uma posição entre os países líderes da economia mundial.

Desde a estabilização do real, observa-se uma preocupação do governo federal com o estímulo à inovação e à competitividade da empresa brasileira, o que favorece mais uma política de desenvolvimento tecnológico, com ênfase na melhoria do parque industrial, do que uma política para a ciência propriamente dita. Barros (2001, p. 85) corrobora esse entendimento ao argumentar que as declarações governamentais recentes[67] "embora tentem paralelamente salvaguardar a área científica, indicam uma preocupação maior com a questão tecnológica".

Salerno e Kubota (2008, p. 17) argumentam que para a Organização para Cooperação Econômica e Desenvolvimento (OECD), "as políticas de inovação constituem um amálgama das políticas de ciência, de tecnologia e industrial". Elas envolvem "a relação entre a ciência e a sua produção, a tecnologia e sua geração, assim como a inovação por parte das empresas."

A respeito do assunto, Schwartzman (2001) considera que o modelo linear de pesquisa foi perdendo espaço. Consequentemente, o apoio para a pesquisa básica foi perdendo terreno, quando esta não está associada a resultados e produtos previamente identificáveis. Rezende (2010) aborda essa questão ao criticar os fundos setoriais.

A questão mostra-se complexa, pois para gerar inovação radical faz-se necessário um considerável investimento em pesquisa e desenvolvimento, que por sua vez implica estímulo à pós-graduação e à produção de ciência básica. Pois, embora ciência e tecnologia seja um campo distinto da educação, eles, sem nenhuma dúvida, constituem campos complementares. Nesse

[67] O comentário do autor refere-se ao período em que Ronaldo Sardenberg foi Ministro de Estado da Ciência e Tecnologia, durante o segundo mandato do presidente Fernando Henrique Cardoso (1999-2002).

cenário, é oportuno ressaltar que o estímulo à pesquisa perpassa pela questão do acesso à informação científica e tecnológica.

Nesse sentido, ao se apoiar o fortalecimento da pesquisa e do desenvolvimento, fala-se, ainda que indiretamente, no apoio à gestão dos dados científicos, dentre eles os dados *online* inerentes a projetos de *e-science.* O fato é que a gestão dos dados científicos precisa ser norteada por uma política de tratamento da informação que oriente as questões de armazenamento, de curadoria, recuperação e colaboração de dados de forma a contribuir para o avanço da pesquisa no Brasil.

A respeito do assunto, Cunha (2001, p. vii) argumenta que:

> [...] entre os fatores que distinguem os países em desenvolvimento (agora emergentes) está o acesso à informação. Realmente os países mais desenvolvidos possuem acesso mais rápido à informação científica e tecnológica, ampliando, cada vez mais o que Jean-Jacques Servan Schreiber chamou de 'fosso tecnológico'.

Corrobora esse aspecto, o fato de o então Ministro da Ciência, Tecnologia e Inovação, Aluísio Mercadante, ao apresentar, em 2012, a Estratégia Nacional de Ciência, Tecnologia e Inovação (ENCTI) comentar que "o principal desafio que o Brasil terá de enfrentar se quiser se transformar em um país efetivamente desenvolvido, com uma economia eficiente e competitiva, é preparar-se para a sociedade do conhecimento" (BRASIL, 2012, p. 9). A ENCTI entende ciência, tecnologia e inovação como eixo estruturante do desenvolvimento do Brasil. Ao se falar em uma sociedade competitiva, retomamos o conceito de sociedade do conhecimento, onde a gestão de dados científicos se torna fundamental para o desenvolvimento do país.

Compreender o processo de desenvolvimento da política científica e tecnológica do Brasil significa, antes de tudo, conhecer os percalços pelos quais os pesquisadores e cientistas passaram ao induzirem a criação dos programas de pós-graduação, dos centros de pesquisa, das associações científicas e instituições de fomento à C&T no Brasil, além de permitir conhecer os interesses políticos que permearam o caminho da ciência e do desenvolvimento tecnológico no Brasil. Essa breve análise histórica corrobora a proposta de *campo científico* de Bourdieu (2003; 2004), bem como reforça o conceito de autonomia relativa proposto pelo autor, pois o que Bourdieu propôs foi um meio termo entre autonomia e ausência de autonomia da ciência. Em suas palavras: "escapar à alternativa da 'ciência pura', totalmente livre de qualquer necessidade social, e da 'ciência escrava', sujeita a todas as demandas político-econômicas" (BOURDIEU, 2004, p. 21).

## 2.10 POLÍTICA DE INFORMAÇÃO NO EXTERIOR E NO BRASIL

Na medida em que a sociedade sofreu a transição do modo de produção capitalista para a chamada sociedade da informação, emerge o campo da Ciência da Informação, bem como as preocupações inerentes ao acesso e uso da informação, o que culmina posteriormente, com a elaboração de políticas para a área. Na percepção de Bramam (2011), a política de informação surge como um campo distinto durante as últimas décadas do Século XX.

O país que se destaca no assunto é os Estados Unidos, em função da própria maneira com que se beneficiaram da informação durante a Segunda Guerra. É no seio dessa modificação ocorrida na sociedade e no cenário pós Segunda Guerra Mundial que nasce a própria Ciência da Informação.

Borko (1968), em um contexto histórico de pós Segunda Guerra e início da Guerra Fria, apresenta reflexões sobre o que viria a ser a Ciência da Informação e comenta sobre o seu caráter interdisciplinar. Já Wersig e Neveling (1975) contextualizam sobre o advento da Ciência da Informação como uma disciplina que nasce para responder problemas práticos, oriundos da documentação e amplamente discutidos no artigo de Bush (1945).

Pode-se afirmar que a partir das preocupações inerentes ao acesso à informação, trazidas à tona por Bush (1945) e discutidas, em momento posterior, por Borko (1968), Wersig e Neveling (1975), Belkin (1978), dentre outros, culminaram com a necessidade de refletir-se sobre os aspectos epistemológicos, técnicos e políticos da nova área. Dentre essas reflexões, surgem as questões de acesso à informação científica e tecnológica, segurança nacional, liberdade de expressão, acesso à informação governamental, transparência nas ações do governo que trouxeram à tona a necessidade de se fomentar políticas de informação.

Apesar do cenário exposto acima, Rosenberg (1982b) defende que a ideia de uma política nacional para governar a criação, distribuição e uso da informação não é nova. O autor fundamenta sua colocação no fato de que a primeira e a quarta emenda da Constituição dos Estados Unidos[68] têm mais de duzentos anos.

Rowlands (1998), por sua vez, argumenta que o surgimento de uma abordagem sistemática para a formulação de políticas de informação nos Estados Unidos tem início na década de 1960, devido à influência do *Weinberg Report*, publicado em 1963, que tratava sobre transferência de informação científica e técnica nos EUA. De acordo com a autora:

> [...] durante a década de 1960 e início de 1970, as questões e preocupações em torno da informação científica e técnica assumiram um foco geopolítico

[68] A Constituição Americana foi promulgada em 17 de setembro de 1787.

> preciso contra o pano de fundo da Guerra Fria e da corrida para a dominação do espaço sideral. O surgimento posterior de processamento de dados em grande escala e os volumes crescentes de informações que foram mantidas em linguagem de máquina no início da década de 1970 obrigaram a uma resposta política pública acordada, que se estendeu muito além das fronteiras da comunidade científica e técnica. Pela primeira vez, os problemas em áreas tão diversas como a privacidade, a liberdade de expressão, o sigilo, o acesso às informações do governo e a segurança nacional estavam sendo levados ao conhecimento dos decisores políticos, em uma única construção que integra - o poder do computador (e, por implicitamente, informações) para mudar radicalmente a sociedade.

Na percepção de Rosenberg (1982b), desde a publicação do *Weinberg Report*, talvez a resposta mais significativa, do poder executivo do governo americano para política de informação tenha sido o *Relatório Rockefeller*, publicado em 1975.

Rowlands (1998) considera que questões inerentes à informação têm tido espaço constante na política pública dos EUA há pelo menos 30 anos, e, na percepção da autora, o final da década de 1990 traz à tona a necessidade de se fomentar um conjunto de medidas de política de informação coerentes para melhorar a qualidade da prestação dos serviços públicos, para garantir o acesso justo e continuado à informações do governo, bem como para aumentar a criatividade nacional, a produtividade e a geração de riqueza.

Para Rosenberg (1982a; 1982b), o termo "política de informação" também está relacionado ao corpo de leis e aos regulamentos que regem a indústria de telecomunicações e jornalismo, incluindo questões como concessões de rádio e TV, exibição de filmes estrangeiros e tecnologia empregada na disseminação da informação. Já Fung (1986), também relaciona o termo às questões de privacidade e liberdade de informação, bem como a leis e políticas que afetam as editoras, durante a impressão de uma publicação, e as bibliotecas.

Rowlands (1998) adverte que uma definição sucinta para política de informação tende a ser problemática. Para a autora:

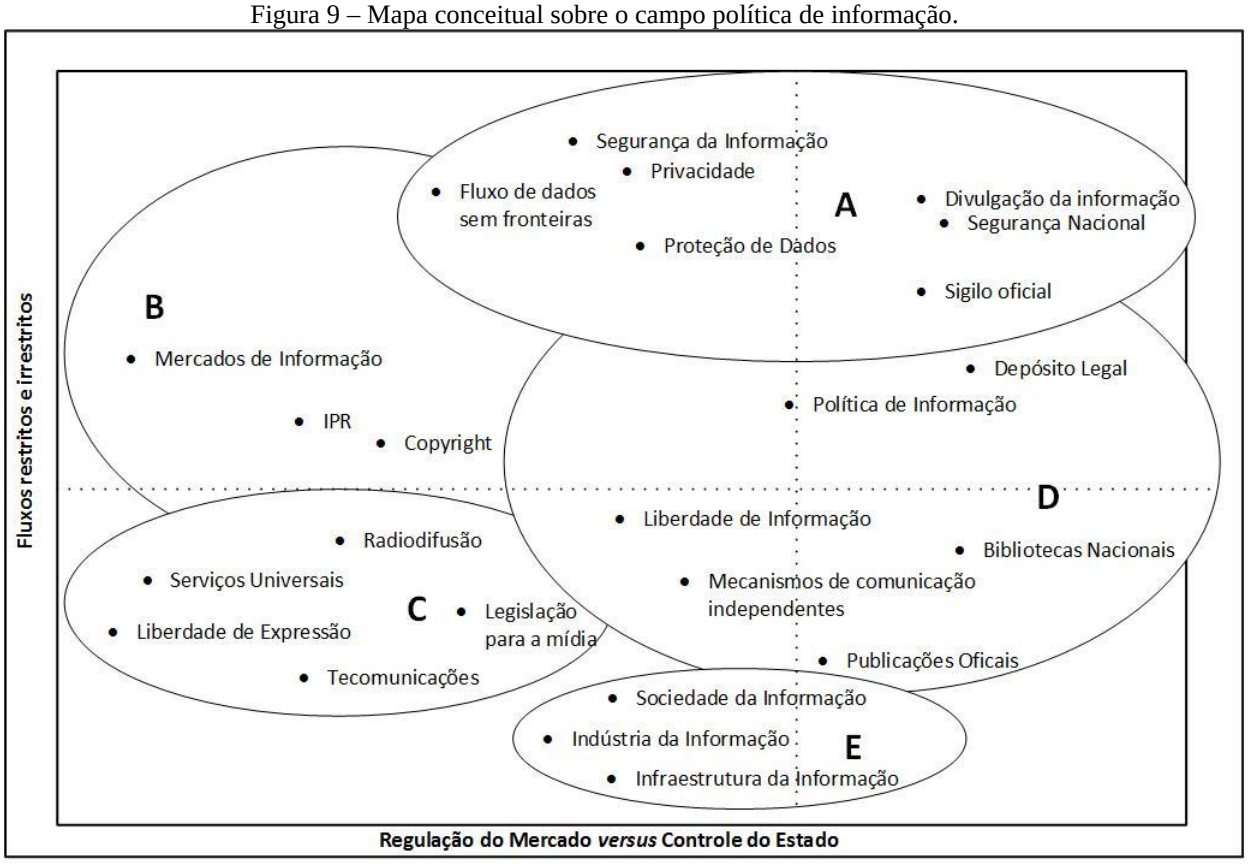
> [...] historicamente, as políticas de informação tendem a se concentrar sobre questões específicas e áreas problemáticas, tais como pesquisa e desenvolvimento, o desenvolvimento do mercado da informação, a liberdade de acesso à informação governamental e os aspectos legais, como privacidade, direitos autorais e direitos de propriedade intelectual (ROWLANDS, 1998, p. 232).

Nesse sentido, Rowlands (1998) considera que alguns autores se mostram relutantes em definir a natureza e o alcance da política de informação por medo de serem rotulados como reducionistas. A autora demonstra apreço pela definição de política de informação proposta por Hernon e Relyea (1968 *apud* ROWLANDS, 1998, p. 232) que diz:

> [é] um campo que engloba tanto a ciência da informação como as políticas públicas, [que] trata informações tanto como uma mercadoria - aderente à teoria econômica dos direitos de propriedade – e como um recurso a ser coletado, protegido, compartilhado, manipulado e gerenciado. Contudo, a literatura muitas vezes se refere à política de informação no singular, mas não existe uma política abrangente simples.

A partir do exposto, Rowlands (1998) alerta que o verdadeiro problema para desenvolver uma política nacional de informação é entender o que precisamente ela pretende atingir. A autora apresenta um mapa com conceitos de política de informação e os relaciona uns com os outros em um espaço bidimensional, conforme demonstra a Figura 9.

Figura 9 – Mapa conceitual sobre o campo política de informação.




Fonte: Rowlands (1998, p. 232, tradução nossa).

A partir do mapa conceitual, Rowlands (1998) propõe cinco agrupamentos de conceitos para representar os subdomínios, dentro do campo mais amplo da política de informação, conforme pode ser observado no Quadro 10.

Quadro 10 – Subdomínios da política de informação.

| CLUSTER | SUBDOMÍNIO | INTERPRETAÇÃO |
|---|---|---|
| **A** | Protecionismo da informação | Regulamentos e mecanismos que controlam o acesso à informação e divulgação na esfera pública (ex.: segredo oficial) e em mercados de informação (ex.: proteção de dados). |
| **B** | Mercado da informação | Leis e regulamentos que protegem o investimento na criação de conteúdo informacional (ex.: direitos autorais) e permitem trocas no mercado. |
| **C** | Radiofusão e telecomunicações | Políticas públicas que regulam os meios de comunicação de massa, equilibrando interesses comerciais e do cidadão (ex.: acesso universal). |
| **D** | Acesso público à informação oficial | Políticas e regulamentos que moldam o acesso à informação de cidadãos, arquivada pelo governo (ex.: liberdade de informação). |
| **E** | Sociedade da informação e infraestrutura | Políticas públicas que promovem o investimento (ou encorajam o setor privado a investir) na infraestrutura de informação. |

Fonte: Rowlands (1998, p. 233).

No Brasil, a política de informação é discutida, no âmbito da informação científica e tecnológica por Aguiar (1980), Garcia (1980), Lemos (1986), Gomes (1988), Costa (1991) e Cunha (2005); da informação arquivística, por Jardim (1999), da informação sobre telecomunicações por Marques e Pinheiro (2011); da informação para a leitura por Silva, Bernardino e Nogueira (2012), e no âmbito de uma política de informação geral, por Amaral (1991), Silva (1991), Aun (1999), Unger e Freire (2008). Além destes, Gonzáles de Gómez (1987; 2002) traz importantes contribuições teóricas para o cenário brasileiro.

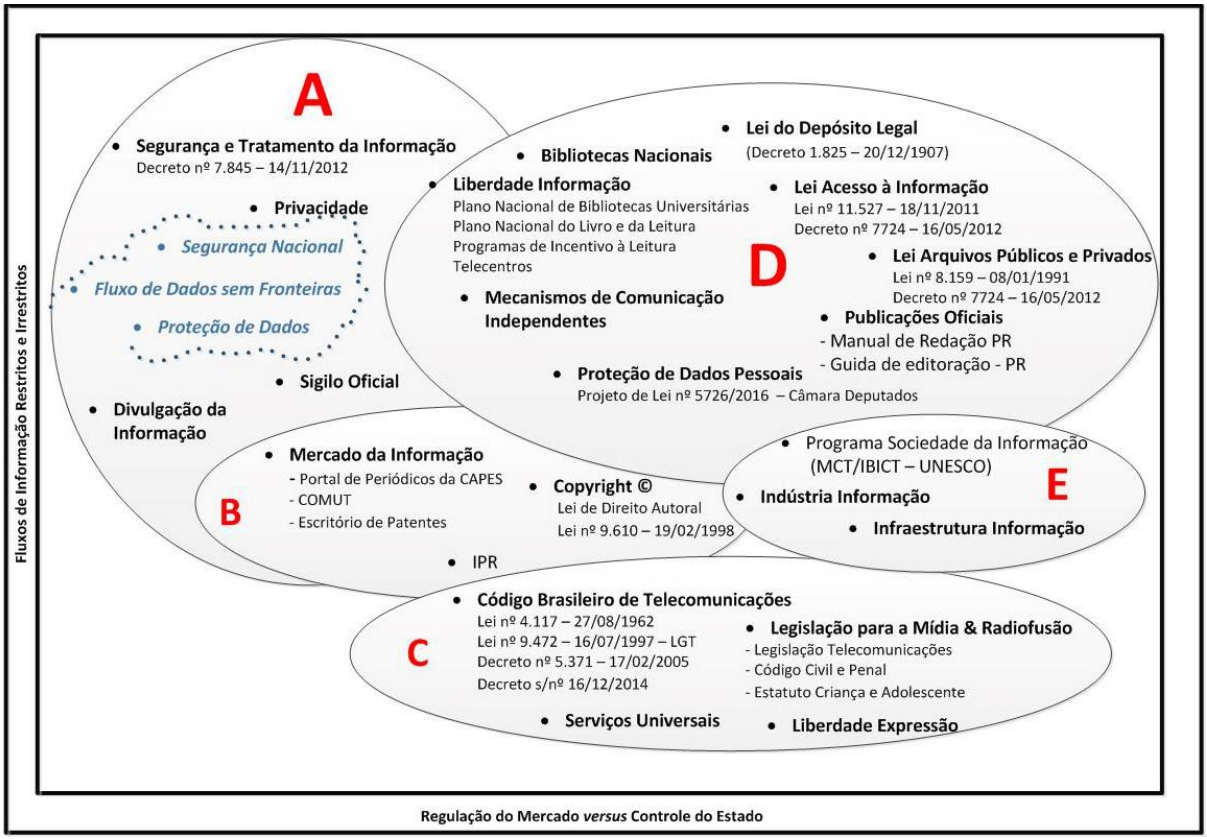
Uma das primeiras iniciativas brasileiras, quanto à política de informação, foi a publicação do Decreto 1.825 em 20 dezembro de 1907 que dispõem sobre a remessa de obras impressas à Biblioteca Nacional. Essa iniciativa ficou conhecida como Lei do Depósito Legal e tinha como objetivo registrar e preservar a memória intelectual do país por meio do recolhimento e da preservação dos livros publicados no Brasil, ou, publicados por autores brasileiros no exterior. O Decreto vigorou até o dia 14 de dezembro de 2004, data em que foi publicada a Lei 10.994 que dispõe sobre o depósito legal de publicações, na Biblioteca Nacional e dá outras providências. Em 14 de janeiro de 2010 é publicada a Lei 12.192, que dispõe sobre o depósito legal de obras musicais na Biblioteca Nacional.

A Lei do Depósito Legal coaduna com a percepção de Rosenberg (1982a; 1982b), Fung (1986), Gomes (1988) e Rowlands (1998) no que diz respeito à preservação da memória intelectual de um país, não se restringindo apenas aos livros, mas incluindo também a indústria

cultural. Nesse sentido, cabe a reflexão de que a iniciativa brasileira além de se preocupar com a preservação da memória, procurou estabelecer mecanismos para que o país exercesse o controle bibliográfico nacional, fato que viabilizou a produção de bibliografias nacionais na década de 1980.

Ao que parece, o Brasil não teve outra iniciativa que se aproximasse de uma política de informação até a década de 1950, marcada pela busca do desenvolvimento científico e tecnológico do país. Reproduzindo o mapa conceitual de Rowlands (1998), no contexto brasileiro, conforme ilustra a Figura 10, se obtém uma macro visão do marco legal sobre a política de informação, bem como áreas que carecem de regulamentação, a exemplo da questão sobre o fluxo de dados e a própria proteção de dados.

Figura 10 – Mapa conceitual sobre o campo política de informação no Brasil.




Fonte: A autora a partir da adaptação ao contexto brasileiro a partir de Rowlands (1998, p. 232).

A literatura revela que a política de informação no Brasil mostra-se imbricada com a política de desenvolvimento científico e tecnológico, ocorrida na década de 1950, em função da necessidade do acesso à informação científica e tecnológica. O fato que marca a relação entre as duas políticas é uma das atribuições do CNPq, quando de sua criação, qual seja, "manter relação com instituições nacionais e estrangeiras para intercâmbio de documentação técnico-

científica" (IBICT, 2013). Essa atribuição culminou com a publicação do Decreto n° 35.124, em 27 de fevereiro de 1954, que criou o Instituto Brasileiro de Bibliografia e Documentação (IBBD), órgão vinculado ao CNPq.

Sobre o assunto, Gomes (1988, p. 105) já defendia a ideia de que "as orientações fundamentais relativas à definição e à implantação de uma política nacional de informação científica e tecnológica já estão implícitas no planejamento econômico e social em geral e, particularmente, na política de desenvolvimento científico e tecnológico".

A criação do IBBD foi oportuna sob duas perspectivas, a primeira refere-se à necessidade de acesso à informação científica e tecnológica em prol de alavancar-se o desenvolvimento do país. A segunda, por sua vez, refere-se ao período de apoio da UNESCO, por meio do Comitê Internacional de Bibliografia, para a criação de centros nacionais de informação.

De acordo com Silva (1994), a UNESCO teve papel de destaque na criação dos sistemas e serviços nacionais de informação na América Latina. Para Silva (1994), a atuação da UNESCO divide-se em três fases: a primeira ocorreu na década de 1950 e compreendeu a criação de centros nacionais de informação no México, Brasil e Uruguai efetivamente. Nos anos 1970, teve início a segunda fase, quando a UNESCO envidou esforços para criar os sistemas nacionais de informação. Na terceira fase, nos anos 1980, essa instituição procurou auxiliar os países a formularem suas respectivas políticas nacionais de informação, sem alcançar êxito na Argentina, no Brasil e no México.

Dentre os serviços de informação em C&T oferecidos à época pelo IBBD, destacavam-se: as pesquisas bibliográficas, as bibliografias especializadas, a criação do Catálogo Coletivo Nacional (CCN) em 1954, a criação da "Revista Ciência da Informação" em 1972, a criação do Programa de Comutação Bibliográfica (COMUT) em 1980, dentre outros. Dentre as atividades idealizadas pela Presidente do IBBD – Lydia de Queiroz Sambaquy, destaca-se o projeto de criação da Biblioteca Nacional de Ciência e Tecnologia (ANDRADE; OLIVEIRA, 2005).

Em 1976 o IBBD enfrentou transformações que culminaram com a mudança do seu nome para Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT[69]), mantendo-se vinculado ao Conselho Nacional de Desenvolvimento Cientifico e Tecnológico. Nesse ano, o Instituto consolidava-se, então, como órgão que coordenaria, no Brasil, as atividades de informação em C&T.

[69] A Resolução nº 20, de 25 de março de 1976 cria o IBICT.

Sobre o IBICT é pertinente comentar que Lemos (1986), Tarapanoff (1992) e Cunha (2005) consideram o Instituto o órgão responsável pela política brasileira de informação científica e tecnológica.

Cunha (2005) apresenta uma breve história da trajetória do IBICT acompanhada do contexto da política de desenvolvimento científico e tecnológico. O autor acredita que as mazelas enfrentadas pelo setor científico e tecnológico na década de 1970 e 1980 foram refletidas nas atividades do Instituto, iniciando assim:

> [...] um período de difícil transição no qual se destaca a rotatividade de seus dirigentes. Tal fato pode ter ocasionado uma possível descontinuidade administrativa, com reflexos negativos nos diversos níveis da cadeia hierárquica, trazendo em seu bojo alguns transtornos, haja vista o novo perfil que cada dirigente trazia ao assumir, resultando em interrupções total ou parcial de projetos, ou na geração de novas ações sem uma adequada análise (CUNHA, 2005, p. 7).

A década de 1990 trouxe a necessidade de o IBICT refletir sobre a Internet, os periódicos digitais e a própria biblioteca digital. Em 1999 foi criado o Programa Sociedade da Informação, que já chegou trazendo rumores sobre a absorção das atividades do Instituto pelo Programa (CUNHA, 2005).

O IBICT é considerado hoje referência em projetos voltados ao movimento do acesso livre ao conhecimento. Exemplo desse compromisso é a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), lançada em 2002, que utiliza a tecnologia de arquivos abertos e integra sistemas de informação de teses e dissertações de instituições de ensino e pesquisa brasileiros. Além da BDTD, o Instituto apoia, desde 2009, em conjunto com a FINEP, a criação de repositórios institucionais abertos em universidades públicas e centros de pesquisas financiados com recursos públicos. Outra iniciativa de destaque é a formação da Rede de Serviços de Preservação Digital – Cariniana, responsável pela preservação dos periódicos eletrônicos na plataforma OJS/SEER no Brasil.

A respeito das iniciativas do IBICT de acesso aberto à informação científica, merece ser ressaltado que o Instituto tem concentrado seus esforços nas teses, dissertações e nos periódicos. Entretanto, até o momento, não há iniciativas por parte IBICT relacionadas ao armazenamento, tratamento, à curadoria, preservação e difusão da informação para os dados coletados no âmbito da *e-science*, que antecedem o processo de publicação científica, seja ele uma tese, uma dissertação ou publicação em periódico.

Retomando o aspecto cronológico do texto, em 1973 é criado o Sistema Nacional de Informação Científica e Tecnológica (SNICT), que tinha como objetivo planejar e coordenar,

em âmbito nacional, os trabalhos de informação científica e tecnológica. Sobre o SNICT, observa-se que IBBD ia atuar como um órgão de apoio, assim como a Biblioteca Nacional.

De acordo com Tarapanoff (1992, p. 151), "o tempo perdido entre a expressão da ideia, a criação do sistema [SNICT] e o esboço do decreto que deveria regulá-lo foi de três anos". A autora ressalta que "o rascunho do decreto foi entregue à Presidência da República, mas nunca foi submetido à sua aprovação". Nesse sentido, o SNICT "nunca passou de uma ideia".

Durante o ano de 1999 foi formado, em Brasília, um grupo de discussão sobre o contorno e as diretrizes de um programa que pretendia introduzir o Brasil na chamada sociedade da informação. A partir das discussões desse grupo o Ministério da Ciência e Tecnologia criou o Programa Sociedade da Informação. Esse programa teve três estágios. O primeiro compreendeu os estudos preliminares que conduziram ao lançamento formal do Programa em dezembro de 1999. O segundo consistiu na elaboração do Livro Verde que continha a proposta do Programa Sociedade da Informação. Em seguida, a terceira fase é representada pelo detalhamento do Programa feito no Livro Branco.

Infere-se que o Programa Sociedade da Informação foi o embrião que deu origem ao Programa de Governo Eletrônico (e-Gov), que dentre seus objetivos procura aproximar o governo da sociedade e promover a transparência das ações do governo.

Para Amaral (1991), até o início da década e 1990, o governo e a sociedade ainda não compreendiam a importância para a criação de uma política nacional de informação no Brasil. A autora considera que tanto o governo, como a classe dominante do país não tinham interesse em implantar tal política.

Na percepção de Silva (1991), as políticas de informação no Brasil têm sido propostas, mas não são prioridade para o governo, e não apresentam a articulação necessária com o contexto cultural e educacional. Nesse sentido, a autora converge sua opinião com Herrera (1995) no que diz respeito à existência de uma política implícita, difusa nas demais políticas públicas governamentais.

Costa (1991), por sua vez, considera que o assunto tem sido debatido há anos, assim como iniciativas têm sido envidadas. Porém, o autor considera que os resultados não têm sido satisfatórios e dentre os motivos para tanto, está o fato de que os planos, ou programas de ação nacional, muitas vezes não saem do papel.

Silva (1991, p. 11) critica o acesso à informação científica e tecnológica no Brasil. Para a autora, "o ciclo da transferência de informação tecnológica é quebrado", o que ocorre é a importação de *kown how* e manuais de serviço pelo país.

Para Silva (1991), uma política de informação ideal, mas não utópica, deve considerar em seu escopo: a) a multifacetada realidade brasileira (o computador e a enxada convivendo juntos); b) o direito de acesso à informação pelos menos favorecidos; c) a integração da sociedade aos avanços científicos e tecnológicos de forma participativa e; d) a participação dos diversos seguimentos da sociedade brasileira.

A percepção de Jardim (ENANCIB, p. 6) é de que "uma política de informação é mais que a soma de um determinado número de programas de trabalho, sistemas e serviços". Para o autor, "é necessário que se defina o universo geográfico, administrativo, econômico, temático, social e informacional a ser contemplado pela política de informação".

Interessante observar que Gomes (1988), Amaral (1991) e Jardim (2008) consideram que uma política de informação deve ser um conjunto explícito de diretrizes. A crítica de Jardim (2008, p. 6, grifo nosso) é severa ao considerar que "um estado democrático é, por princípio, incompatível com políticas públicas de saúde, educação, habitação ou informação, que não sejam explícitas".

Sobre o assunto, cabe retomar-se o conceito proposto por Herrera (1995) sobre política implícita e política explícita no âmbito das políticas científicas. A primeira é a política oficial, expressa em leis, declarações governamentais, regulamentos e instituições de governo responsáveis pelo planejamento da ciência. Já a política implícita é a que determina o papel da ciência na sociedade, é muito mais difícil de ser identificada porque não tem estrutura formal; essencialmente expressa as demandas científicas e tecnológicas do "projeto nacional" de cada país, é a que está verdadeiramente em ação. O autor assinala que essas políticas (explícita e implícita) não são necessariamente contraditórias e divergentes. Porém, argumenta que quando o país está em crise, elas normalmente apresentam contradições.

Para Gomes (1988, p. 105) a política de informação deve considerar o que "seria desejável e realizável para um país em matéria de produção, transferência e acesso à informação, levando-se em conta os recursos informacionais e de infraestrutura existentes, os recursos desejados, as necessidades dos usuários e, de modo geral, da própria sociedade em sua totalidade".

As reflexões propostas por Rosenberg (1982a; 1982b), Gomes (1988), Rowlands (1998) e Braman (2011) têm em comum o fato de que a política de informação é complexa, pois rege a forma com que a informação vai interagir com a sociedade.

Na percepção de Rosenberg (1982b, p. 37), "tanto os países industrializados como os países em desenvolvimento formulam suas políticas de informação com o objetivo de proteger

interesses que consideram vitais". Para o autor, o resultado prático da ação "é quase sempre o aparecimento de barreiras ao livre fluxo da informação para dentro e para fora do país".

A argumentação exposta acima por Rosenberg (1982a) é perfeitamente observada no acesso à informação em países como, por exemplo, a China, onde o governo tende a censurar *sites*, mecanismos de busca e redes sociais estrangeiros.

Rosenberg e Cunha (1983) acreditam que o acesso à informação científica e tecnológica pode ser considerado essencial ao desenvolvimento econômico de um país. Sobre o assunto, os autores refletem que ainda não está claro se a falta de informações é resultado da carência de desenvolvimento econômico, ou o contrário – se é causa dessa carência.

Na percepção de Marques e Pinheiro (2011, p. 69) "uma análise da relevância da informação na sociedade contemporânea, especificamente, políticas de informação nacionais atuais, não seria possível sem a apreensão das novas dinâmicas que surgem com a difusão das TIC".

Rosenberg (1982a) já havia alertado, na década de 1980, que os avanços da tecnologia de informação, bem como, as novas formas de comunicação precisavam lançar um novo olhar sobre as regras, regulações, leis e políticas que permeiam o mundo da informação. Que dirá agora em épocas de *big data* a dimensão das modificações necessárias.

De volta à problemática desta pesquisa, é imperioso comentar que o movimento de *big data* impõe a necessidade de haver um repositório de dados tratados, com possibilidade de recuperação e difusão da informação para a utilização em novas pesquisas. Portanto, faz-se necessário que o Brasil implemente uma política de gestão de dados científicos, como parte de uma política nacional de informação. Para ampliar o entendimento sobre o tema, é preciso discutir sobre políticas de acesso à informação de forma explícita.

Nesse cenário, é importante ressaltar a necessidade de clareza da política informacional sobre a reutilização dos dados coletados no *big data*. Qual o limite para essa reutilização? Quais as regras para a reutilização desses dados? Quando se tratar de dados sobre a vida, quais as questões que emergem quanto à reutilização? O fato é que as atuais políticas de informação mostram-se obsoletas no que diz respeito aos meios legais de reutilização de dados e proteção à privacidade. Por outro lado, o conceito de política implícita, proposto por Herrera (1995) também merece profunda atenção em questões tão delicadas, tais como os dados sobre a ciência da vida. Nesse aspecto, a política de informação deve ser explícita.

A partir das considerações de Rosenberg (1982a; 1982b) e Fung (1986) entende-se que o Brasil possui um marco legal quanto à política de informação. Fazem parte da política explícita, portanto do marco legal, a Lei de Depósito Legal, a Lei de Acesso à Informação e as

legislações inerentes ao Conselho Nacional de Arquivos e todo o marco legal apresentado na Figura 10, desenvolvido a partir do mapa conceitual sobre política de informação de Rowlands (1998).

Na visão de Gray (2007), os formuladores de políticas públicas precisam estimular tanto a criação de *softwares* de análise de dados, como o apoio a essas ferramentas. Entende-se aqui que o apoio pressupõe as atividades de sustentação de um sistema já em operação, o que viabiliza o aprimoramento contínuo desses *softwares*.

Borgman (2013) apresentou reflexões sobre questões inerentes ao compartilhamento de dados, como, por exemplo: quais dados devem ser compartilhados, quando os dados podem ser compartilhados, e de que forma os dados podem ser compartilhados? Do ponto de vista do tratamento técnico da informação, Borgman (2013) tece as seguintes considerações – como atribuir crédito para dados (atribuição) e como fazer referência a dados (citação) de forma que os outros possam identificar, descobrir e recuperá-los?

As questões inerentes a gestão de dados científicos no Brasil são muitas e complexas. É necessário que o profissional da informação se aprofunde no tema e proponha diretrizes, em conjunto com pesquisadores de cada área do conhecimento, para a construção de um modelo teórico que atenda a um conjunto mínimo de diretrizes para a descrição de dados científicos, tratamento técnico, temporalidade de armazenamento (incluindo formas de preservação e regras para descarte), regras de reutilização e citação do dado original.

# 3 A CONSTRUÇÃO DA METODOLOGIA

Esta pesquisa iniciou em 2013, à época, apenas três autores brasileiros haviam publicado sobre o tema *e-science* no âmbito da Ciência da Informação. A pesquisa sobre o tema em bases de dados internacionais, apesar de trazer mais resultados, por vezes também não se mostrou promissora em função do desconhecimento dos termos indexadores que estavam sendo utilizados nas respectivas bases de dados.

Assim, após busca exaustiva nas bases de dados internacionais: ERIC, LISA, LISTA e Web of Science, foi verificado que ainda não havia trabalho na literatura sobre Ciência da Informação atrelado ao fenômeno da *e-science* que identificasse a sua comunidade, os autores que mais publicam, bem como os países que produzem essa literatura. Ou seja, não havia um mapeamento a respeito do tema que facilitasse a identificação dos canais de comunicação utilizados pela comunidade.

A identificação da comunidade científica que está trabalhando em *e-science* mostrou-se fundamental para nortear a revisão de literatura, bem como para auxiliar na identificação de um país para realizar o doutorado sanduíche e, assim, permitir que a pesquisadora se aproximasse das atividades que envolvem a gestão dos dados científicos operacionalmente e não apenas pela leitura.

A partir do exposto, a primeira etapa da pesquisa foi a realização de um estudo bibliométrico em todos os registros bibliográficos das bases de dados *Library and Information Science Abstracts* (LISA) e *Information Science & Technology Abstracts* (LISTA), conforme ilustrado na Figura 11 e descrito a seguir.

Figura 11 – Primeira etapa da pesquisa.


Fonte: a autora.

Nesse sentido, foi realizada uma pesquisa descritiva e de levantamento (*survey)*, que utilizou a bibliometria, um método quantitativo baseado em análises estatísticas, para análise de dados. Portanto, a análise de dados referente a primeira etapa da pesquisa, assumiu um caráter extremamente quantitativo. Essa análise procurou alcançar os três primeiros[70] objetivos específicos desta tese, para tanto o estudo bibliométrico procurou:

- identificar quantitativamente os autores em ciência da informação que mais publicaram sobre *e-science* e verificar o país de origem desses autores;
- identificar quantitativamente os periódicos que mais publicaram sobre *e-science* e verificar o país de origem desses periódicos;
- identificar as instituições e associações de classe em Ciência da Informação que já publicaram sobre *e-science* e verificar o país de origem dessas instituições e associações;
- identificar a distribuição de autores e instituições por país;

A pesquisa bibliográfica foi realizada no período de 19/03/2013 a 19/06/2013 nas bases de dados internacionais *Library and Information Science Abstracts* (LISA), *Information Science & Technology Abstracts* (LISTA), *Web of Science*, *Education Resources Information Center* (ERIC). No âmbito nacional, a pesquisa foi realizada na *Scientific Electronic Library Online* (SciELO), Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI) e ABCDM.

A base de dados LISA foi selecionada por permitir acesso a artigos de periódicos, desde 1969 até a presente data, de mais 440 títulos de periódicos relacionados com Biblioteconomia e Ciência da Informação, conferindo exaustividade à busca.

A base de dados LISTA foi selecionada por indexar mais de 675 das principais revistas científicas, além de livros, relatórios de pesquisas e protocolos. O período de cobertura dessa base de dados remonta a meados dos anos 1960, o que propicia uma pesquisa exaustiva sobre o tema.

No âmbito nacional, a BRAPCI foi escolhida por ser uma base de dados referencial que relaciona os 27 títulos de periódicos brasileiros em Ciência da Informação, sejam estes correntes ou não, bem como apresenta informações técnicas sobre os respectivos títulos de periódicos.

---

[70] **OE 1** – Identificar os países desenvolvidos que possuem ações de governo para a gestão de dados científicos.
**OE 2** – Analisar as ações de governo de países desenvolvidos sobre a gestão de dados científicos nos países identificados.
**OE 3** – Identificar os principais problemas e as soluções inerentes à construção de uma política estruturada para a gestão de dados científicos.

Para complementar a busca nos periódicos brasileiros também foi realizada uma pesquisa na Base de Dados ABCDM (ex-ABCID), uma base referencial, criada em 2001, que cobre os artigos de periódicos das revistas publicadas no Brasil e em Portugal nas áreas de Arquivologia, Biblioteconomia, Ciência da Informação, Documentação, Museologia e as áreas afins como Administração, História e Tecnologia.

Em função dos índices de revocação e precisão nos resultados das buscas realizadas aleatoriamente, optou-se, na busca estruturada, por utilizar apenas o termo *e-science*, aumentando assim o índice de revocação. O período de cobertura do levantamento bibliográfico compreendeu os anos de 2003 a 2013. O resultado desse levantamento é observado na Tabela 1.

Tabela 1 – Levantamento *sobre e-science*.

| Base de Dados | Nome da Base | 2003/2004 | 2005/2006 | 2007/2008 | 2009/2010 | 2011/2012 | Total de Estudos |
|---|---|---|---|---|---|---|---|
| **Internacionais** | LISA | 11 | 18 | 30 | 35 | 20 | 114 |
| | LISTA | 10 | 22 | 44 | 51 | 19 | 148* |
| | ERIC | 3 | 3 | 3 | 5 | 2 | 16 |
| **Nacionais** | BRAPCI | -- | -- | 1 | 1 | 2 | 4 |
| | SCIELO | -- | 1 | 5 | 6 | 13 | 25 |
| | Base de Dados ABCDM | -- | -- | 1 | 2 | -- | 3 |

Fonte: a autora. Dados coletados até 10/06/2013.[71]

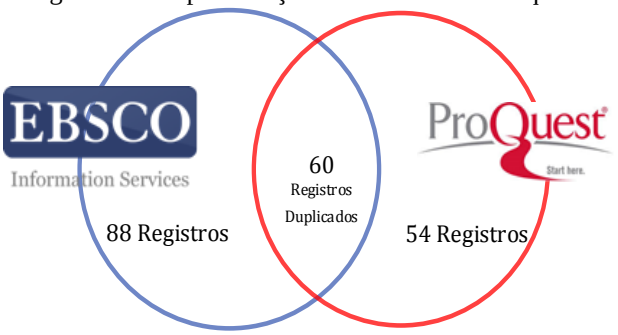
## 3.1 BUSCA BIBLIOGRÁFICA SOBRE A E-SCIENCE NO CONTEXTO MUNDIAL NA CIÊNCIA DA INFORMAÇÃO

Para nortear a pesquisa bibliográfica sobre *e-science* na Ciência da Informação, foi necessário estabelecer um universo de pesquisa bibliográfica. Esse universo foi representado pelos artigos indexados nas bases de dados LISA e LISTA. Em função do total de artigos encontrados ser de 262 sem a eliminação dos registros duplicados, optou-se por trabalhar com o universo da pesquisa, o que permitiu uma maior compreensão sobre a literatura publicada até maio de 2013.

Durante o processo de identificação de registros duplicados, em função da junção dos registros das duas bases de dados pesquisadas, foram eliminados 60 registros, conforme ilustra a Figura 12. Assim, o universo da pesquisa é representado por um total de 202 registros bibliográficos (n = 202).

[71] Dois registros são de 2013.

Figura 12 – Representação do Universo de Pesquisa.




Fonte: a autora.

As variáveis desta análise foram: V1 – Autor; V2 – Ano; V3 – Título do Periódico e V4 – Instituição de origem do autor.

Constitui uma limitação desta análise o fato de não se ter considerado uma base de dados francófona como fonte de dados. Assim, autores franceses e seus respectivos periódicos certamente não serão encontrados de forma expressiva no resultado deste estudo.

Também constitui uma limitação o fato de só terem sido analisadas as bases de dados relacionadas com a Ciência da Informação (LISA e LISTA). Portanto, esta análise compreende apenas uma pequena faceta do tema *e-science*. Há que se considerar que na *Web of Science* foram encontrados 300 registros para a busca. Assim, apenas essa base já superou o quantitativo oriundo da LISA e LISTA. Em razão do exposto, a extrapolação de resultados deve ser feita com cautela no que diz respeito ao tema *e-science* na Ciência da Informação. Apesar de ser um estudo quantitativo, os resultados aqui apresentados não podem ser extrapolados para o tema *e-science* e sua relação com as demais áreas do conhecimento.

### 3.1.1 Procedimentos metodológicos na análise bibliométrica

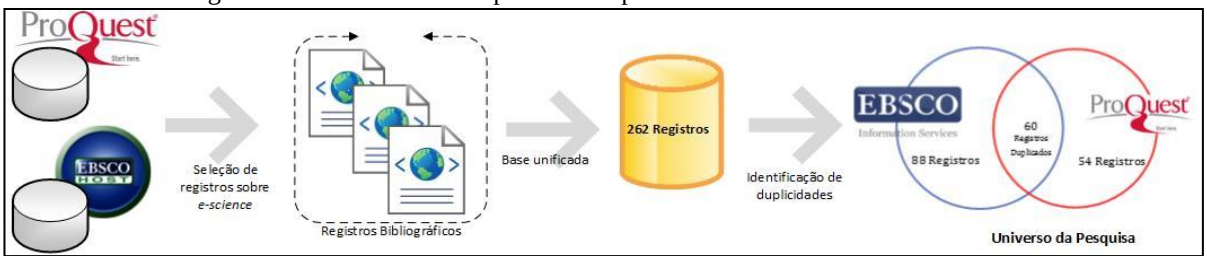
A primeira etapa da pesquisa foi fazer a busca bibliográfica nas bases de dados selecionadas. Na sequência, foram feitos os *downloads* dos arquivos de dados (extensão .txt) da base LISTA e LISA respectivamente. O arquivo da LISTA continha 148 registros, enquanto o da LISA continha 114 artigos, totalizando assim 262 registros.

Na sequência, foi feita a limpeza dos arquivos .txt separadamente e padronizando as *tags* de registro (REG), título do artigo (TI), autor (AU), periódico (PR), data (DT), palavras-chave (PC), organização (AD). Após a limpeza dos arquivos, estes foram salvos em formato .csv, – um arquivo para os registros da LISTA e outro para os registros da LISA.

A etapa seguinte consistiu em copiar os dados do arquivo .csv da LISA para o arquivo .csv da LISTA. Consequentemente, os registros originários da LISA precisaram ser

renumerados sequencialmente a partir do registro 148 (último da LISTA). Após a renumeração, salvou-se o arquivo unificado em formato .xls para permitir ordenação pelo título de forma a identificar os registros duplicados. Após a conferência do título do artigo, foi verificado o nome do periódico e autor. Uma vez confirmada a semelhança, um dos registros foi considerado duplicado e eliminado do arquivo unificado. Eliminaram-se assim 60 registros. As etapas operacionais acima relatadas podem ser observadas na Figura 13.

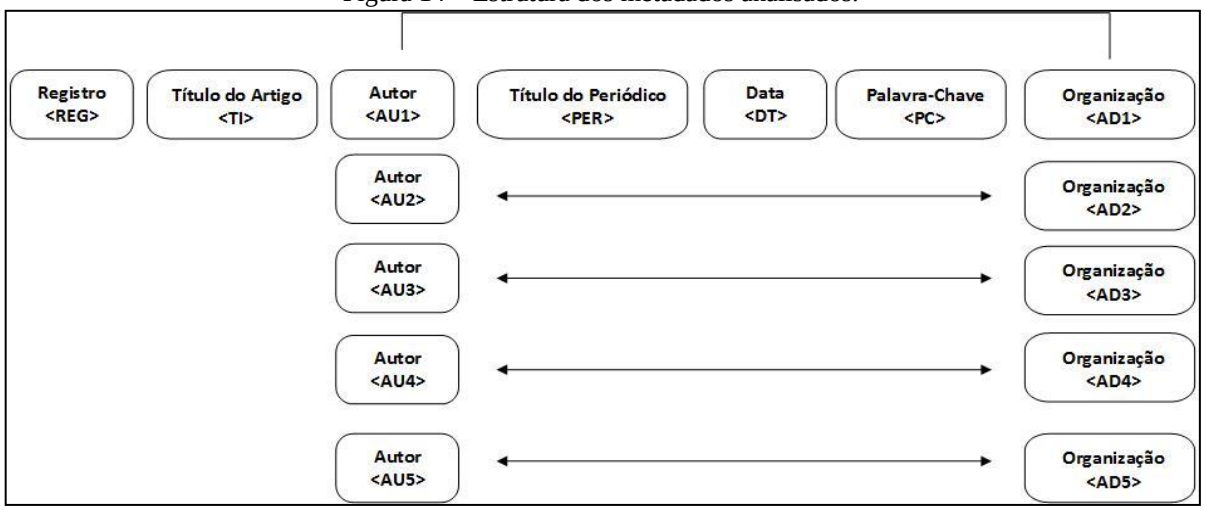
Figura 13 – Procedimentos operacionais para a análise bibliométrica – Fase 1.




Fonte: a autora.

A quarta etapa teve início com a organização de uma tabela matriz que tivesse os dados de número de registro, título artigo, autor (es), periódico, data e organização. Também foi elaborada uma planilha só com o número de registro e o autor, uma vez que um artigo pode ter mais de um autor. Consequentemente, foi elaborada mais uma planilha que continha o número do registro e a instituição de filiação do autor. Assim, para analisar os dados, foram elaboradas inicialmente três planilhas – uma matriz, uma de autor(es) e outra de instituição do(s) autor(es). A estrutura completa dos metadados analisados é observada na Figura 14.

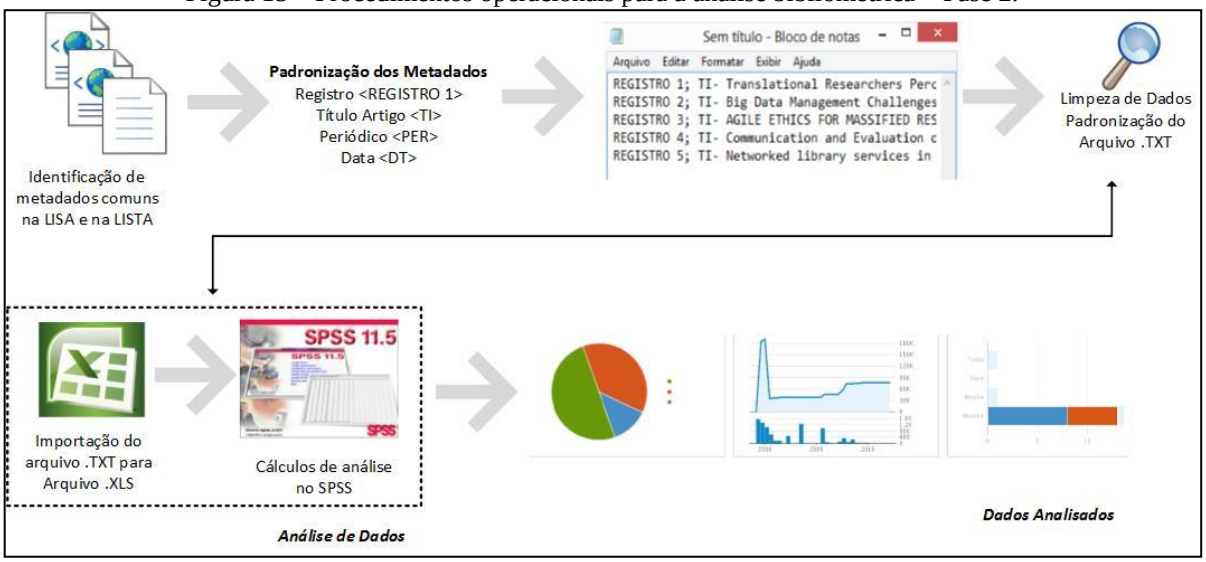
Figura 14 – Estrutura dos metadados analisados.




Fonte: a autora.

A etapa seguinte foi realizada com o *softwar*e SPSS que importou os dados da planilha matriz em formato .xls. Com o auxílio do SPSS foi possível identificar o número de publicações por ano e o periódico que mais publicou sobre o tema. A planilha de autores permitiu identificar os autores que mais produziram artigos sobre *e-science*, enquanto a planilha de instituições permitiu identificar as instituições com destaque no tema. Todas as etapas até aqui mencionadas foram feitas com o auxílio do SPSS, porém, os gráficos foram elaborados no Excel. A Figura 15 ilustra as demais etapas operacionais que viabilizaram a análise bibliométrica.

Figura 15 – Procedimentos operacionais para a análise bibliométrica – Fase 2.




Fonte: a autora.

A identificação das palavras-chaves mais usadas na indexação das bases foi realizada com o apoio do Excel e ilustrada em uma nuvem de *tags* com o apoio do aplicativo *Wordle*[72] disponível gratuitamente na internet.

Foram recuperados nas buscas 31 registros referentes a artigos publicados nos anos de 2011 e 2013. Esses registros não foram levados no momento de se contabilizar as publicações sobre o tema por ano. Tais artigos foram desprezados em função dos respectivos anos ainda não terem sido completamente indexado pelas bases dados selecionadas no estudo. Merece ser comentado que a eliminação desses registros não interfere no quadro de autores que mais publicaram sobre o tema, tão pouco no quadro de periódicos que mais publicaram sobre *e-science*.

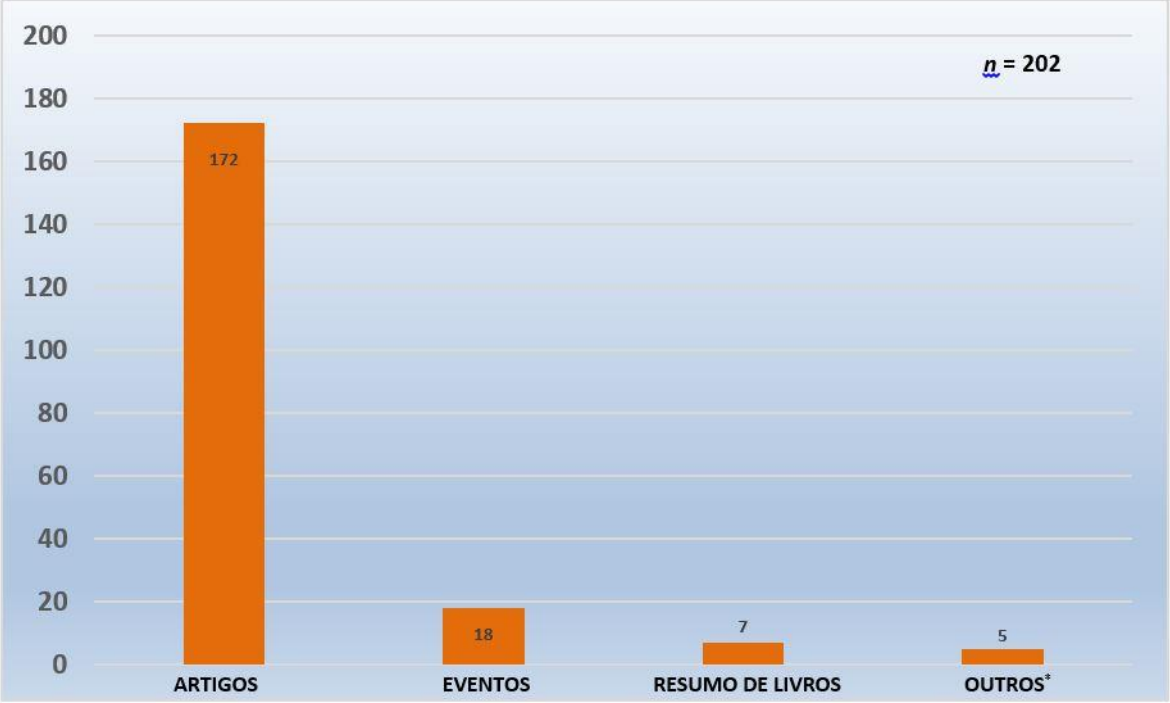
[72] Wordle. Disponível em: <http://wordle.net>.

### 3.1.2 Resultado da análise bibliométrica[73]

Os dados foram analisados com o apoio dos *softwares* Excel e SPSS. O SPSS foi utilizado para realizar as análises descritivas, calculando seus percentuais, percentuais válidos e valores acumulados. O Excel, por sua vez, foi utilizado para elaborar os gráficos. Já as nuvens de *tags* foram criadas com o aplicativo *Wordle*.

O Gráfico 2 ilustra os tipos de publicações indexados pelas bases de dados LISA e LISTA. Estão agrupados em *conference* os *proceedings*, *conferences reports* e *conferences*. O agrupamento outros é formado por: editoriais, opiniões (*opinion*), cartas (*letters*) e um estudo de caso (*case study*) que não foi indexado como artigo.

Gráfico 2 - Tipos de documentos indexados pela LISA e LISTA (2003 – 2013).



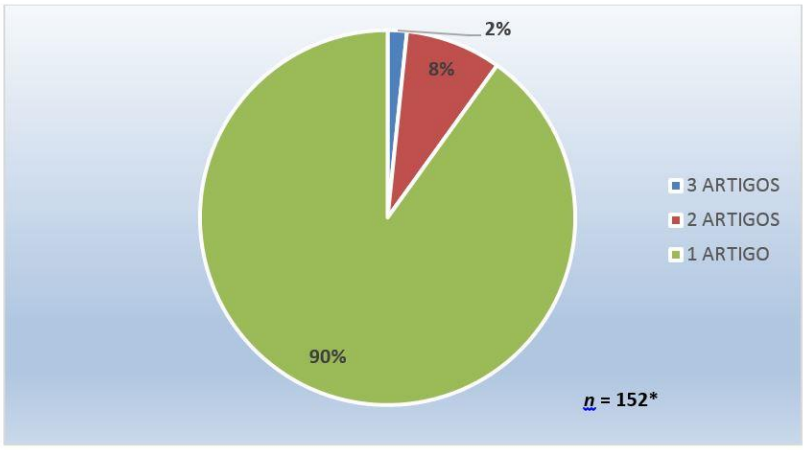

Fonte: a autora.
*Outros (editorial, opinião, cartas).

A análise do Gráfico 2 indica que há uma preponderância de artigos científicos no conteúdo das bases de dados. Merece ser comentado que dos sete resumos de livros, seis tratavam do livro "*The data deluge: can libraries cope with e-science*?" – publicado em 2010, pelos editores Denna B. Marcum e Gerald George.

[73] Os resultados deste estudo bibliométrico foi publicado como artigo de periódico (COSTA; CUNHA, 2015) na Revista Encontros Bibli (Revista Eletrônica de Biblioteconomia e Ciência da Informação).

O Gráfico 3 demonstra que apenas 10% dos autores publicaram mais de um artigo sobre *e-Science* no período de 2003 a 2013. Ressalta-se que foram levados em consideração nesse gráfico apenas os artigos científicos, ou seja, foram excluídos da análise os documentos agrupados como *conference*, *book review* e outros. O Gráfico confirma o fenômeno da Lei de Lotka, também conhecida como lei de frequência da produtividade científica, que considera que poucos autores produzem muito e, por outro lado, muitos autores produzem pouco.

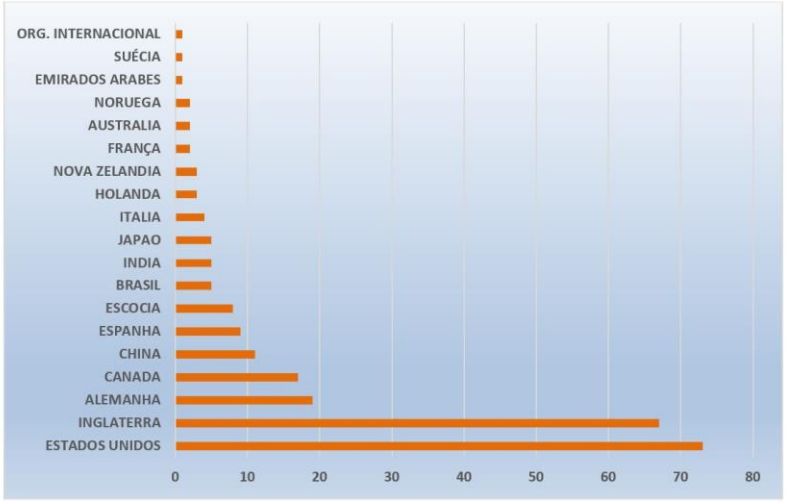
Gráfico 3 - Distribuição da quantidade de autores *versus* artigos publicados sobre eScience - Período (2003/2013).




Fonte: a autora.
*Dos 172 artigos, 20 não possuíam informação sobre o autor, logo, n = 152.

Com fundamento nos autores que informaram sua instituição de origem no artigo, é possível afirmar que apesar do termo *e-science* ter surgido na Inglaterra, autores vinculados a instituições nos Estados Unidos já produziram mais sobre o tema, conforme pode ser observado no Gráfico 4, onde na sequência aparece a Inglaterra.

Gráfico 4 - Distribuição dos autores por País (com fundamento na instituição informada no artigo).




Fonte: a autora.

Há que se ressaltar, na leitura do Gráfico 3, que nem todos os autores informaram a sua instituição (um total de 73 registros), o que impossibilitou identificar o País de origem da produção. Por outro lado, uma minoria de registros (20 registros) não possuía informação de autoria, logo também não possuía informação da instituição. Incluem-se nestes casos alguns resumos de livros, cartas, dentre outros.

Ainda sobre o Gráfico 3, constata-se que os três maiores produtores de documentos sobre *e-science* são membros do G8. Por outro lado, a produção da França, apesar de membro do G8, é menor que a do Brasil. Nesse aspecto, é importante lembrar que as bases de dados americanas pouco indexam documentos franceses. Assim, o Gráfico 3 deve ser analisado cuidadosamente, pois para evitar essas disparidades, essa pesquisa precisava ter incluído os registros da base de dados Francis.

Dentre os demais países, percebe-se a presença da produção de artigos entre os membros do BRICS2, representando assim a emergência desses países tanto em termos industriais quanto no desenvolvimento científico.

Dentre as instituições, obtiveram destaque pelo número de autores e publicações, as relacionadas na Tabela 2.

Tabela 2 - Instituições com maior produção de documentos sobre *e-science* (2003-2013).

| PAIS | INSTITUIÇÃO | OCORRÊNCIAS (n = 202) |
|---|---|---|
| Inglaterra | University of Oxford | 13 |
| EUA | Purdue University | 10 |
| EUA | University of California | 7 |
| Inglaterra | University of Manchester | 7 |
| China | Chinese Academy of Sciences | 6 |
| Escócia | University of Edinburgh | 6 |
| Inglaterra | Arts and Humanities e-Science Support Centre | 5 |
| Inglaterra | Oxford eResearch Centre | 5 |
| EUA | University of Washington | 5 |
| EUA | Microsoft Corporation | 4 |

Fonte: a autora.

A partir da análise da Tabela 2, fica evidente a maturidade da Inglaterra no tratamento de dados oriundos da *e-Science*, pois na tabela figuraram não apenas universidades, mas também centros especializados em *e-Science* e *e-Research*. O resultado não causa surpresa, afinal no ano 2000 a Inglaterra já possuía o National e-Science Center.

Outro centro da Inglaterra que apareceu na análise, mas em menor evidência, foi o e-Science Centre. Também apareceu o e-Science Core Programe. Esse programa é gerido pelo Conselho de Pesquisa em Ciências da Engenharia e Física, em nome das comunidades de todos os Conselhos de Pesquisa. Tem apoiado o desenvolvimento de tecnologias genéricas, como o *software* conhecido como *middleware* – necessário para permitir que diferentes recursos trabalhem de forma integrada por meio de redes, bem como criem grids computacionais.

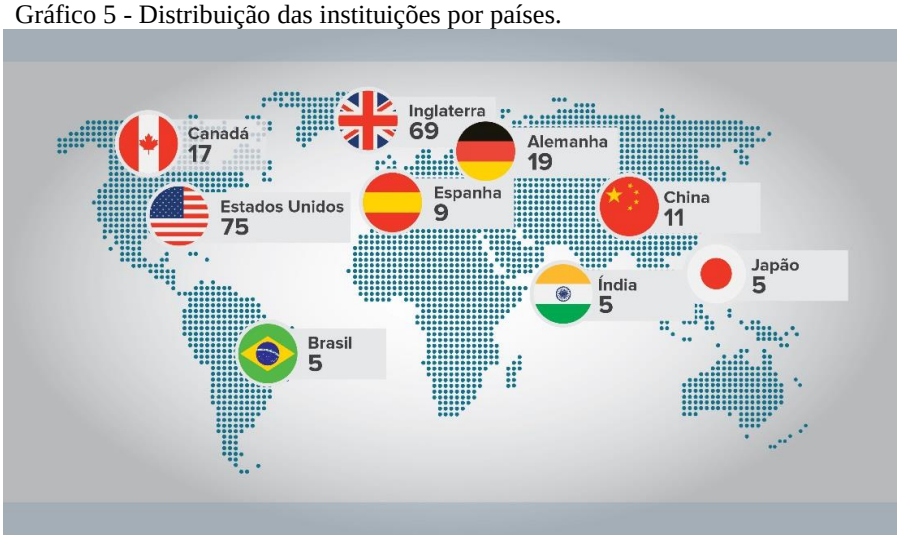
O fato da Inglaterra, já em 2001, ter lançado, de forma pioneira, um programa, conferiu-lhe certa maturidade no tratamento de dados oriundos da *e-Science*. Tal afirmação se justifica pelo volume de trabalhos vinculados ao tema pelas Universidades de Oxford e Manchester. Também apresentam trabalhos significativos sobre o tema o Arts and Humanities e-Science Support Centre, o Oxford eResearch Centre, o National e-Science Centre, o eScience Centre, bem como o e-Science Core Programme.

Já no contexto dos Estados Unidos destacam-se os trabalhos vinculados às universidades de Purdue e de Washington. Além disso, percebe-se um interesse de grandes corporações, como a Microsoft, pelo tema, destacando-se o fato de Tony Hey, atual vice-presidente da área de pesquisa da Microsoft, ter sido o diretor do e-Science Core Programme no Reino Unido.

A única associação de classe que foi recuperada pelos registros da LISA e LISTA foi a Association of Research Libraries (ARL). Foi constatado que sete publicações envolviam a ARL, são elas: 1) *ARL launches e-Science agenda*; 2) *e-Science and data support service: a study of ARL members institutions*; 3) *e-Science in research libraries: new ARL reports to research libraries*; 4) *Reinventing librarianship: themes from ARL-CNI Forum*; 5) *ARL activities*; 6) *ARL e-science survey* e 7) *Libraries and changing research 137 practice: a report of the ARL-CNI Forum on e-research and cyberinfrastructure*.

O número de instituições e sua distribuição por alguns dos países que se destacaram na análise dos dados pode ser observado no Gráfico 5. Esse gráfico sintetiza os países que mais produziram sobre e-science no âmbito da Ciência da Informação, portanto, permitiu alcançar satisfatoriamente o primeiro objetivo específico desta tese.

Gráfico 5 - Distribuição das instituições por países.




Fonte: a autora.

A análise do campo referente ao local de publicação permitiu fazer um ranking dos periódicos que mais publicaram sobre o tema *e-science*. O ranking pode ser observado na Tabela 3.

Tabela 3 - Periódicos que mais publicaram sobre *e-science.*

| PERIÓDICO | FREQ. | % | n = 202 % VÁLIDO | % ACUMULADO |
|---|---|---|---|---|
| *Ariadne* | 10 | 5,0 | 5,0 | 5,0 |
| *IEEE Intelligent Systems* | 7 | 3,5 | 3,5 | 8,4 |
| *D-Lib Magazine* | 6 | 3,0 | 3,0 | 11,4 |
| *Information Services & Use* | 6 | 3,0 | 3,0 | 14,4 |
| *Library Hi Tech News* | 6 | 3,0 | 3,0 | 17,3 |
| *Information Today* | 5 | 2,5 | 2,5 | 19,8 |
| *International Journal of Web Services Research* | 5 | 2,5 | 2,5 | 22,3 |
| *Issues in Science & Technology Librarianship* | 5 | 2,5 | 2,5 | 24,8 |
| *Sci-Tech News* | 5 | 2,5 | 2,5 | 27,2 |
| *Electronic Library* | 4 | 2,0 | 2,0 | 29,2 |
| *Information, Communication & Society* | 4 | 2,0 | 2,0 | 31,2 |
| *Journal of Information Processing & Management* | 4 | 2,0 | 2,0 | 33,2 |
| *ABI Technik* | 3 | 1,5 | 1,5 | 34,7 |
| *ARL: Bimonthly Rep. on Res. Library Issues & Actions* | 3 | 1,5 | 1,5 | 36,1 |
| *Communications of the ACM* | 3 | 1,5 | 1,5 | 37,6 |
| *International Journal on Digital Libraries* | 3 | 1,5 | 1,5 | 39,1 |
| *Journal of Library Administration* | 3 | 1,5 | 1,5 | 40,6 |
| *Journal of the Medical Library Association* | 3 | 1,5 | 1,5 | 42,1 |
| *Library Hi Tech* | 3 | 1,5 | 1,5 | 43,6 |

Fonte: a autora com fundamento em dados da LISA e LISTA em junho de 2013.

A Tabela 3 revela que o periódico que mais publicou sobre o tema foi o Ariadne, periódico de origem inglesa voltado para profissionais da informação. Sua área de concentração é sobre bibliotecas digitais, abordando os temas de evolução dos serviços de informação e redes de informação. Em segundo lugar aparece o periódico IEEE Intelligent Systems, periódico acadêmico voltado para engenheiros de *software*, *designers*, gerentes de sistemas de informação, engenheiros de conhecimento, pesquisadores e profissionais em áreas como finanças, manufatura, medicina, defesa e as ciências. A publicação é produzida pela IEEE Computer Society e o seu patrocínio é oferecido pela Association for the Advancement of Artificial Intelligence (Associação para o Avanço da Inteligência Artificial).

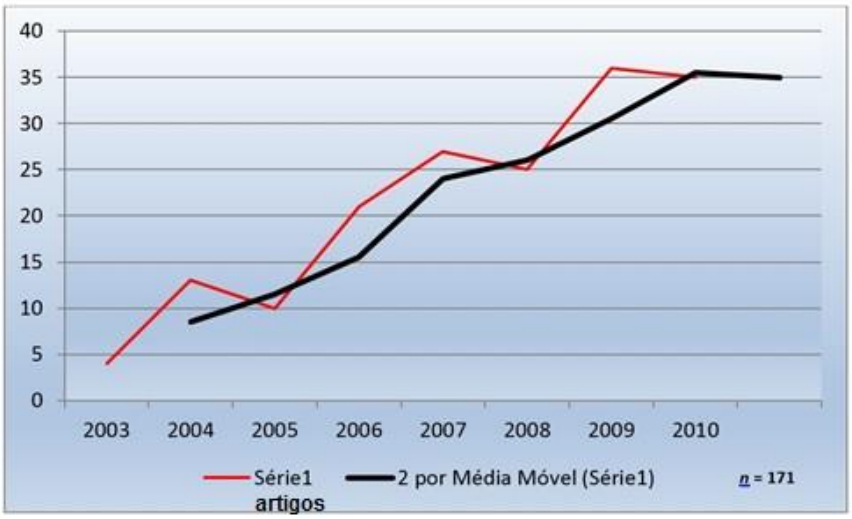
Os demais periódicos estão voltados para a ciência da informação. Sendo que cinco deles são periódicos 'profissionais' – "Library Hi Tech News", "Information Today", "Comunication of the ACM", "Journal of the Medical Library Association", "Library Hi Tech" e "ARL: Bimonthly Report on Research Library Issues & Actions", este último publicado pela ARL.

No que diz respeito aos eventos, a IATUL[74] Annual Conference Proceedings foi a que mais teve ocorrência nas bases de dados. Além deste evento, foram identificados os seguintes: eSciDoc, International Bielefeld Conference, NFAIS[75] Annual Conference e a ICSTI Annual Conference.

Para saber se o tema *e-Science* está crescendo dentro da Ciência da Informação, os registros da base de dados foram organizados por ano, de forma a identificar se ocorre um crescimento anual da produção literária da área. O Gráfico 6 ilustra o crescimento da produção até o ano de 2010.

[74] Internacional Association of Scientific and Technological Libraries.
[75] National Federation of Advanced Information Services.

Gráfico 6 - Produção de artigos por ano.




Fonte: a autora com fundamento em dados da LISA e LISTA em junho de 2013.

Foram desconsiderados para o Gráfico 6 a produção referente a 2011[76] e 2012[77] em função do período de atualização das bases de dados. Ou seja, há um entendimento de que o volume de publicações indexados em 2011 e 2012 ainda não reflete a realidade em função do *gap* que ocorre entre o período de publicação do volume de um periódico e o tempo que as bases de dados levam para incluí-lo e indexá-lo.

Merece ser ressaltado que o Gráfico 6 demonstra um leve decréscimo no ano de 2010 em relação a 2009 e uma posterior estagnação no volume de publicações. Uma possível explicação para a leve queda do gráfico é a criação de periódicos específicos sobre *e-Science*, que, dada a sua recente criação, ainda não foram indexados pelas bases de dados LISA e LISTA. Como exemplo, cita-se os periódicos "International Journal of Digital Curation" e o "Journal of eScience Librarianship", criados em 2006 e 2012, respectivamente.

Para identificar qual seria o melhor termo para se realizar uma busca sobre o tema, foi realizada uma análise das palavras-chave indexadas nos 202 registros. A Tabela 4 ilustra os 20 termos mais indexados.

[76] Em junho de 2013, as bases de dados possuíam apenas 17 registros referentes a 2011.
[77] Em junho de 2013, as bases de dados possuíam apenas 12 registros referentes a 2012.

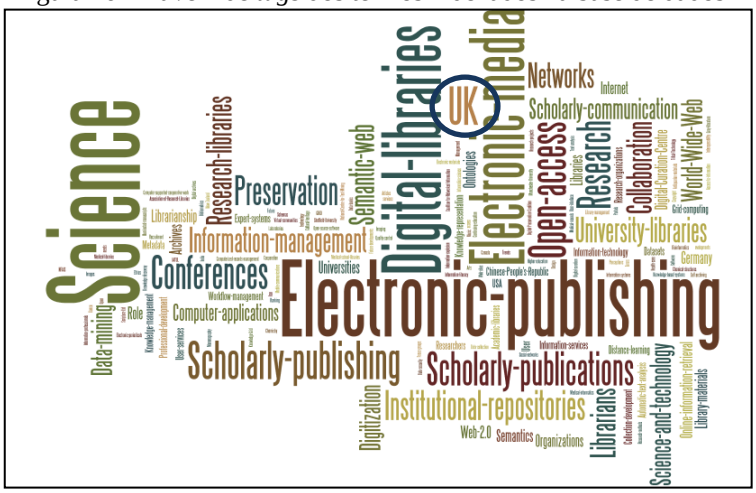
Tabela 4 - Vinte palavras-chave mais indexadas.

| PALAVRA-CHAVE | FREQUENCIA | % | % VALIDO | % ACUMULADO |
|---|---|---|---|---|
| Science | 30 | 3,4 | 3,4 | 3,4 |
| Digital-libraries | 22 | 2,5 | 2,5 | 5,9 |
| Electronic publishing | 22 | 2,5 | 2,5 | 8,4 |
| Research | 21 | 2,4 | 2,4 | 10,8 |
| Electronic-media | 13 | 1,5 | 1,5 | 12,3 |
| Librarians | 12 | 1,4 | 1,4 | 13,7 |
| Information resources management | 10 | 1,1 | 1,1 | 14,8 |
| Computer network resources | 9 | 1 | 1 | 15,8 |
| Conferences | 9 | 1 | 1 | 16,9 |
| Libraries | 9 | 1 | 1 | 17,9 |
| Open access | 9 | 1 | 1 | 18,9 |
| Scholarly publishing | 9 | 1 | 1 | 20 |
| UK | 9 | 1 | 1 | 21 |
| E-science | 8 | 0,9 | 0,9 | 21,9 |
| Research libraries | 8 | 0,9 | 0,9 | 22,8 |
| Scholarly publications | 8 | 0,9 | 0,9 | 23,7 |
| Collaboration | 7 | 0,8 | 0,8 | 24,5 |
| Information resources | 7 | 0,8 | 0,8 | 25,3 |
| Information science | 7 | 0,8 | 0,8 | 26,1 |
| Information services | 7 | 0,8 | 0,8 | 26,9 |

Fonte: a autora com fundamento em dados da LISA e LISTA em junho de 2013.

Interessante destacar que a análise dos 114 registros recuperados na base de dados LISA revela que o termo *e-Science* não foi usado como indexador (palavra-chave) em nenhum dos registros, conforme ilustrado na Figura 16.

Figura 16 - Nuvem de *tags* dos termos indexados na base de dados LISA.



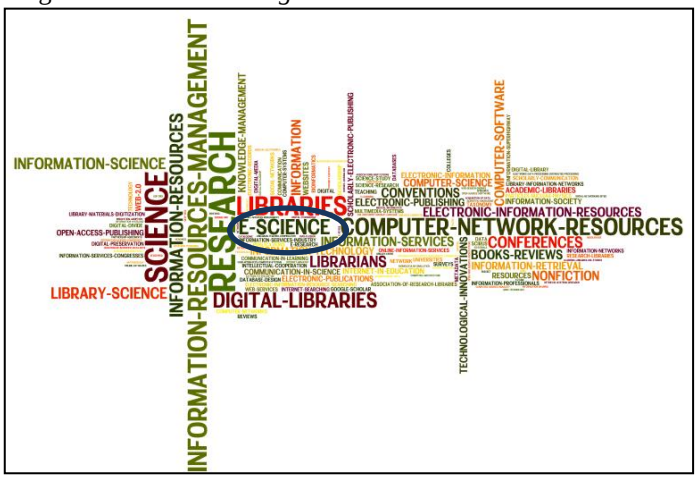

Fonte: a autora.

Percebe-se que as palavras-chave utilizadas na LISA são mais genéricas, por exemplo: ciência, publicação eletrônica, bibliotecas digitais, em 141 publicações acadêmicas dentre outros. Esses termos certamente aumentariam a revocação da busca, mais diminuiriam sua precisão. Infere-se que há uma possibilidade de o termo não ter sido utilizado por ser muito novo e não constar ainda nos vocabulários controlados.

Interessante observar o uso do termo UK (United Kingdom) como palavra-chave. Infere-se que esse termo revela o lastro da Inglaterra em projetos na área de e-Science.

Por outro lado, em um mesmo período de tempo, verificou-se que a base de dados LISTA usou o termo *e-Science* como indexador, conforme ilustra a Figura 17.

Figura 17 - Nuvem de *tags* dos termos indexados na base de dados LISTA.




Fonte: a autora.

Na LISTA também é observado o uso de termos mais genéricos como palavras-chave. A exemplo cita-se: ciência da informação, bibliotecas digitais, bibliotecários. Porém, o termo *e-Science* aparece com razoável destaque na nuvem, o que indica a quantidade de vezes que foi utilizado.

Para compreender de fato porque a LISA não utilizou o termo e a LISTA sim, é necessário fazer uma entrevista com os responsáveis pela política de indexação das respectivas bases de dados. Qualquer inferência mais precisa pode ser leviana.

O que se pode afirmar, com fundamento nos dados, é que a busca realizada pela pesquisadora teve um retorno de 262 registros (202 excluindo os registros duplicados) pelo fato de não ter se delimitado o campo para a busca. Se a delimitação tivesse sido feita na palavra-chave, o resultado seria frustrante, pois o termo *e-Science* não foi indexado pela LISA e não teve absoluta expressão na busca realizada na LISTA. Entende-se que os registros foram recuperados pelo termo estar no título ou no resumo.

Merece ser comentado que os termos: dilúvio de dados, ciberinfraestrutura e *big data* que seriam mais precisos, também não foram utilizados nas respectivas bases de dados. Nesse cenário, é interessante relembrar que o termo *cyberinfrastructure* foi cunhado pela National Science Foundation. Assim, e possível inferir que o termo criado no Reino Unido (eScience) tenha ganhado mais adeptos, levando-o a ser um termo indexador.

Conclui-se, portanto, que ao se realizar buscas relacionadas ao fenômeno da *e-science* é preciso aumentar o índice de revocação, pois uma busca precisa certamente levará a um resultado frustrante. Aparentemente o resultado mais efetivo da busca deu-se em função da recuperação do termo *e-science* nos campos de título e do resumo.

A análise dos dados revelou que dois são os autores que mais publicaram sobre e-Science, sendo que todos eles publicaram três artigos no período de 2003-2011. A lista dos autores pode ser observada no Tabela 5.

Tabela 5 - Autores que mais publicaram sobre *e-science*.

| AUTOR | QTD TOTAL | QTD. ARTIGOS | QTD OUTROS |
|---|---|---|---|
| Blanke, Tobias[78] | 5 | 3 | 2 |
| Mullins, James L[79]. | 4 | 3 | 1 |
| Aschenbrenner, Andreas | 3 | 3 | -- |
| Becker, Carolin | 3 | 2 | 1 |
| Dovey, Matthew J. | 3 | 3 | -- |
| Gore, Sally A. | 3 | 3 | -- |
| Guo Jing | 3 | 3 | -- |
| Hai Zhuge | | 3 | -- |
| Kallerbon, Reiner | 3 | 2 | 1 |
| Lu Xiaobin | | | |
| Paterson, Lorraine | 3 | 3 | -- |
| Rusnak, Ute | 3 | 2 | 1 |

Fonte: a autora.

Interessante observar que em pesquisa no Google Acadêmico sobre *eScience* e bibliotecas, os artigos em destaque são de Hey, o primeiro "The data deluge: an e-science perspective", citado por 367 e; o segundo: "E-science and its implications for the library community", citado por 64. Esse segundo artigo também é recomendado na bibliografia sobre o tema elaborada por Szigeti e Wheeler (2011). Porém, na busca realizada na LISA e no LISTA o Hey aparece como autor de apenas dois artigos (HEY, 2004; 2006), por isso não foi representado na Tabela 5.

[78] Autor cuja a instituição está vinculado ao Reino Unido.
[79] Autor cuja instituição está vinculado ao Estados Unidos.

O único autor que figura na lista dos que mais publicaram e está relacionado na bibliografia elaborada por Szigeti e Wheeler (2011) é James L. Mullins[80], professor da Purdue University. Os resultados apresentados na Tabela 5 confirmam que nem sempre o autor que mais publica é o mais citado.

### 3.1.3 Considerações sobre o resultado da análise bibliométrica

As contribuições da Ciência da Informação para a *e-Science* ainda são incipientes. Tal afirmativa se faz possível em função da quantidade de trabalhos publicados na base de dados Library Information Science Technology Abstracts (LISTA), bem como na base de dados Library Information Science Abstracts (LISA). No entanto, os dados revelam uma curva crescente no volume de trabalhos sobre o tema.

A análise bibliométrica revelou que os países que mais têm iniciativas na *e-science* são os Estados Unidos e o Reino Unido – o que permitiu alcançar o primeiro objetivo específico desta tese, bem como nortear a revisão de literatura sobre políticas para a gestão de dados científicos nesses respectivos países, contribuindo, dessa forma, para alcançar o segundo e terceiro objetivos específicos desta tese.

O Reino Unido se destaca no que diz respeito a centros de tratamento de dados da *e-Science* em relação aos outros países. Atribui-se esse destaque ao pioneirismo da Inglaterra no tratamento desses dados. Sobre o assunto, merece ser comentado que a composição de membros dos Comitês Consultivos (Technical Advisory Board, Secretariat, Organisational Advisory Board) revela uma participação predominante de países como Estados Unidos, Reino Unido e Austrália no ambiente de *e-science*.

A única associação profissional que emergiu da análise de dados foi a Association of Research Libraries, o que denota um campo ainda desconhecido para os profissionais da informação. Para atualizar-se recomenda-se a participação no IATUL, bem como a leitura dos títulos de periódicos Ariadne e IEEE Intelligent Systems.

Os autores que se destacam em termos de publicação sobre o tema são Tobias Blanke (Reino Unido) e James L. Mullins (EUA). No entanto, recomenda-se a leitura dos trabalhos de Tony Hey (Reino Unido), autor pioneiro na área de *e-science*. Apesar de não aparecerem nos

[80] Enabling international access to scientific data sets: creation of the distributed data curation center.

resultados do estudo bibliométrico em função dos termos utilizados na indexação, na Ciência da Informação, destacam-se os trabalhos de Cristine Borgman (EUA).

## 3.2 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo encontram-se o material e o método que foram utilizados para a realização deste trabalho.

### 3.2.1 Caracterização da Pesquisa

A partir do resultado do estudo bibliométrico explanado no Capítulo 3.4 – *A construção da metodologia*, realizado em 2013, chegou-se à conclusão que à época havia um número incipiente de estudos que tratassem sobre a gestão de dados científicos, bem como, os problemas e desafios inerentes à construção de uma política estruturada para a gestão desses dados. Também se mostrou incipiente o número de estudos referentes às necessidades de informação de cientistas que trabalham com dados científicos em larga escala (*big data*, *e-science*).

Logo, esta tese, quanto aos seus objetivos, apresenta-se como exploratória, pois há um consenso na literatura de que quando há pouco conhecimento acumulado e sistematizado sobre o assunto a ser pesquisado, a pesquisa, quanto aos seus objetivos, é exploratória (GIL, 2006; MATIAS-PEREIRA, 2007; VERGARA, 2004).

A maneira pela qual o pesquisador coleta e analisa seus dados configura o processo de pesquisa. Este processo pode ter um enfoque qualitativo, quantitativo ou misto (COLLIS; HUSSEY, 2005). Esta pesquisa priorizou o caráter qualitativo na coleta de seus dados, uma vez que não empregou dados estatísticos como centro do processo de análise do problema de pesquisa. No entanto, quando necessário, os métodos quantitativos foram utilizados, como, por exemplo, na análise bibliométrica sobre a literatura em *e-science*, assim como na análise de construções de perguntas com opções de resposta em escala Likert aplicadas durante a segunda etapa da pesquisa em amostra intencional de pesquisadores doutores brasileiros envolvidos com o tema e com funcionários de agências de fomento no Brasil.

Segundo Minayo (1996; 2007), é a pesquisa qualitativa que permite a revelação de processos sociais ainda pouco conhecidos referentes a grupos particulares e que propicia a construção de novas abordagens, revisão e criação de novos conceitos e categorias.

A pesquisa exploratória explanada nesta tese apresentou uma visão geral da política de gestão de dados científicos em países desenvolvidos, o que lhe conferiu um caráter descritivo. Para Vergara (2004) e Collis e Hussey (2005), a pesquisa descritiva identifica, obtém e expõe informações sobre as características de um problema ou questão. Ela vai além da pesquisa exploratória ao examinar um problema, uma vez que avalia e descreve as características das questões pertinentes.

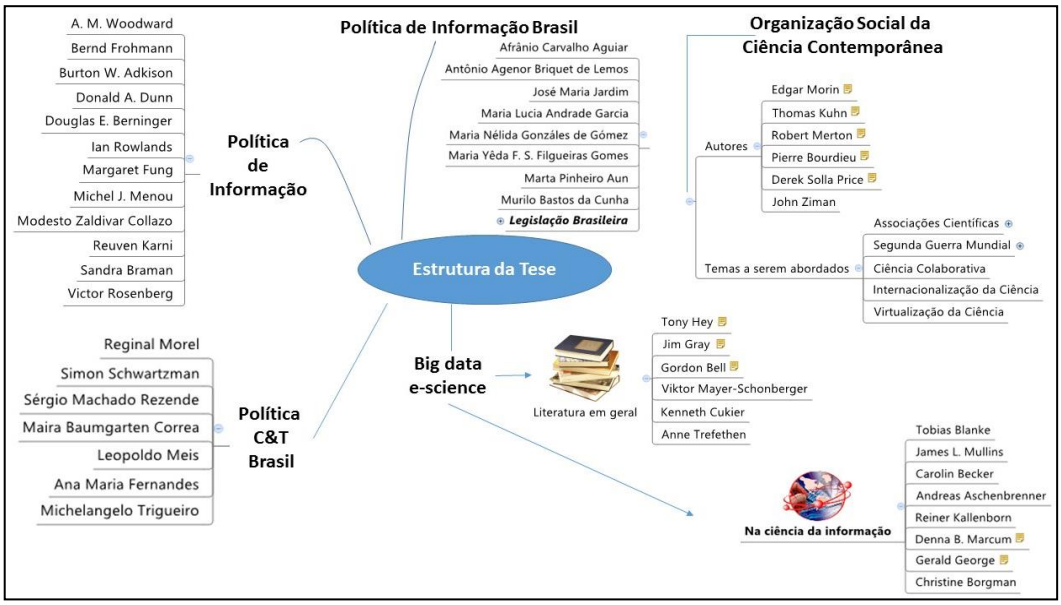
Para melhor descrever as políticas selecionadas como objeto de estudo, o procedimento utilizado foi a avaliação comparativa, que permitiu a identificação dos principais desafios e das soluções que os países desenvolvidos enfrentaram para desenvolver sua política. Para Marconi e Lakatos (2007), o método comparativo estuda semelhanças e diferenças em diversos tipos de grupos e permite uma análise do dado concreto, deduzindo do mesmo elementos constantes, abstratos e gerais.

A abordagem de investigação qualitativa utilizada foi a *Grounded Theory* (também chamada de *Teoria Fundamentada em Dados)* que permitiu abarcar a riqueza e a complexidade da análise comparativa das respostas dos funcionários de agências de fomento versus doutores envolvidos com o tema (*e-science*). O *software* de apoio a análise qualitativa à luz da Teoria Fundamenta em Dados foi o Nvivo 10[81].A Teoria Fundamentada em Dados procura desenvolver uma teoria a partir dos dados sistematicamente recolhidos e analisados, ou fazer descrições muito úteis (CHARMAZ, 2009; STRAUSS; CORBIN, 2008).

Esta tese, considerando a etapa de construção da metodologia, se deu em três etapas. A primeira, referente ao processo de construção da metodologia, compreendeu a fase de ampla revisão bibliográfica da literatura internacional sobre aspectos relacionados a *e-science*, bem como a política de informação e gestão de dados científicos em países desenvolvidos. No contexto brasileiro, foram realizas buscas na BRAPCI sobre a política de desenvolvimento científico e tecnológico e a política nacional de informação. Dessa forma, o marco teórico do estudo se configurou conforme a ilustração da Figura 18.

[81] O Nvivo é um *software* de análise qualitativa que permite criar categorias, codificar respostas, filtrar dados e questionar os dados sobre as perguntas de investigação. Por meio do *software* é possível manter um registro histórico de processo de investigação realizado durante a pesquisa. O *software* é desenvolvido pela QSR International.

Figura 18 – Configuração do marco teórico da tese.

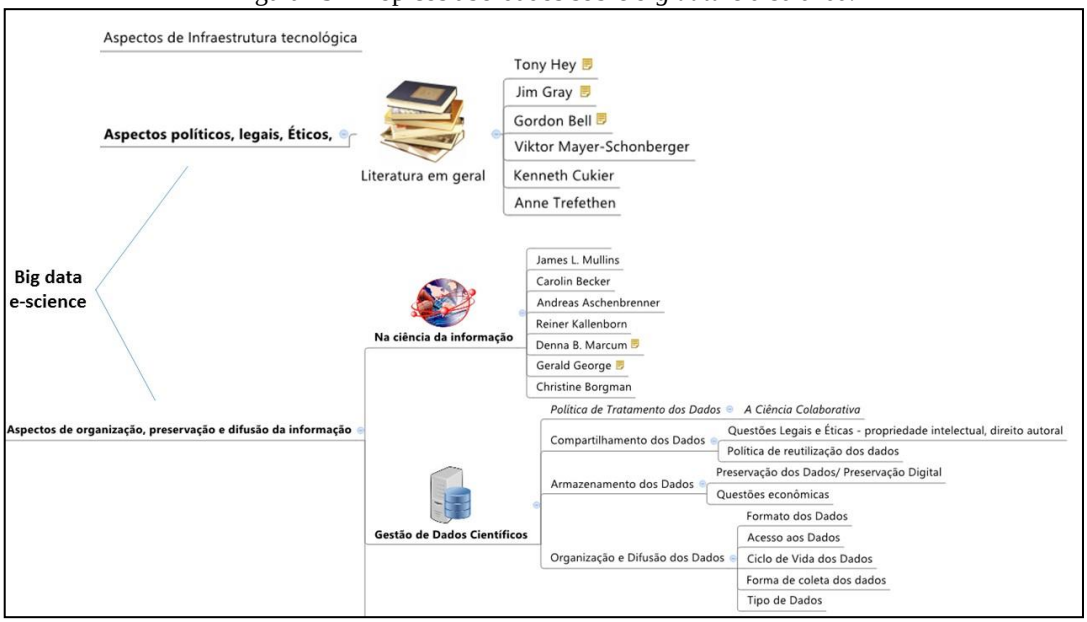



Fonte: a autora.

No que diz respeito à estrutura de tópicos a serem abordados sobre *big data* e *e-science*, esta pesquisa não teve a intenção de abordar aspectos inerentes à infraestrutura tecnológica necessária para apoiar os dados oriundos da *e-science*.

Com isso, a revisão de literatura se concentrou nas questões políticas, legais e éticas necessárias para configurar uma política de gestão de dados científicos. Além disso, quanto aos aspectos sobre a gestão da informação, o estudo abordou questões intrínsecas a coleta, organização, preservação e difusão da informação, conforme ilustrado na Figura 19.

Figura 19 – Tópicos abordados sobre *big data* e *e-science*.




Fonte: a autora.

O resultado de uma pesquisa pode ser classificado como aplicado ou puro, sendo que a pesquisa aplicada é aquela motivada pela necessidade de se resolver um problema concreto (COLLIS; HUSSEY, 2005). Logo, o fato de a pesquisa desenvolver diretrizes políticas para a gestão de dados científicos no Brasil, confere a ela um caráter aplicado.

Pelo exposto, este trabalho caracteriza-se, quanto aos seus objetivos, como uma pesquisa exploratória e descritiva. No que diz respeito ao processo de pesquisa ela se caracteriza como quali-quantitativa (mista) e quanto ao seu resultado ela é uma pesquisa aplicada.

### 3.2.2 A Teoria Fundamentada em Dados

A Teoria Fundamentada em Dados, originalmente conhecida como *Grounded Theory*, foi desenvolvida pelos sociólogos Barney G. Glaser e Anselm L. Strauss durante a realização de estudos sobre o processo de morte em hospitais. O primeiro livro sobre o tema foi publicado em 1967 – "The Discovery of Grounded Theory: strategies for qualitative research".

Glaser e Strauss (1965; 1967) opuseram-se aos pressupostos metodológicos quantitativos positivistas da década de 1960 e ofereceram estratégias sistemáticas para a prática da pesquisa qualitativa. Assim, a grande contribuição do livro foi defender o desenvolvimento de teorias a partir da pesquisa baseada em dados, em vez da dedução de hipóteses analisáveis a partir de teorias existentes (CHARMAZ, 2009; STRAUSS; CORBIN, 2008).

A Teoria Fundamentada em Dados é potencialmente utilizada na área de enfermagem, onde se originou. Porém, percebe-se que no Brasil a teoria já vem sendo utilizada em outras áreas tais como: Administração (FARIAS, 2008) e Ciência da Informação (CRESPO; CAREGNATO, 2006; COSTA, 2011; GASQUE, 2008; SIMÕES, 1997; SOARES, 2003).

A *Grounded Theory*, como instrumento de investigação qualitativa, mostra-se adequada para compreender o fenômeno desta pesquisa, pois de acordo Glaser e Strauss (1967, p. viii), essa abordagem é "[...] um método geral de análise comparativa constante". Já Strauss e Corbin (2008, p. 84) argumentam – "[...] fazer comparações é uma característica essencial de nossa metodologia". Por sua vez, Flick (2009) comenta que a codificação dos dados envolve comparações constantes entre fenômenos, casos e conceitos.

Os métodos da Teoria Fundamentada favorecem a exploração das ideias sobre os dados por meio de uma redação analítica já na fase inicial da pesquisa. O pesquisador deve começar pela coleta de dados e concluir com redações de análises e reflexões sobre todo o processo (CHARMAZ, 2009). Nesse aspecto, a pesquisa bibliográfica é realizada à medida que o

pesquisador precisa de apoio para compreender os dados coletados; pois, para Glaser e Strauss (1967), a literatura pode ser utilizada como fonte de análise comparativa entre a teoria gerada pelos dados e as teorias tradicionais já estabelecidas.

Charmaz (2009) defende a coleta de dados relevantes, detalhados e completos, estabelecendo-os em seus contextos situacionais e sociais relevantes. Para tanto, a teoria apresenta três grandes atividades, quais sejam: a coleta de dados, a codificação dos dados e a posterior análise. Porém, merece destaque o fato de que "o processo de pesquisa não é tão linear, os pesquisadores [...] param e escrevem sempre que as ideias lhes ocorrem" (CHARMAZ, 2009, p. 25).

Como método de coleta de dados, a literatura revela que a *Grounded Theory* utiliza como instrumentos: a pesquisa bibliográfica, entrevistas, observação participante, análise textual (textos extraídos, textos existentes), documentos e diário de campo (CHARMAZ, 2009; GLASER; STRAUSS, 1967; STRAUSS, CORBIN, 2008). Neste estudo foram utilizados: a pesquisa bibliográfica, a pesquisa documental (análise textual/ documentos), questionário (entrevistas) e a observação participante da pesquisadora durante a aplicação das entrevistas.

A respeito da codificação de dados, Strauss e Corbin (2008) a consideram uma das principais atividades da Teoria Fundamentada. Charmaz (2009), em consonância com os autores acima citados, comenta – "a codificação gera os ossos da análise [...] representa mais do que um começo; ela define a estrutura analítica a partir da qual [o pesquisador] constrói a análise [...] é o elo fundamental entre a coleta de dados e o desenvolvimento de uma teoria emergente" (CHARMAZ, 2009, p. 70).

O processo de codificação apresenta pequenas variações de acordo com o autor abordado. Glaser e Strauss (1967) em sua proposta original dividiram a codificação em fase inicial e fase focalizada. A codificação inicial "envolve a denominação de cada palavra, linha ou segmento de dado" (CHARMAZ, 2009, p. 72). Já a codificação focalizada "[...] utiliza os códigos iniciais mais significativos ou frequentes para classificar, sintetizar, integrar e organizar grandes quantidades de dados" (CHARMAZ, 2009, p. 72). Um outro nível de codificação apresentado por Glaser e Strauss (1967) é a codificação teórica – "[...] um nível sofisticado de codificação que segue os códigos selecionados [pelo pesquisador] durante a codificação focalizada" (GLASER, 1978 apud CHARMAZ, 2009, p. 94).

Strauss e Corbin (2008) trabalham com a terminologia de codificação aberta, axial e seletiva, que será utilizada neste trabalho e explicada a seguir.

A codificação aberta é definida como "[...] processo analítico por meio do qual os conceitos são identificados e suas propriedades e dimensões são descobertos nos dados"

(STRAUSS; CORBIN, 2008, p. 103). A primeira tarefa da codificação aberta é a *nomeação de conceitos*. Strauss e Corbin (2008) argumentam que para descobrir qualquer coisa nova nos dados e ganhar um melhor entendimento deve-se proceder uma análise detalhada, chamada de *microanálise*. Durante a *microanálise*, cada incidente é comparado a outro incidente no nível de propriedade, em busca de similaridade e diferenças. Os incidentes são posteriormente agrupados em uma *categoria*. Esse processo é registrado pelo pesquisador em *memorandos*[82].

Diante do exposto, entende-se que a partir da *microanálise* é que os conceitos são agrupados em categorias. Strauss e Corbin (2008, p. 114) reforçam que agrupar conceitos em categorias é importante porque permite ao analista reduzir o número de unidades com as quais trabalha.

A última etapa da codificação aberta consiste em desenvolver as categorias em termos de propriedades e dimensões. Strauss e Corbin (2008, p. 117) comentam que é necessário "[...] dar especificidade à categoria por meio da definição de suas características particulares". Nesse processo, o pesquisador diferencia uma categoria de outra e dá a ela precisão.

Finalizada a codificação aberta, dar-se-á início à codificação axial - que tem como objetivo recompor os dados em um todo coerente. É definida como – "[...] o ato de relacionar categorias com subcategorias ao longo das linhas de suas propriedades e suas dimensões" (STRAUSS, CORBIN, 2008, p. 124). O objetivo da codificação axial é começar o processo de reagrupamento dos dados que foram divididos durante a codificação aberta. É importante ressaltar que "sempre começa a surgir na codificação aberta um sentido de como as categorias se relacionam" (STRAUSS; CORBIN, 2008, p. 124).

Para Charmaz (2009, p. 90) "[...] a codificação axial especifica as propriedades e dimensões de uma categoria". Vale ressaltar que enquanto as categorias representam o fenômeno pesquisado, as subcategorias, identificadas na codificação axial, "representam questões sobre o fenômeno, como, por exemplo, quando, onde, por que, quem, como, e com que consequências" (STRAUSS; CORBIN, 2008, p. 214-215).

Para Strauss (1987 apud STRAUSS; CORBIN, 2008, p. 126) as tarefas básicas da codificação axial são:

---

[82] Para Strauss e Corbin (2008) os memorandos podem variar em tipo e formato. Eles devem ser analíticos e conceituais e não descritivos. Podem ter diversos formatos: notas de codificação, notas teóricas, notas operacionais etc.

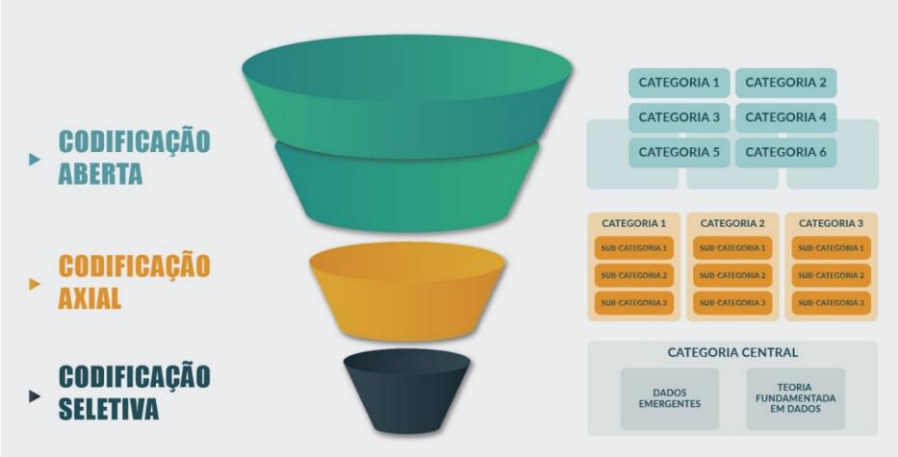
- organizar as propriedades de uma categoria e suas dimensões, uma tarefa que começa na codificação aberta;
- identificar a variedade de condições, ações/interações e consequências associadas a um fenômeno;
- relacionar uma categoria à subcategoria por meio de declarações que denotem como elas se relacionam umas com as outras;
- procurar nos dados pistas que denotem como as principais categorias podem estar relacionadas umas com as outras.

As tarefas acima descritas também são registradas pelos pesquisadores nos memorandos. Outra forma apresentada pelos autores é o uso de *miniestruturas* e *diagramas conceituais*[83], "[...] ambos criados para mostrar a relação entre os conceitos" (STRAUSS; CORBIN, 2008, p. 139).

Uma vez finalizada a codificação axial, terá início a codificação seletiva, cuja definição é "[...] o processo de integrar e refinar a teoria" (STRAUSS; CORBIN, 2008, p. 143). Aqui, o primeiro passo da integração é decidir a categoria central (também chamada de categoria básica) dos dados. Para refinar a teoria, é necessário que o pesquisador reveja o esquema teórico dominante (categoria central e subcategorias) em busca de consistência interna e de falhas na lógica, completando as categorias mal desenvolvidas e podando os excessos. A partir daí se valida o esquema teórico.

A Figura 20 representa o processo de análise dos dados de acordo com a *Grounded Theory*.

Figura 20 – Representação do processo de análise dos dados na Teoria Fundamentada.




Fonte: a autora com fundamento em Strauss e Corbin (2008).

A Figura 20 demonstra que os memorandos são elaborados pelo pesquisador durante todo o processo de codificação dos dados. A respeito do assunto, Charmaz (2009) comenta que:

[83] Para Strauss e Corbin (2008) os diagramas são memorandos visuais e não escritos. São mecanismos que representam a relação entre os conceitos. Os diagramas podem ser feitos manualmente ou usando *softwares* como o ATLAS ou NUD-IST.

> [...] a redação do memorando é a etapa intermediária fundamental entre a coleta de dados e a redação da pesquisa. [...] A redação dos memorandos constitui um método crucial da teoria fundamentada, porque ela incentiva [o pesquisador] a analisar os seus dados e códigos no início do processo da pesquisa (CHARMAZ, 2009, p. 107).

Outro conceito abordado pela Teoria Fundamentada é a *amostragem teórica*. Essa amostragem difere da amostra inicial do estudo, que fornece um ponto de partida para coleta de dados, mas não de refinamento teórico. Charmaz (2009, p. 139) argumenta que "[...] a amostragem inicial na teoria fundamentada é onde você começa, ao passo que a amostragem teórica é o que orienta para onde ir." Consequentemente, os critérios para a amostragem inicial se distinguem daqueles que o pesquisador invoca quando realiza a *amostragem teórica*.

De acordo com Charmaz (2009, p. 140), "[...] o objetivo da amostragem teórica é a obtenção de dados para ajudar [o pesquisador] a explicar suas categorias". A autora vai além ao argumentar que "[...] a amostragem teórica diz respeito apenas ao desenvolvimento conceitual e teórico, ela não tem relação com a representação de uma população ou com a elevação da capacidade de generalização estatística dos seus resultados".

Hood[84] (*apud* CHARMAZ, 2009, p. 141) explica que:

> [...] há uma diferença sutil entre a amostragem teórica e outros tipos de amostragem intencional. A amostragem teórica é uma amostragem intencional, mas o é de acordo com as categorias que alguém desenvolve a partir de suas análises, e essas categorias não são baseadas em cotas, mas sim em preocupações teóricas.

A partir da amostragem teórica o pesquisador obtém a *saturação teórica*, ou seja, o pesquisador chega ao ponto no qual a coleta de dados sobre uma categoria teórica não revela nenhuma propriedade nova nem permite *insights* teóricos novos sobre a teoria emergente.

É justamente a amostragem teórica que faz com que o pesquisador colete dados no início da sua pesquisa. A partir da análise (codificação aberta, axial e seletiva) desses dados que serão desenvolvidas categorias e subcategorias que precisarão ser analisadas em mais profundidade. A análise das categorias, subcategorias e suas relações é possibilitada pela amostragem teórica. Nas palavras de Hood[85](*apud* CHARMAZ, 2009, p. 145) "[...] você recua e avança entre a coleta dos dados e a análise e, à medida que sua teoria se desenvolve pelo método comparativo constante, você fica sabendo, por meio de cada uma das etapas, quais os dados você precisa coletar para refinar a sua teoria".

---

[84] Jane Hood em conversa com Kathy Charmaz. A conversa foi reproduzida no livro de Charmaz (2009).
[85] Jane Hood, em entrevista à Kathy Charmaz, sobre como ela realizou a amostragem teórica.

### 3.2.3 Procedimentos operacionais da pesquisa

A partir dos resultados preliminares da literatura sobre *e-science* nas bases de dados LISTA e LISA, foi constado que os países que mais publicam sobre o tema foram Inglaterra e Estados Unidos, respectivamente. Tal constatação permitiu que fosse priorizado na revisão de literatura os artigos produzidos nesses países.

Em função dessa constatação, a pesquisadora procurou selecionar uma universidade americana que possuísse uma Pós-Graduação em Ciência da Informação atenta aos fenômenos digitais de tratamento, recuperação e preservação da informação, bem como preocupada com o arcabouço que envolve uma política de informação.

Nesse sentido, destacou-se a School of Information – University of Michigan por ser uma escola reconhecida internacionalmente pelo seu pioneirismo nos estudos interdisciplinares entre Ciência da Informação e Tecnologia da Informação. Dentre os professores da escola, destacou-se no âmbito de política de informação o trabalho de Rosenberg (1982a) que pesquisa há, pelo menos 25 anos sobre o tema, com destaque por ter analisado a política de informação de C&T no Brasil na década de 1980. A atuação de Rosenberg é notória e reconhecida internacionalmente, uma vez que dissertou sobre o assunto para o "Annual Review of Information Science and Technology" (ROSENBERG, 1982b) em 1982.

Uma vez identificada a escola e um possível orientador, a pesquisadora iniciou seu contato com o professor Victor Rosenberg na expectativa de que ele a aceitasse como pesquisadora visitante pelo período de agosto de 2014 a abril de 2015. A carta de aceite da University of Michigan foi enviada para a pesquisadora em setembro de 2013. Na sequência, a pesquisadora entrou em contato com a CAPES para participar do Programa de Doutorado Sanduíche (PDSE). A carta de concessão do estágio de doutorando no exterior foi recebida pela pesquisadora em 22 de maio de 2014.

Assim, a segunda fase da pesquisa foi realizada na School of Information da University of Michigan, localizada na cidade de Ann Arbor, no estado de Michigan nos Estados Unidos. Essa etapa compreendeu um período de aproximadamente 4 meses (agosto a novembro de 2014). Dentre as atividades realizadas destacaram-se o incremento da revisão de literatura em função do acesso a livros e artigos de periódicos sobre o tema, disponíveis no acervo da universidade, bem como a realização de entrevistas, conforme ilustra a Figura 21.

Figura 21 – Segunda etapa da pesquisa.

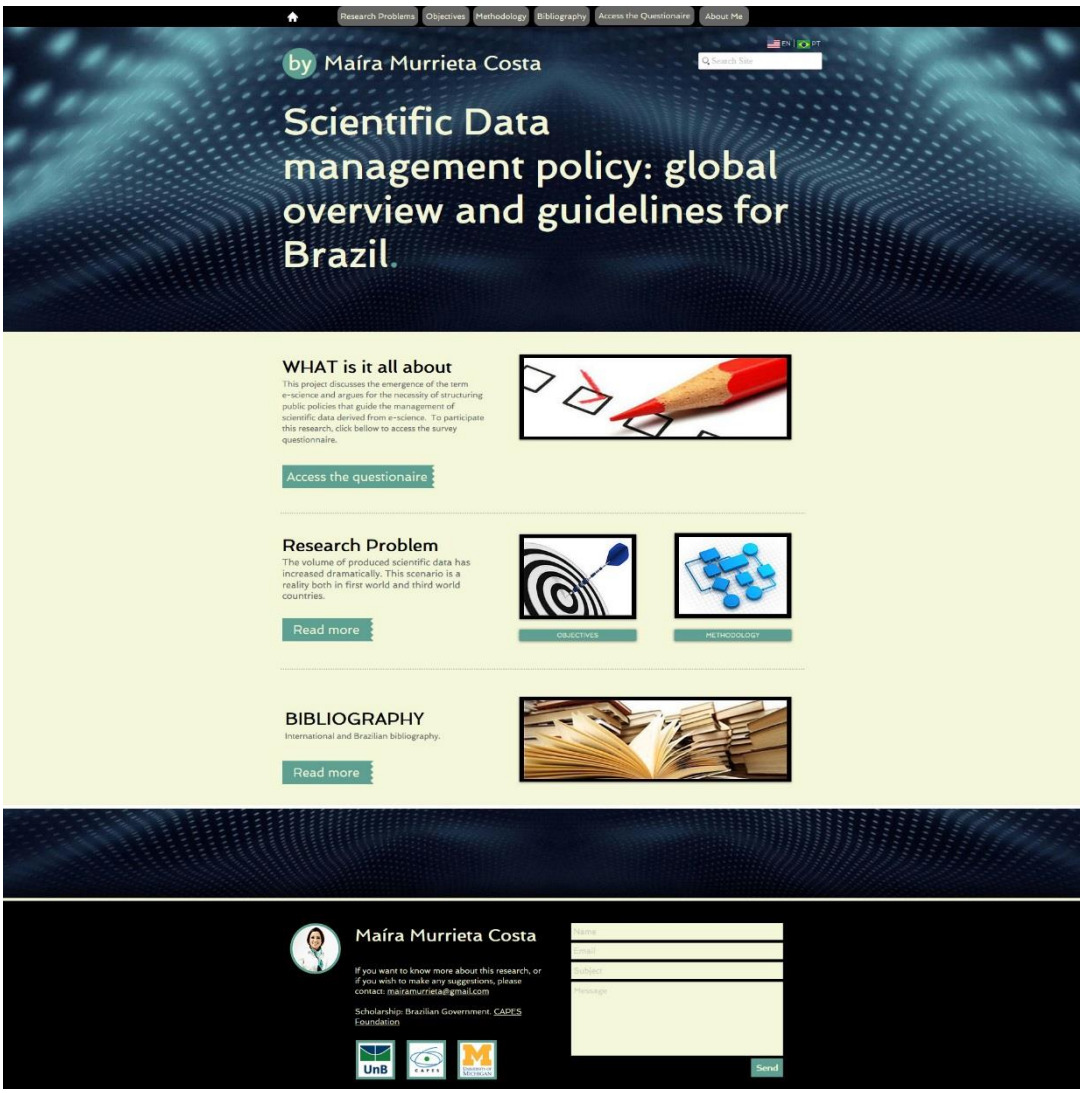

Fonte: a autora.

Com o objetivo de identificar o comportamento dos profissionais e pesquisadores em relação ao compartilhamento de dados científicos, bem como identificar se há nos EUA uma política para a gestão de dados científicos *online*, foi desenvolvido o Questionário 01 (Apêndice 1). Este questionário foi disponibilizado no *blog* sobre a pesquisa[86], para uma amostra intencional de professores da School of Information.

O *blog* da pesquisa foi desenvolvido na plataforma *wix*, com a finalidade de apresentar de forma sucinta os problemas da pesquisa e seus objetivos, a metodologia utilizada e as referências bibliográficas que sustentam a tese, bem como hospedar o questionário *on-line*. Além disso, foi desenvolvido um pequeno texto sobre a pesquisadora e seu vínculo com a Universidade de Michigan, conforme demonstra a Figura 22.

[86] Disponível em: <http://mairam66.wix.com/maira-costa>.

Figura 22 – Home Page do Blog da Pesquisa.




Fonte: a autora.

O questionário, disponível para consulta no Apêndice 1, foi desenvolvido no idioma inglês, utilizando como ferramenta de apoio o Google Forms[87].

O pré-teste do questionário foi iniciado em 08/10/2014 e finalizado após quatro dias, portanto no dia 11/10/2014. Participaram dele um doutor e quatro doutorandos, todos fluentes em inglês, sendo um americano nato, um sul americano e três brasileiros. Eles foram selecionados tanto pela proficiência no idioma, como pelo fato de estarem coletando dados de pesquisa e enfrentando no dia a dia a dificuldade com a gestão desses dados. Porém, apenas um deles tinha conhecimento teórico sobre o problema pesquisado nesta tese. A partir das sugestões o instrumento de coleta de dados sofreu diversas alterações que levaram a concluí-lo apenas em meados de novembro de 2014, perto do final do ano letivo na School of Information. Face ao

[87] Disponível em - https://docs.google.com/forms/d/1I2YltFmL5_kdtfJQiWBRSF8tsVWEYCkDVOvqSfVnl9k/viewform

exposto, optou-se por enviá-lo apenas no ano de 2015, após as festividades natalinas no intuito de obter um maior índice de participação para a pequena amostra ao qual foi aplicado.

A amostra intencional é classificada como não probabilística, selecionada pelo critério de acessibilidade[88] (VERGARA, 2004), composta por sete professores da School of Information – University of Michigan.

Ressalta-se que essa coleta de dados teve um caráter exploratório, seu objetivo foi identificar problemas inerentes a política de gestão de dados em um país de primeiro mundo, de forma a nortear o desenvolvimento de coleta de dados no Brasil. Assim, apesar de ter sido aplicado a amostra pequena, os resultados mostraram-se satisfatórios em função da *expertise* dos respondentes sobre o tema. A partir desses conjuntos de dados iniciais a pesquisadora teve conhecimento do Projeto DataONE e questões intrigantes sobre a gestão de dados no âmbito de dados coletados, por exemplo, pela NASA, o que permitiu um novo aprimoramento do questionário que foi posteriormente aplicado no Brasil.

Em função do exposto, a segunda fase da pesquisa permitiu que o segundo[89] e o terceiro[90] objetivo desta tese fossem alcançados satisfatoriamente. O acesso ao texto integral de artigos e livros relacionados ao tema viabilizou a construção de uma ampla categoria de dados à luz da Grounded Theory. Além disso, o roteiro de entrevista aplicado aos professores da UMICH permitiu o alcance do terceiro objetivo no âmbito dos EUA.

A terceira e última etapa de coleta de dados teve início em 2016. Essa última fase, vinculada aos objetivos específicos 4[91] e 5[92], identificou os atores estratégicos na gestão de dados científicos no Brasil. Para uma melhor compreensão do fenômeno pesquisado, esses atores estratégicos foram divididos em duas categorias: pesquisadores envolvidos com o tema (doutores ou alunos de doutorado) e funcionários de agências de fomento no Brasil. Foi desenvolvido um instrumento de coleta de dados específico para cada categoria, conforme ilustra a Figura 23.

---

[88] Longe de qualquer procedimento estatístico, seleciona elementos pela facilidade de acesso à eles (VERGARA, 2004, p. 51).
[89] **OE 2** – Analisar as ações de governo de países desenvolvidos sobre a gestão de dados científicos nos países identificados.
[90] **OE 3** – Identificar os principais problemas e as soluções inerentes à construção de uma política estruturada para a gestão de dados científicos.
[91] **OE 4** – Identificar a postura das agencias de fomento no Brasil com relação ao tema.
[92] **OE 5** – Identificar o posicionamento dos pesquisadores brasileiros envolvidos com o tema.

Figura 23 – Terceira etapa da pesquisa.


Fonte: a autora.

Ambos os instrumentos de coleta de dados procuraram esmiuçar a problemática da gestão de dados e desenvolver um esboço de diretrizes para a gestão desses dados no país a partir de diferentes perspectivas (funcionários de agências de fomentos *versus* pesquisadores).

A coleta de dados que foi realizada com funcionários das agências de fomento ocorreu entre maio e agosto de 2016, foi totalmente realizada em ambiente *web* por meio do aplicativo SurveyMonkey. O formulário pode ser consultado no Apêndice 2[93]. O SurveyMonkey conferiu agilidade a essa coleta pelo fato de gerenciar os respondentes e emitir alerta de *e-mail* para aqueles que ainda não haviam respondido.

Já a coleta realizada com pesquisadores brasileiros (doutores ou alunos de doutorado) envolvidos com o tema *e-science* ocorreu durante todo o ano de 2016. Essa etapa foi mais longa, pois a pesquisadora inicialmente procurou realizar apenas entrevistas, o que se mostrou extremamente difícil em função da agenda dos selecionados para compor a amostra. Por fim, em função do baixo número de entrevistas realizadas no primeiro semestre de 2016 (apenas 15), a pesquisadora optou por disponibilizar o instrumento de coleta de dados em forma de questionário no Blog da pesquisa e enviar nova carta convite para os selecionados na amostra, esse processo ocorreu no segundo semestre de 2016. O questionário foi desenvolvido no GoogleForms, conforme Apêndice 3. A ferramenta disponibilizada pela empresa Google mostrou-se com desenvolvimento satisfatório, mas com limitações quando comparada ao SurveyMonkey.

[93] Disponível em - <https://pt.surveymonkey.com/r/PW8FFP5>.

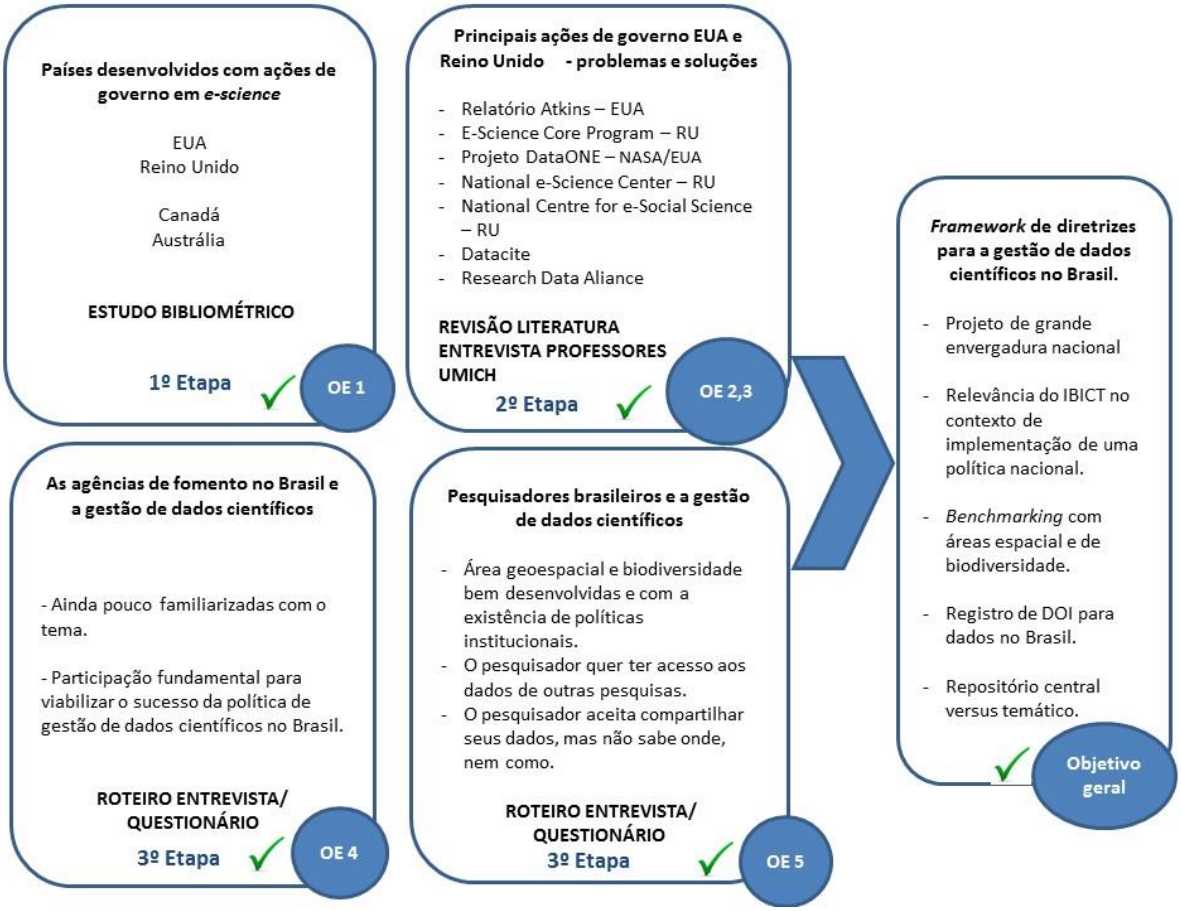
As três etapas de pesquisa são complementares (*análise bibliométrica*, *análise da literatura sobre o tema de forma descritiva*, *pesquisa de campo*), formam uma triangulação de técnicas que teve como objetivo permitir que a pesquisadora construísse uma visão holística do fenômeno, o que certamente contribuiu para uma melhor compreensão do tema estudado. As etapas de pesquisa são sintetizadas na Figura 24.

Figura 24 – Etapas da pesquisa.


Fonte: a autora.

Esta tese se propôs a desenvolver um conjunto de diretrizes para a elaboração de uma política nacional para a gestão de dados científicos no Brasil. Para alcançar esse objetivo geral, foram delimitados cinco objetivos específicos, alcançados por meio diferentes técnicas de pesquisa, já relacionadas durante este capítulo referente a metodologia. A síntese dos objetivos de pesquisa e o instrumento de coleta de dados pode ser observada na Figura 25.

Figura 25 – Visão geral dos objetivos específicos da pesquisa *versus* instrumento de coleta de dados




Fonte: a autora.

### 3.2.4 Definições Operacionais

Nesta pesquisa foram utilizados os termos técnicos abaixo descritos.

**Análise Bibliométrica:** uma técnica quantitativa e estatística de medição de índices de produção da literatura em determinada área do conhecimento, tem como objetivo conhecer a natureza específica de cada grupo produtor do conhecimento. Possui três leis clássicas: a Lei de Lotka, a Lei de Bradford e a Lei de Zipf.

***Big data*:** termo referente à grande quantidade de dados não estruturados, que atualmente são produzidos e disponibilizados em rede móveis e computação em nuvem.

**Ciclo de Vida dos dados:** começa quando o pesquisador ainda está planejando sua etapa de coleta de dados. Os próximos três estágios (coletar, validar, descrever) são a base para o acesso de longo prazo ao dado. Enquanto isso, os três últimos estágios (preservar, descobrir e integrar) representam a descoberta e o uso dos dados.

**Curadoria de dados:** é um serviço que envolve todos os processos aplicados a objetos digitais durante o seu ciclo de vida. Compreende as atividades de estabelecimento de padrões para conjunto de dados, adição de valor, gestão de risco e boas práticas na gestão de dados digitais. Trata-se de um conceito mais inclusivo que o arquivamento digital e a preservação digital.

**Dados abertos do governo:** são dados coletados pelo governo que devem estar acessíveis ao público de forma a garantir os princípios da publicidade e da transparência da administração pública. Dentre os principais sistemas do Governo Federal que possuem dados abertos merecem destaque o Sistema Integrado de Planejamento e Orçamento (SIOP); Sistema de Informações das Estatais – (SIEST); Sistema Integrado de Administração de Pessoas (SIAPE); Sistema de Informações Organizacionais – (SIORG); Sistema Integrado de Administração de Serviços Gerais (SIASG); Sistema de Gestão de Convênios e Contratos de Repasses (SICONV); Sistema Integrado de Administração Patrimonial (SIAPA). O foco do programa de dados abertos do governo está é a prestação de contas de ações governamentais para com a sociedade, seguindo, portanto, os princípios da transparência pública.

***e-science*:** é um conceito que se refere à coleção de instrumentos e tecnologias necessárias para apoiar a pesquisa científica do Século XXI – intrínseca à natureza colaborativa e multidisciplinar, bem como ao grande volume de dados produzidos que precisam estar disponibilizados em rede. Em outras palavras, é a ciência trabalhando com o apoio da tecnologia da informação e comunicação. Também é chamado de ciência orientada a dados, computação fortemente orientada a dados, ciberinfraestrutura, dados científicos ou quarto paradigma.

**Gestão de dados científicos:** conjunto de atividades intrínsecas ao processo de tratamento técnico (curadoria), armazenamento, recuperação, disseminação e preservação dos dados coletados pela *e-science*.

**Metadados:** São dados que descrevem outros dados de um determinado objeto. Continuam atualizados durante toda a vida da informação e se relacionam com a descrição de mais de um objeto. Não precisam, necessariamente, ser digitais.

**Políticas públicas:** Conjunto de ações do governo que irão produzir efeitos específicos na vida do cidadão, bem como na economia de um país. Após formuladas, desdobram-se em planos, programas e projetos que serão submetidos a sistemas de acompanhamento e avaliação.

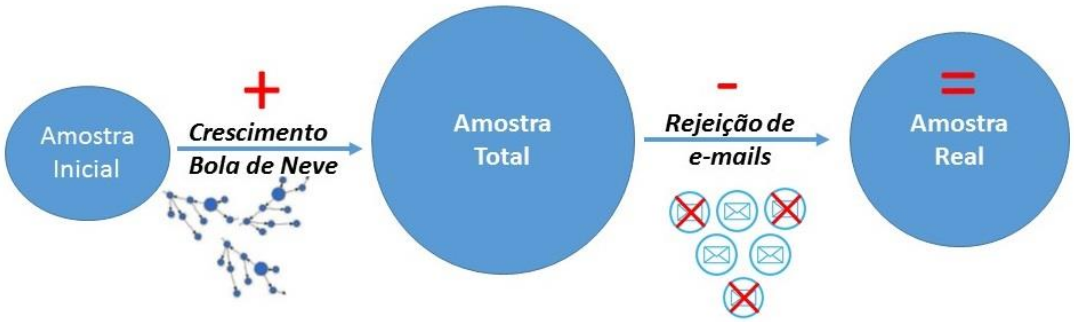
### 3.2.5 Universo e Amostra

Este tópico refere-se à amostra delineada para a terceira etapa desta pesquisa. Conforme relatado no tópico anterior, esta amostra dividiu-se em duas categorias (*funcionários de agências de fomento* e *pesquisadores sobre o tema*). Ambas se classificam como não probabilística, formadas pelo critério de tipicidade (VERGARA, 2004), com crescimento em bola de neve.

Ambas as amostras utilizaram a técnica de bola de neve com o objetivo de incrementar o número de participantes identificados na amostra inicial, conforme ilustra a Figura 26.

A base empírica da pesquisa contou com uma amostra inicial e o crescimento da amostra pela técnica bola de neve. Porém há que se ressaltar que o instrumento de coleta de dados foi enviado por *e-mail*, ou seja, há uma taxa de retorno em função de mudança de endereço, ou mesmo rejeição do *e-mail* em função de regras do provedor do serviço, ou mesmo pelo usuário. Assim, nesta pesquisa surgiu a figura da amostra real, formada pela equação amostra inicial + crescimento bola de neve = Amostra Total. Como a amostra total sobre um índice de perda em função de rejeição do *e-mail*, tem-se a Amostra Real, conforme Figura 26.

Figura 26 – Amostra Real.




Fonte: a autora.

#### *3.2.5.1 – Amostra dos pesquisadores envolvidos com o tema e-science*

A categoria "*Pesquisadores sobre o tema*" teve como critérios de amostra ser doutor ou aluno de doutorado de qualquer área do conhecimento; ter tido contato com grandes quantidades de dados, bem como a necessidade de gerenciar esses dados; ter conhecimento sobre o movimento de *e-science*. Para identificar essa amostra optou-se por fazer um levantamento nas bases de dados do Currículo Lattes, no Diretório do Grupo de Pesquisas do CNPq, em diretórios de eventos relacionados a *e-science*. Em todos os *sites* foram utilizados

como argumento de pesquisa os termos: *e-science, cyberinfrastructure, data deluge, data driven science, data curation, data scientist* – no intuito de aumentar a revocação de um número de pessoas que permitisse enviar o questionário, ainda que diminuísse a precisão. Optou-se por essa forma, pois um doutor que não dominasse o tema, certamente não iria responder o questionário.

A partir dos critérios estabelecidos formou-se uma amostra inicial que contou com 100 pesquisadores. Considerando a técnica de bola de neve e o índice de rejeição de *e-mails* formou-se uma amostra real de 111 pesquisadores.

Para localizar os *e-mails* dessas pessoas novamente foram feitas buscas no Google, no Google Scholar – sendo o argumento de busca ➔ "NOMECOMPLETO" + EMAIL. Em alguns casos, quando o nome da pessoa era mais comum, o argumento de busca foi ➔ "NOME COMPLETO" + EMAIL + INSTITUIÇÃO.

A recuperação dos *e-mails* foi alta, pois, ainda que, em um primeiro momento, não se localize no Google, o Google Scholar traz artigos publicados pelos doutores e ou alunos de doutorado. Certamente recuperar tais *e-mails* foi um trabalho operacional árduo, mas viável.

Um fato observado na pesquisa e que dificultou o levantamento de *e-mails*, é que os *sites* na *web*, para evitar a proliferação dos *spams* nos endereços de *e-mail* tem divulgado o endereço eletrônico da seguinte maneira: nome.sobrenome[at]gmail[dot]com[dot]br, como por exemplo ➔ mairamurrieta[at]gmail[dot]com.

*3.2.5.2 – Amostra dos funcionários de agências de fomento*

Já a categoria "*Funcionários de agência de fomento*" teve como critérios de amostra ser servidor de agência de fomento ou fundação de amparo a pesquisa, ou ser consultor, membro de conselho diretor ou consultivo das agências de fomento ou fundações de amparo à pesquisa no Brasil. Para identificar essa amostra optou-se por fazer um levantamento de *e-mails* nos *sites* das agências de fomento (CAPES, CNPq) e nos *sites* das Fundações de Amparo à Pesquisa. A partir dos critérios estabelecidos formou-se uma amostra inicial que contou com 151 pesquisadores. Considerando a técnica de bola de neve e o índice de rejeição de *e-mails* formou-se uma amostra real de 169 pesquisadores.

**3.2.6 Procedimentos de Coleta de Dados**

Os procedimentos de coleta de dados selecionados para compor a pesquisa exploratória priorizaram a coleta de dados qualitativa, mas quando se julgou necessário, dados quantitativos

também foram coletados. Assim, cada objetivo específico necessitou de instrumentos adequados, conforme descrito no Quadro 11.

Quadro 11 – Objetivos da pesquisa *versus* os instrumentos de coleta de dados.

| Objetivo Específico | Instrumento de Coleta de Dados | Fonte de Coleta de Dados |
|---|---|---|
| **OE 1** – Identificar os países desenvolvidos que possuem ações de governo para a gestão de dados científicos. | Estudo bibliométrico nas bases de dados LISA e LISTA | Metadados dos registros bibliográficos das bases de dados LISA e LISTA. |
| **OE 2** – Analisar as ações de governo de países desenvolvidos sobre a gestão de dados científicos nos países identificados. | Análise da literatura sobre o tema a partir do resultado do estudo bibliométrico. | • Bases de dados LISA, LISTA.<br>• OPAC[94] da University of Michigan Library |
| **OE 3** – Identificar os principais problemas e as soluções inerentes à construção de uma política estruturada para a gestão de dados científicos. | Análise da literatura sobre o tema a partir do resultado do estudo bibliométrico. | • Bases de dados LISA, LISTA.<br>• OPAC da University of Michigan Library |
| **OE 4** – Identificar a postura das agências de fomento no Brasil com relação ao tema. | Entrevista com servidores de agências de fomento e fundações de amparo à pesquisa no Brasil | • Amostra intencional de servidores das Agências de Fomento e Fundações de Amparo à Pesquisa |
| **OE 5** – Identificar o posicionamento dos pesquisadores brasileiros envolvidos com o tema. | Entrevista com doutores e ou alunos de doutorado | • Amostra intencional composta por doutores ou alunos de doutorado que tenham conhecimento da e-science e necessidade de gerenciar dados científicos. |

Fonte: a autora.

### 3.2.7 Formulários de Coleta de Dados

Para atender o OE[95] 4 foi desenvolvido o formulário de coleta de dados que consta no Apêndice 2[96]. Este instrumento foi aplicado aos funcionários de agências de fomento e fundações de amparo à pesquisa no Brasil. Já para atender o OE[97] 5 foi desenvolvido o formulário de coleta de dados que consta no Apêndice 3[98]. Este instrumento foi aplicado com os doutores envolvidos com o tema *e-science*.

A coleta de dados que foi realizada com funcionários das agências de fomento (OE 4|) ocorreu entre maio e agosto de 2016, foi totalmente realizada em ambiente *web* por meio do aplicativo SurveyMonkey. O SurveyMonkey conferiu agilidade a essa coleta pelo fato de

[94] *Online public access catalog* /catálogo público de acesso em linha.
[95] Identificar a postura das agencias de fomento no Brasil com relação ao tema
[96] Disponível online em https://pt.surveymonkey.com/r/PW8FFP5
[97] Identificar o posicionamento dos pesquisadores brasileiros envolvidos com o tema.
[98] Disponível online em https://pt.surveymonkey.com/r/KYR828H

gerenciar os respondentes e emitir alerta de *e-mail* para aqueles que ainda não haviam respondido. Considerando a amostra total desta categoria, o convite para participar da pesquisa foi enviado para 183 pesquisadores. Sendo que destes, 12 *e-mails* retornaram e outros dois tiveram o recebimento cancelado pelo provedor de *e-mail* ou pelo usuário, assim, a amostra real desenhou-se em 169 funcionários de agências de fomento.

Já a coleta realizada com doutores ou alunos de doutorado envolvidos com o tema *e-science* ocorreu durante todo o ano de 2016. Essa etapa foi mais longa, pois a pesquisadora inicialmente procurou realizar apenas entrevistas, o que se mostrou extremamente difícil em função da agenda dos selecionados para compor a amostra. Por fim, em função do baixo número de entrevistas realizadas no primeiro semestre de 2016, a pesquisadora optou por disponibilizar o instrumento de coleta de dados em forma de questionário no Blog da pesquisa e enviar nova carta convite para os selecionados na amostra, esse processo ocorreu no segundo semestre de 2016.

Assim, o universo e a amostra desta fase constituíram-se conforme descrito no Quadro 12.

Quadro 12 – Composição da amostra intencional da pesquisa.

| | **Universo** | **Amostra inicial** | **Bola de neve** | **Amostra total** | **Índice rejeição** | **Amostra real** | **Qtd de respostas** |
|---|---|---|---|---|---|---|---|
| **Questionário 1 – Doutores ou alunos de doutorado** | Todos os doutores ou alunos de doutorados de universidades brasileiras | 100 | 36 | 136 | 25 | 111 | 40 (36%) |
| **Questionário 2 – Agências de Fomento** | Funcionários de agência de fomento brasileira, incluindo os representantes de conselhos consultivos das agências | 151 | 32 | 183 | 14 | 169 | 22 13% respostas |

Fonte: a autora.

## 3.3 LIMITAÇÕES DA TESE

Esta tese não teve como objetivo trabalhar os dados abertos do governo, seu objetivo foi direcionado para dados científicos, ou seja, produzidos por pesquisas científicas. Assim, apesar dos dados abertos do governo serem, muitas vezes, fonte de dados secundários para pesquisas científicas, eles são coletados por entidades governamentais com o objetivo de avaliar programas de governos, gerar indicadores governamentais e prestar contas com a sociedade, portanto, fogem ao escopo da tese – qual seja, trabalhar dados oriundos da *e-science*, objeto da pesquisa colaborativa do Século XXI, coletados por sensores especializados e em grande escala (*big data*).

# 4 ANÁLISE DE DADOS

Neste capítulo são apresentados os resultados da terceira etapa da pesquisa que procurou mapear a situação da gestão de dados de pesquisa no Brasil, a qual corresponde aos objetivos específicos 4[99] e 5[100], respectivamente. Ressalta-se que os objetivos específicos 1, 2 e 3 tiveram seus resultados alcançados e apresentados na análise bibliométrica sobre o estudo do termo *e-science* nas bases de dados LISA e LISTA, bem como a segunda etapa da pesquisa que foi realizada na School of Information da University of Michigan, conforme já descrito no Capítulo 3.2.3 – referente aos procedimentos operacionais da pesquisa.

Esta etapa da pesquisa foi conduzida com dois grupos de amostra, ambos selecionados de forma não probabilística e intencional. Cabe ressaltar que a compreensão do fenômeno exigiu que fosse observado o comportamento tanto de pesquisadores brasileiros envolvidos com o assunto, quanto de funcionários de agências de fomento e fundações de amparo à pesquisa no Brasil. Ao considerar-se o exposto, os dados obtidos nessa etapa indicam tendências e, uma vez que a amostra não foi probabilística, estes dados não devem ser extrapolados.

Os resultados são apresentados por grupo de entrevistados. No Capítulo 4.1 consta a análise dos dados referentes aos pesquisadores doutores envolvidos com questões inerentes aos dados científicos no Brasil, seguindo-se o Capítulo 4.2 referente à análise dos dados relativos às Agências de Fomento. No Capítulo 4.3 é apresentada a análise sobre o entendimento do termo curadoria de dados e gestão dos dados, bem como sobre a necessidade de uma política nacional de gestão de dados científicos. Ao final, no Capítulo 4.4, é apresentada a teoria fundamentada em dados, sintetizada na proposta de um *framework* que contém diretrizes para a elaboração de uma política de gestão de dados científicos. Esse *framework* foi elaborado a partir da reflexão da codificação das respostas qualitativas de ambos os instrumentos de coleta de dados, bem como da compreensão da literatura revisada sobre a gestão de dados científicos. A Figura 27 ilustra os diferentes instrumentos de pesquisa, com o número de respondentes para cada um deles.

[99] **OE 4** – Identificar a postura das agências de fomento no Brasil com relação ao tema.
[100] **OE 5** – Identificar o posicionamento dos pesquisadores brasileiros envolvidos com o tema.

Figura 27 – Instrumento de coleta de dados *versus* quantidade de resposta.


Fonte: a autora.

Para apoiar a análise qualitativa dos dados foi utilizado o *software* Nvivo versão 10. O objetivo de se utilizar o *software* foi construir as categorias de informações à luz da Teoria Fundamentada em Dados. Para tanto, as entrevistas dos doutores, bem como as respostas dos questionários das agências de fomento foram transcritas e posteriormente alimentadas em uma planilha em Excel. Essa planilha foi importada para o Nvivo10 com todas as respostas tanto das entrevistas, como dos questionários.

O *software* se mostrou de difícil utilização, pois a versão 10 é pouco amigável em relação à versão 11, além de não apresentar as mesmas opções de geração de gráficos, análise de *cluster*, construção de mapas mentais dentre outros recursos de análise. Assim, o uso do Nvivo10 nesta tese se restringiu à construção das categorias de informação no que o *software* denomina Fonte Interna de Dados, ou seja, na base empírica da pesquisa – entrevistas e questionários.

Há que se ressaltar que apesar do *software* vender uma facilidade no processo de codificação dos dados, o fato é que a codificação só ocorre quando o pesquisador consegue alcançar familiaridade com a sua massa de dados. E a familiaridade, por sua vez, é alcançada por meio da reflexão, independentemente do uso de *softwares* ou não. Nesse sentido, tudo que foi feito nesta tese com o uso do Nvivo10 poderia ter sido feito manualmente. Cabe ressaltar que a dificuldade de gerar gráficos levou a pesquisadora a exportar suas categorias de informação para o SPSS de forma a viabilizar a análise da frequência das categorias por resposta qualitativa, o que gerou uma estatística descritiva.

A experiência no uso do Nvivo10 permite afirmar que seu potencial se dá quando o pesquisador inicia sua utilização na categorização de fontes externas (literatura, documentos, etc.) e durante sua análise de dados na categorização de fontes internas, o que permitiu uma comparação entre ambas.

Nesta tese, as poucas questões que foram analisadas exclusivamente no Nvivo10 foram as referentes ao entendimento do conceito de curadoria de dados e gestão de dados científicos. Além dessas, analisou-se ainda a questão referente ao delineamento das características e do perfil profissional do cientista de dados. Porém, as saídas gráficas não se mostraram visualmente harmônicas, o que levou a pesquisadora a procurar o auxílio de um *designer* para redesenhar os gráficos com fundamento nos dados gerados pelo Nvivo.

Além do Nvivo10, foram utilizados nesta tese os *softwares* SPSS e Excel no processo de geração de análises descritivas e geração de gráficos. O SPSS se mostrou um facilitador no cruzamento de dados dos respondentes e permitiu a análise entre a área de formação do pesquisador, geração à qual pertence, *versus* comportamento das respostas, o que permitiu algumas inferências por parte da pesquisadora.

## 4.1 ANÁLISE DOS DADOS REFERENTES AOS PESQUISADORES DOUTORES ENVOLVIDOS COM QUESTÕES INERENTES AOS DADOS CIENTÍFICOS NO BRASIL

A primeira parte da entrevista procurou delinear um perfil do respondente. Assim, as primeiras perguntas, até mesmo para deixar o entrevistado mais à vontade, referiam-se aos dados demográficos e continham as seguintes variáveis: ano de nascimento, instituição, tempo de trabalho na instituição, função na instituição, área de formação e área de pesquisa. Por meio dessas perguntas o entrevistado se apresentava de uma forma geral e, ao mesmo tempo, a pesquisadora explicava o contexto do instrumento de coleta de dados e contextualizava a pesquisa.

O maior número de respondentes foi da área de Ciências Sociais Aplicadas, seguido de Ciências Exatas e da Terra (Moda = 3). Entre os respondentes, a média de trabalho na instituição é de 12 anos, fato que revela maturidade dos respondentes. O maior número de entrevistas se encontra na USP, UnB e IEN – o que reflete a proximidade da pesquisadora com a Universidade de Brasília, bem como as visitas técnicas realizadas em 2016 à USP/SP e ao IEN/RJ, conforme revelam os dados da Tabela 6.

Tabela 6 – Número de participantes da pesquisa por instituição *versus* área atuação– Brasil.

| | | Área de pesquisa/atuação | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ciências Exatas /Terra | Ciência Computação | Ciências Sociais Aplicadas | Eng.. | Ciências Humanas | Ciências Biológicas | Ciências Saúde | Ciências Agrárias | Mult. | |
| ANTAQ | Cont. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| UCB | Cont. | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 |
| UFMG | Cont. | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| INPA | Cont. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| IBICT | Cont. | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 3 |
| USP | Cont. | 1 | 0 | 0 | 2 | 1 | 1 | 0 | 0 | 0 | 5 |
| UnB | Cont. | 0 | 0 | 4 | 0 | 0 | 0 | 1 | 0 | 0 | 5 |
| CGEE | Cont. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| UFJF | Cont. | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| UFG | Cont. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| UFF | Cont. | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| UFSC | Cont. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| UFPB | Cont. | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| SENADO | Cont. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| EMBRAPA | Cont. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| EMILIO GOELDI | Cont. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| IEN | Cont. | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 4 |
| ICMBIO | Cont. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| ELSEVIER | Cont. | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| UFRGS | Cont. | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **TOTAL** | **Cont.** | **12** | **1** | **13** | **5** | **4** | **1** | **1** | **2** | **1** | **40** |
| | **% Total** | **30,0%** | **2,5%** | **32,5%** | **12,5%** | **10,0%** | **2,5%** | **2,5%** | **5,0%** | **2,5%** | **100,0%** |

Fonte: a autora.

Dentre os respondentes, 47,5% pertencem à Geração X[101], seguido da *Baby Boomer*[102] o que corrobora o tempo de trabalho na instituição e a participação de pessoas maduras profissionalmente nesta pesquisa, conforme dados apresentados na Tabela 7. Além disso, mostra que nas instituições, ainda não houve a transição para pesquisadores da Geração Y[103] na amostra da pesquisa. Para esta análise, o coeficiente de correlação de Pearson foi muito baixo (0,023), portanto não se pode afirmar que existe uma relação entre essas duas variáveis para essa amostra de pesquisa.

[101] Nascidos entre 1965-1980. A idade dessa geração varia de 31 a 45 anos. São considerados pioneiros no domínio dos computadores, adotam a tecnologia e a internet como uma maneira de controle das suas vidas.
[102] Nascidos entre 1946 – 1964. A idade dessa geração varia entre 45 a 64 anos. Eles têm familiaridade inconsciente com a *web* 2.0, dependendo das necessidades do trabalho do indivíduo e suas preferências pessoais.
[103] Nascidos após 1980. Essa geração possui menos de 30 anos. São completamente envolvidos com a tecnologia da internet. São descritos como impacientes, destemidos, talentosos e com baixa tolerância a críticas.

Tabela 7 – Área de pesquisa *versus* ano de nascimento – Brasil.

| | | Baby Boomer - (1951 - 1964) | X - (1965 - 1980) | Y - (1981 - 1991) | Total |
|---|---|---|---|---|---|
| Ciências Exatas e da Terra | Cont. | 5 | 4 | 3 | 12 |
| | % Total | 12,5% | 10,0% | 7,5% | 30,0% |
| Ciência da Computação | Cont. | 0 | 1 | 0 | 1 |
| | % Total | 0,0% | 2,5% | 0,0% | 2,5% |
| Ciências Sociais Aplicadas | Cont. | 4 | 7 | 2 | 13 |
| | % Total | 10,0% | 17,5% | 5,0% | 32,5% |
| Engenharias | Cont. | 1 | 2 | 2 | 5 |
| | % Total | 2,5% | 5,0% | 5,0% | 12,5% |
| Ciências Humanas | Cont. | 2 | 2 | 0 | 4 |
| | % Total | 5,0% | 5,0% | 0,0% | 10,0% |
| Ciências Biológicas | Cont. | 0 | 1 | 0 | 1 |
| | % Total | 0,0% | 2,5% | 0,0% | 2,5% |
| Ciências da Saúde | Cont. | 0 | 0 | 1 | 1 |
| | % Total | 0,0% | 0,0% | 2,5% | 2,5% |
| Ciências Agrárias | Cont. | 1 | 1 | 0 | 2 |
| | % Total | 2,5% | 2,5% | 0,0% | 5,0% |
| Multidisciplinar | Cont. | 0 | 1 | 0 | 1 |
| | % Total | 0,0% | 2,5% | 0,0% | 2,5% |
| **Total** | **Conta.** | **13** | **19** | **8** | **40** |
| | **% Total** | **32,5%** | **47,5%** | **20,0%** | **100,0%** |

Fonte: a autora.

Aproximadamente 50% dos respondentes são professores pesquisadores. Seguidos por alunos de doutorado (20%). Dentre os participantes, 37,5% dizem que o tipo de dado que produzem é de observação, conforme dados apresentados na Tabela 8. O que se mostra coerente (*correlação*) com o número de participantes das Ciências Sociais Aplicadas.

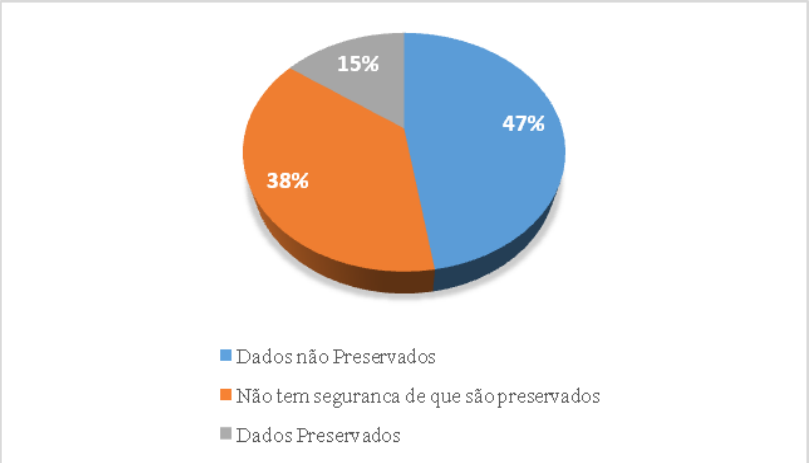
Tabela 8 – Tipo de dado produzido pelo pesquisador – Brasil.

| | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| Observação | 15 | 37,5 | 37,5 | 37,5 |
| Simulação | 3 | 7,5 | 7,5 | 45,0 |
| Produzidos em laboratórios | 3 | 7,5 | 7,5 | 52,5 |
| Coletados automaticamente | 1 | 2,5 | 2,5 | 55,0 |
| Abertos do governo | 6 | 15,0 | 15,0 | 70,0 |
| Observação e dos abertos do governo | 6 | 15,0 | 15,0 | 85,0 |
| Simulação e coletados automaticamente | 3 | 7,5 | 7,5 | 92,5 |
| Observação e coletado automaticamente | 1 | 2,5 | 2,5 | 95,0 |
| Simulação e laboratório | 1 | 2,5 | 2,5 | 97,5 |
| Observação e simulação | 1 | 2,5 | 2,5 | 100,0 |
| **Total** | **40** | **100,0** | **100,0** | |

Fonte: a autora.

No que diz respeito à preservação de dados de suas respectivas pesquisas (P2[104]), do total de respondentes, apenas 15% defendem que seus dados de pesquisa são preservados. Merece ser destacado que aproximadamente 37% dos respondentes não possuem segurança em afirmar que seus dados de pesquisa estão preservados, conforme revelam os dados do Gráfico 7.

Gráfico 7 – Preservação dos dados produzidos pela pesquisa – Brasil.




Fonte: a autora.

Pela análise das respostas qualitativas referentes à P9, percebeu-se que quinze participantes deixaram a opção de resposta em branco. Dentre os que responderam, quatro respostas foram categorizadas como desconexas. Por outro lado, outros seis participantes externalizaram coerentemente o que corresponde a um sistema de administração de dados brutos de pesquisa. Os sistemas relatados foram: Specify, desenvolvido pela Universidade do Kansas; Moporã – sistema desenvolvido pela instituição; GITHUB; GITLAB; DSpace; SISBIO – desenvolvido internamente pela instituição; Lime Survey, Sistema Interno da Elsevier, conforme demonstra o Quadro 13.

[104] a pergunta foi realizada em escala de três pontos e os participantes tinham como opção de resposta uma barra de rolagem com as opções <discordo>, <indiferente> e <concordo>. O primeiro ponto <discordo> correspondia ao intervalo de 1 a 49 na barra de rolagem, enquanto a opção <indiferente> foi representada pelo de 45 a 55 e a opção concordo pelo intervalo de 56 a 100.

Quadro 13 – Sistema de administração dos dados brutos da pesquisa – Brasil.

| Perfil Entrevistado | Resposta | Códigos |
|---|---|---|
| INPA | ***Specify*** *- Sistema de busca internacional desenvolvido pela University of Kansas* ***Moporã*** *- sistema desenvolvido internamente* | ***Preservação de dados***<br>• Specify – Kansas<br>• Moporã – INPA |
| USP | *Os códigos fontes e artefatos gerados são armazenados em plataformas GIT, como o* ***GITHUB*** *e o* ***GITLAB****, e podem ser recuperados pelo site.* | ***Preservação de dados***<br>• Plataforma GIT<br>• GITUB<br>• GITLAB |
| IEN | *O* ***DSpace*** *recupera pelo título do dado, palavras –chave. É um sistema de busca institucional para a produção científica (não apenas para dados). Porém os depósitos dos dados brutos ainda não são obrigatórios. O que se tem de obrigatoriedade no IEN hoje é o depósito das teses e dissertações produzidas pela instituição. Estamos colocando a produção de artigos e periódicos e congressos, mas ainda não há uma política institucional que discipline o depósito dos dados brutos. Os pesquisadores se quiserem depositar os dados, eles já sabem que existe a plataforma para tal finalidade.* | ***Preservação de dados***<br>• DSpace<br>• Repositório |
| ICMBIO | *Sim, para os dados que foram tratados pelo* ***SISBIO****, é possível acessa-los pelo Portal da Biodiversidade. (...) Por outro lado, os dados armazenados localmente (em hardwares locais), não contam com sistema para recuperação dos dados neles armazenados.* | ***Preservação de dados***<br>• SISBIO |
| ELSEVIER | ***A Elsevier tem um sistema de busca que recupera dados brutos associados a artigos de journals que ela edita***. Ela já possui 5 ou 6 iniciativas nesse sentido. ***As mais conhecidas hoje são soluções que anexam dados aos artigos e aí eles podem ser indexados com os próprios artigos e existe um produto importante dentro da linha que eu trabalho que é o PURE, que consiste numa ferramenta de desenvolvimento de websites institucionais****. Há a possibilidade de agregar e estes websites institucionais os dados dos pesquisadores que estão com suas páginas ali naquele portal, semelhante a um grande lattes em nível institucional, em que estão ali todas as publicações e também os dados. Pode ser acessado internamente e também externamente. A Elsevier já possui produtos em seu portfólio e está desenvolvendo novas tecnologias. As soluções que estão disponíveis não seriam necessariamente produtos, mas sim um benefício para você ter seu periódico editado na Elsevier.* ***Se você tiver seu periódico editado na Elsevier, os seus autores vão poder anexar os dados junto com seus artigos. São recursos para atrair editores a publicar sua revisa, seu periódico com a Elsevier.*** | ***Preservação de dados***<br>• PURE – Elsevier<br>• Atrair editores<br>• Atrair autores |
| EMBRAPA | *Sim, nós temos acesso aos dados brutos. Aqui utilizamos o* ***Lime Survey****. Na nossa organização, esta ferramenta é utilizada de forma institucional. [...] Porém, eu não tenho acesso aos dados de pesquisa de outra pessoa. Eu tenho os dados de pesquisa apenas com os quais trabalho.* | ***Preservação de dados***<br>• LimeSurvey |
| INSTITUTO DE PESQUISA DO MCTIC | *É um repositório de dados de pesquisa* | ***Preservação de dados***<br>• Repositório |

Fonte: a autora.

Outros 10 participantes se manifestaram qualitativamente, mas as suas respostas não indicaram precisamente a existência de um sistema para recuperação de dados brutos. Por exemplo, alguns pesquisadores responderam que o Google recupera seus dados, outros que o sistema de busca institucional recupera dados, ou ainda, citaram sistema de administração de dados estatísticos, conforme relatos apresentados no Quadro 14.

Quadro 14 – Percepção equivocada sobre a existência de sistema de administração de dados brutos da pesquisa – Brasil.

| Perfil Entrevistado | Resposta | Códigos |
|---|---|---|
| GOVERNO FEDERAL | *Isolada, se utiliza o Statistic* | ***Sistema adm. dados estatísticos***<br>• *Statistic* |
| UNIVERSIDADE FEDERAL | *Olha a gente trabalha com pesquisa de recuperação de informação então a gente tem alguns alunos que desenvolvem modelos e protótipos para recuperar dados. Mas de uma forma geral, não temos um sistema genérico que recupere os dados das pesquisas desenvolvidas na Universidade. Isso, de fato, nós não temos. / A comunicação científica está muito bem elaborada em termos de sistemas. Vemos isso por exemplo no sistema de busca de periódicos nas bibliotecas digitais de teses e dissertações, mas não acompanhamos a mesma evolução com o armazenamento dos dados brutos da pesquisa nos sistemas.* | ***Ausência de um sistema para recuperação de dados brutos***<br>• *Não há um sistema genérico para a universidade*<br>• *Iniciativas isoladas* |
| UNIVERSIDADE FEDERAL | *Imagino que sim: os buscadores automáticos, como o Google, que utilizo para salvar meus dados.* | ***Ausência de um sistema para recuperação de dados brutos***<br>• Mecanismo de busca - Google |
| UNIVERSIDADE FEDERAL | *No momento não há forma de recuperação sistematizada. Tenho usado o DSpace para alguns grupos de dados e penso que vá continuar usando por um bom tempo. Os demais ficarão na nuvem (isso ainda não está claro). Quanto ao DSpace, acho que é (informalmente) institucionalizado na UnB. Penso que minha iniciativa é isolada (não conheço outros projetos relacionados na UnB).* | ***Preservação de Dados***<br>• DSpace<br>• Repositório<br>• Não é sistema de recuperação de dados |
| GOVERNO FEDERAL | *Organização de pasta digitais.* | ***Ausência de um sistema para recuperação de dados brutos***<br>• *Organização de pastas - Windows* |
| UNIVERSIDADE FEDERAL | *Na verdade, acabamos por construir um sistema referencial, que hoje está em transição institucional.* | ***Ausência de um sistema para recuperação de dados brutos***<br>• *Sistema referencial* |
| INSTITUTO DE PESQUISA DO MCTIC | *Todos os cadernos de pesquisa têm registrado os dados brutos gerados pelos instrumentos laboratoriais utilizados. Além disso, há dados que foram repassados para planilhas em Excel.* | ***Ausência de um sistema para recuperação de dados brutos***<br>• *Cadernos pessoais*<br>• *Dados manuais*<br>• *Dados em planilhas* |

| Perfil Entrevistado | Resposta | Códigos |
|---|---|---|
| INSTITUTO PESQUISA DO MCTIC | *Eu tenho a impressão de que se você digitar em alguns motores de busca como Google você tem acesso ao conteúdo do site. Mas isso não está estruturado em nenhum banco de dados para uma melhor recuperação da informação.* | ***Ausência de um sistema para recuperação de dados brutos***<br>• Mecanismo de busca - Google |
| UNIVERSIDADE | *Estão de forma manual, em papel, disquete, pendrive, entre outros.* | ***Ausência de um sistema para recuperação de dados brutos***<br>• *Cadernos pessoais*<br>• *Dados manuais*<br>• *Dados analógicos* |
| INSTITUTO PESQUISA DO MCTIC | *Sim, o mecanismo de busca do repositório. Apesar de não pesquisar por tipo de dado, ele pesquisar por título autor dentre outros. Por esse fato, marco 4 na escala.* | ***Preservação de Dados***<br>• DSpace<br>• Repositório<br>• Não é sistema de recuperação de dados |
| UNIVERSIDADE FEDERAL | *Eu ainda pretendo disponibilizar uma página na web, como professora da UFRGS, com meus dados de pesquisa.( )De uma forma geral o raw data dos dados abertos do governo estão disponíveis para consulta na web. Mas é preciso saber trabalhá-los. O Prof. Luís Carlos Herpen Bona de Curitiba, tem trabalhado com a Rede de Preservação Digital de Longo Prazo na Nuvem.* | ***Ausência de um sistema para recuperação de dados brutos*** |

Fonte: a autora.

Interessante observar que ainda há instituições que possuem dados brutos em cadernos de papel. Gray (2007) argumentava sobre *a importância do caderno pessoal do pesquisador* e que *antigamente os dados brutos de pesquisas estavam perdidos nesses cadernos*, mas que a ciência do Século XXI exige que esses dados estejam *online*.

Outro ponto que merece destaque entre os comentários qualitativos é a notícia de que há um professor em Curitiba trabalhando com uma rede de preservação digital de longo prazo na nuvem. Nesse aspecto, mostra-se relevante uma articulação entre a Rede Cariniana e a Rede desse professor. Afinal, no Brasil por não haver uma instituição alavancando as iniciativas em *e-science* de forma articulada, há uma duplicidade de esforços e a falta de recursos torna-se mais evidente. Nesse sentido, sugere-se que líderes de projetos e iniciativas para a gestão de dados científicos formem uma comunidade para a troca de ideias, alinhamento de esforços e expectativas. Certamente, a troca de ideias entre os líderes impulsionará acordos de parceria entre as instituições. Sugere-se também um papel mais ativo da Rede Cariniana para ampliar o quadro de participantes.

Quando questionados sobre a existência de um sistema de buscas que recuperasse os dados brutos de suas pesquisas (P4 [105]), dentre os participantes, 22,5% afirmaram que com precisão não possuem um sistema com essas características. E, 42,5% não conseguem ter essa mesma precisão para afirmar não possuir o sistema de busca – *o que revela dúvida sobre o que seria esse sistema de busca*. Somente 17,5% (7 pessoas), afirmaram que possuem um sistema de busca para seus dados, conforme revela a Tabela 9.

Tabela 9 – Sistema de busca que recupere os dos dados produzidos pela pesquisa – Brasil.

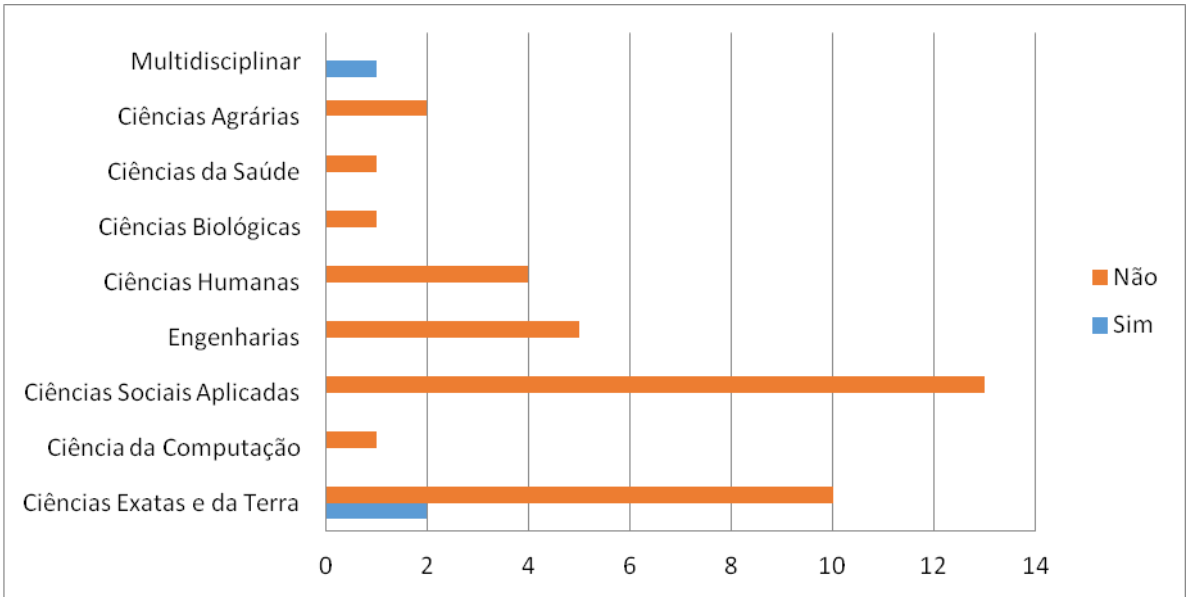
| | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| 0 | 9 | 22,5 | 22,5 | 22,5 |
| 1 | 6 | 15,0 | 15,0 | 37,5 |
| 4 | 3 | 7,5 | 7,5 | 45,0 |
| 5 | 1 | 2,5 | 2,5 | 47,5 |
| 6 | 1 | 2,5 | 2,5 | 50,0 |
| 10 | 1 | 2,5 | 2,5 | 52,5 |
| 38 | 1 | 2,5 | 2,5 | 55,0 |
| 49 | 1 | 2,5 | 2,5 | 57,5 |
| 50 | 3 | 7,5 | 7,5 | 65,0 |
| 69 | 1 | 2,5 | 2,5 | 67,5 |
| 70 | 1 | 2,5 | 2,5 | 70,0 |
| 85 | 1 | 2,5 | 2,5 | 72,5 |
| 91 | 1 | 2,5 | 2,5 | 75,0 |
| 95 | 1 | 2,5 | 2,5 | 77,5 |
| 98 | 1 | 2,5 | 2,5 | 80,0 |
| 99 | 1 | 2,5 | 2,5 | 82,5 |
| 100 | 7 | 17,5 | 17,5 | 100,0 |
| **Total** | **40** | **100,0** | **100,0** | |

Fonte: a autora.

Praticamente todos os participantes (92,5%) afirmaram que não trabalham com *workflow* científico. Apenas 8% declararam que trabalham. Esses são da área Multidisciplinar, seguida da área de Ciências Exatas e da Terra, conforme demonstra o Gráfico 8.

[105] A pergunta foi realizada em escala de três pontos e os participantes tinham como opção de resposta uma barra de rolagem com as opções <discordo>, <indiferente> e <concordo>. O primeiro ponto <discordo> correspondia ao intervalo de 1 a 49 na barra de rolagem, enquanto a opção <indiferente> foi de 50 a 55 e a opção concordo correspondia ao intervalo de 56 a 100.

Gráfico 8 – Pesquisador que utiliza *workflow* científico *versus* área de conhecimento – Brasil.




| Área de Pesquisa/ atuação | | Sim | Não | Total |
|---|---|---|---|---|
| Ciências Exatas e da Terra | Contagem | 2 | 10 | 12 |
| | % do Total | 5,0% | 25,0% | 30,0% |
| Ciência da Computação | Contagem | 0 | 1 | 1 |
| | % do Total | 0,0% | 2,5% | 2,5% |
| Ciências Sociais Aplicadas | Contagem | 0 | 13 | 13 |
| | % do Total | 0,0% | 32,5% | 32,5% |
| Engenharias | Contagem | 0 | 5 | 5 |
| | % do Total | 0,0% | 12,5% | 12,5% |
| Ciências Humanas | Contagem | 0 | 4 | 4 |
| | % do Total | 0,0% | 10,0% | 10,0% |
| Ciências Biológicas | Contagem | 0 | 1 | 1 |
| | % do Total | 0,0% | 2,5% | 2,5% |
| Ciências da Saúde | Contagem | 0 | 1 | 1 |
| | % do Total | 0,0% | 2,5% | 2,5% |
| Ciências Agrárias | Contagem | 0 | 2 | 2 |
| | % do Total | 0,0% | 5,0% | 5,0% |
| Multidisciplinar | Contagem | 1 | 0 | 1 |
| | % do Total | 2,5% | 0,0% | 2,5% |
| **Total** | Contagem | 3 | 37 | 40 |
| | % do Total | 7,5% | 92,5% | 100,0% |

Fonte: a autora.

Os *workflows* que são utilizados pelos participantes são os: Kepler, Taverna e YAWL. Dois (dentre os três) participantes que afirmaram utilizar tal *software* são das Ciências Exatas e da Terra. As respostas vão ao encontro da afirmação de Talia (2012) de que os sistemas de

fluxo de trabalhos científicos mais utilizados são: o Taverna, o Pegasus, o Triana, o Askalon, o Kepler, o GWES e o Karajan e seus maiores usuários são da área de Engenharias.

Quando questionados se os dados de suas pesquisas possuíam alguma classificação quanto ao ciclo de vida (P7[106]), dentre os participantes, 15% afirmam que não possuem nada em termos de ciclo de vida dos dados científicos. Em termos acumulativos, 70% não estão absolutamente convencidos da existência da classificação quanto ao ciclo de vida. Apenas 5% afirmaram categoricamente que seus dados são classificados quanto ao ciclo de vida, conforme demonstram os dados na Tabela 10.

Tabela 10 – Classificação dos dados da pesquisa quanto ao ciclo de vida – Brasil.

| | Frequência | Porcentual | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| **0** | 6 | 15,0 | 15,0 | 15,0 |
| **1** | 11 | 27,5 | 27,5 | 42,5 |
| **2** | 1 | 2,5 | 2,5 | 45,0 |
| **3** | 1 | 2,5 | 2,5 | 47,5 |
| **4** | 1 | 2,5 | 2,5 | 50,0 |
| **5** | 2 | 5,0 | 5,0 | 55,0 |
| **10** | 1 | 2,5 | 2,5 | 57,5 |
| **18** | 1 | 2,5 | 2,5 | 60,0 |
| **20** | 2 | 5,0 | 5,0 | 65,0 |
| **22** | 1 | 2,5 | 2,5 | 67,5 |
| **36** | 1 | 2,5 | 2,5 | 70,0 |
| **49** | 1 | 2,5 | 2,5 | 72,5 |
| **50** | 3 | 7,5 | 7,5 | 80,0 |
| **52** | 1 | 2,5 | 2,5 | 82,5 |
| **60** | 1 | 2,5 | 2,5 | 85,0 |
| **63** | 1 | 2,5 | 2,5 | 87,5 |
| **75** | 1 | 2,5 | 2,5 | 90,0 |
| **79** | 1 | 2,5 | 2,5 | 92,5 |
| **85** | 1 | 2,5 | 2,5 | 95,0 |
| **100** | 2 | 5,0 | 5,0 | 100,0 |
| **Total** | **40** | **100,0** | **100,0** | |

Fonte: a autora.

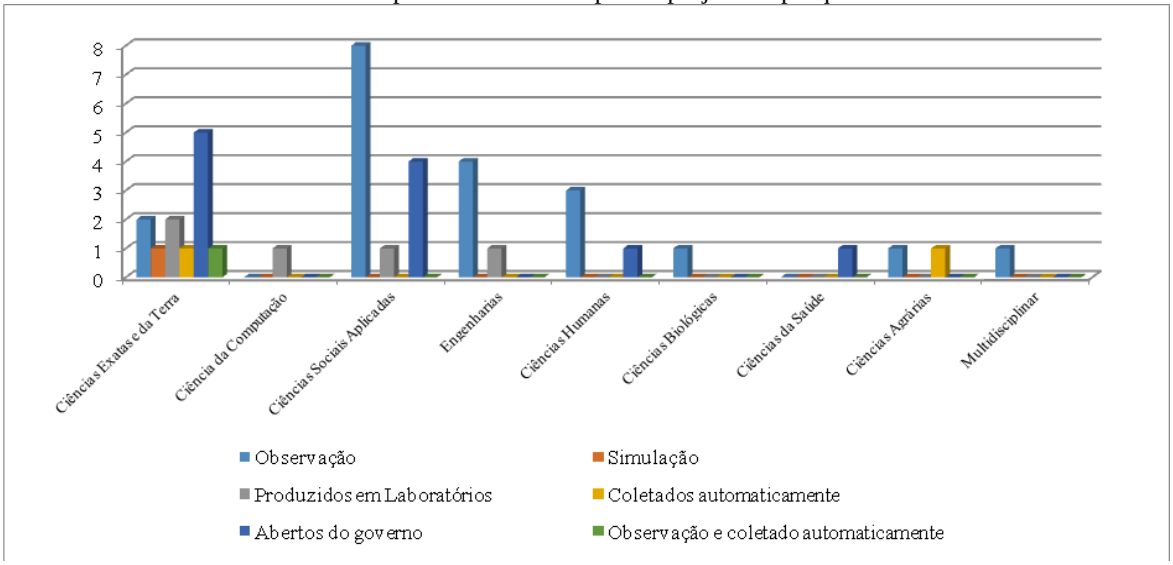
No que diz respeito ao ciclo de vida dos dados de pesquisa, merece ser comentado que o Dataone já possui um *framework* que para ser implementado basta o pesquisador segui-lo. É bem verdade que a temporalidade de guarda dos dados será determinada pelo pesquisador, ou por uma política institucional vigente. Mas, novamente, há que se ressaltar que a literatura para

[106] A pergunta também foi realizada em escala de três pontos e os participantes tinham como opção de resposta uma barra de rolagem com as opções <discordo>, <indiferente> e <concordo>. O primeiro ponto <discordo> correspondia ao intervalo de 1 a 49 na barra de rolagem, enquanto a opção <indiferente> foi representada pelo intervalo de 50 a 55 e a opção concordo intervalo de 56 a 100.

a gestão de dados científicos já possui diretrizes para definir o tempo de armazenamento de acordo com a classificação do dado (natureza do dado, nível de reprodutibilidade do dado, nível de processamento do dado). Assim, o pesquisador brasileiro e as próprias instituições brasileiras já têm um conjunto de diretrizes que podem e devem ser seguidas para implementar o ciclo de vida dos dados produzidos no contexto brasileiro.

Quando questionados sobre qual seria a principal fonte de dados para as suas pesquisas (P9), 50% (20 respondentes) dos participantes afirmaram que a principal fonte de dados corresponde aos dados de observação. Em segundo lugar aparecem como fontes os dados abertos do governo, que representam a resposta de 27,5% dos participantes, conforme demonstra o Gráfico 9.

Gráfico 9 – Principal fonte de dados para o projeto de pesquisa – Brasil.




| | | Observação | Simulação | Produzidos laboratórios | Coletados automat. | Abertos governo | Observação e coletado automat. | Total |
|---|---|---|---|---|---|---|---|---|
| Ciências Exatas e da Terra | Contagem | 2 | 1 | 2 | 1 | 5 | 1 | 12 |
| | % do Total | 5,0% | 2,5% | 5,0% | 2,5% | 12,5% | 2,5% | 30,0% |
| Ciência da Computação | Contagem | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| | % do Total | 0,0% | 0,0% | 2,5% | 0,0% | 0,0% | 0,0% | 2,5% |
| Ciências Sociais Aplicadas | Contagem | 8 | 0 | 1 | 0 | 4 | 0 | 13 |
| | % do Total | 20,0% | 0,0% | 2,5% | 0,0% | 10,0% | 0,0% | 32,5% |
| Engenharias | Contagem | 4 | 0 | 1 | 0 | 0 | 0 | 5 |
| | % do Total | 10,0% | 0,0% | 2,5% | 0,0% | 0,0% | 0,0% | 12,5% |
| Ciências Humanas | Contagem | 3 | 0 | 0 | 0 | 1 | 0 | 4 |
| | % do Total | 7,5% | 0,0% | 0,0% | 0,0% | 2,5% | 0,0% | 10,0% |

| | | Observação | Simulação | Produzidos laboratórios | Coletados automat. | Abertos governo | Observação e coletado automat. | Total |
|---|---|---|---|---|---|---|---|---|
| Ciências Biológicas | Contagem | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| | % do Total | 2,5% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 2,5% |
| Ciências da Saúde | Contagem | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| | % do Total | 0,0% | 0,0% | 0,0% | 0,0% | 2,5% | 0,0% | 2,5% |
| Ciências Agrárias | Contagem | 1 | 0 | 0 | 1 | 0 | 0 | 2 |
| | % do Total | 2,5% | 0,0% | 0,0% | 2,5% | 0,0% | 0,0% | 5,0% |
| Multidisciplinar | Contagem | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| | % do Total | 2,5% | 0,0% | 0,0% | 0,0% | 0,0% | 0,0% | 2,5% |
| **Total** | **Contagem** | **20** | **1** | **5** | **2** | **11** | **1** | **40** |
| | **% do Total** | **50,0%** | **2,5%** | **12,5%** | **5,0%** | **27,5%** | **2,5%** | **100,0%** |

Fonte: a autora.

Dentre os 50% de participantes, cuja principal fonte de dados são os dados de observação, 20% são das Ciências Sociais Aplicadas e 10% das Engenharias.

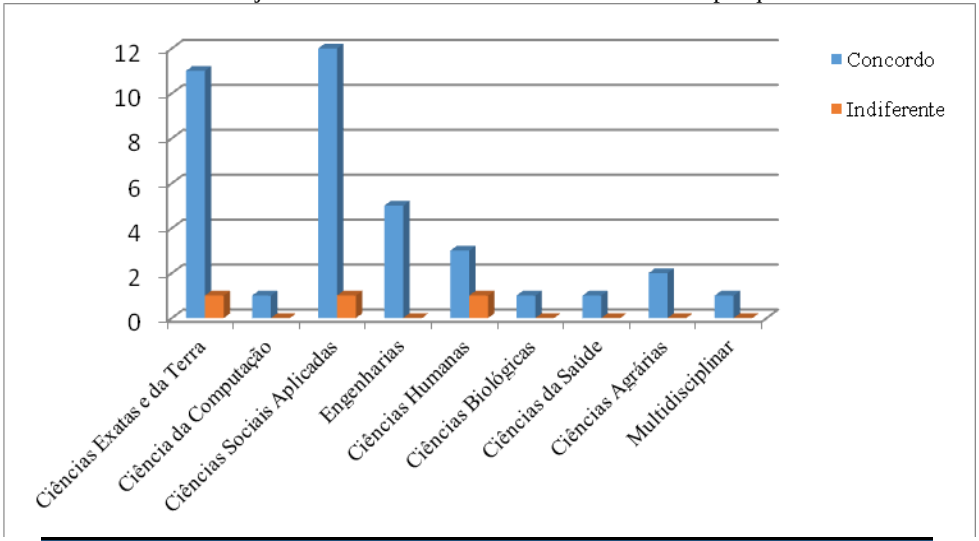
De todos os respondentes de Ciências Sociais Aplicadas, 60%[107] falaram que a principal fonte de dado são os dados de observação. Outros 30,7%[108] (Sociais Aplicadas) responderam que sua principal fonte de dados são os dados abertos de governo.

Dentre os respondentes da área de Engenharias, 80% dos participantes também têm como principal fonte de dados os de observação. Os demais 20% citaram os dados produzidos em laboratório.

No que diz respeito ao acesso aos dados brutos de outras pesquisas, 92,5 % dos participantes concordam que querem ter acesso a esses dados. Dentre esses, 30% dos participantes são das Ciências Sociais Aplicadas, outros 27,5% das Ciências Exatas e da Terra e 12,5% das Engenharias, conforme Gráfico 10.

[107] Total de respondentes = 13. Sendo que 8 falaram que os dados são de observação. X= (8/13)*100.
[108] Total de respondentes = 4. Sendo que 4 falaram que são dados abertos de governo. X= (4/13)*100.

Gráfico 10 – Desejo de ter acesso aos dados brutos de outras pesquisas – Brasil.




| | Concordo Totalmente | Concordo Parcialmente | Indiferente | Total |
|---|---|---|---|---|
| Ciências Exatas e da Terra | 9 | 2 | 1 | 12 |
| Ciência da Computação | 1 | 0 | 0 | 1 |
| Ciências Sociais Aplicadas | 11 | 1 | 1 | 13 |
| Engenharias | 5 | 0 | 0 | 5 |
| Ciências Humanas | 3 | 0 | 1 | 4 |
| Ciências Biológicas | 1 | 0 | 0 | 1 |
| Ciências da Saúde | 0 | 1 | 0 | 1 |
| Ciências Agrárias | 2 | 0 | 0 | 2 |
| Multidisciplinar | 1 | 0 | 0 | 1 |
| **Total** | **33** | **4** | **3** | **40** |

Fonte: a autora.

A respeito do percentual das Ciências Exatas e da Terra, bem como das Engenharias, merece ser comentado que são áreas extremamente internacionalizadas, o que reflete o acompanhamento da tendência mundial em pesquisa e desenvolvimento. Já na área de Sociais Aplicadas, merece ser destacado que a Administração possui tradição em práticas de *benchmarking*, ou seja, troca de informações sobre conjuntos de melhores práticas. Assim, infere-se que o compartilhamento de dados deve seguir essa tendência. A própria Biblioteconomia, integrante das Sociais Aplicadas, também tem tradição no compartilhamento de informações, a exemplo do desenvolvimento de catálogos públicos de acesso em linha, desenvolvidos graças a formatos como o MARC e regras do AACR2 que permitem uma interoperabilidade entre os dados.

No âmbito da gestão de dados científicos, merece ser ressaltada a iniciativa do Institute for Social Research – University of Michigan, que possui iniciativas de armazenamento de dados e compartilhamento desde meados da década de 1940.

Interessante observar que apesar de quererem ter acesso aos dados brutos de outra pesquisa, 37,5% dos participantes afirmam que não obtiveram esse acesso. O mesmo percentual de participantes (37,5%) também afirma que teve acesso aos dados brutos de outras pesquisas. Quando questionados qualitativamente, dentre os 37,5% que tiveram acesso, houve preponderância de casos em que o pesquisador teve acesso aos dados de outras pesquisas pelo fato de fazer parte de um mesmo grupo de pesquisas, ou pelo fato de ser próximo ao pesquisador líder. Tal situação não reflete a questão de compartilhamento de dados na *e-science* (dados *online* disponíveis para qualquer pesquisador). A respeito do assunto, merecem ser ressaltados alguns comentários de pesquisadores entrevistados.

> O acesso a dados brutos de outras pesquisas se dá pela **proximidade com o pesquisador**. Muitas vezes os **dados são compartilhados entre os pesquisadores de um mesmo grupo de pesquisa**. Assim, confia-se na idoneidade do pesquisador e consequentemente em seus dados (grifo nosso).

> Para mim, essa pergunta só faz sentido se for entre os professores que se conhecem que fazem parte de um mesmo ambiente de pesquisa. Se eu falar que vou pegar dados brutos de um professor que está fazendo pesquisa em outra escola, uma pessoa que você nunca viu, que você nem sabe quem é, ou o que nem sabe o que a pesquisa é???? [...] **Compartilharia meus dados apenas com alguém de confiança e dentro do mesmo grupo de pesquisa**. O fato de não ter ideia de como seria utilizado também me impediria de compartilhar (grifo nosso).

> Compartilho não no sentido de um repositório de dados abertos estarem disponibilizando meus dados. Mas, sim, com um colega pesquisador pedindo acesso aos meus dados disponíveis em um simples *pendrive*. **Ou seja, o compartilhamento seria com um pesquisador com o qual tenho uma relação de confiança. Certamente, seria um amigo meu**. **O compartilhamento é feito com fundamento na confiança do pesquisador**. [...] Hoje o compartilhamento é feito com aqueles professores/pares que eu tenho confiança em compartilhar meus dados. [...] Agora usando uma infraestrutura, uma ciberinfraestrutura de pesquisa para compartilhamento de dados a exemplo do que o IBICT quer fazer, ainda não.
> / Você já viu a piada – 'Deus criou o professor do departamento. Aí veio o diabo e criou o colega'(grifo nosso).

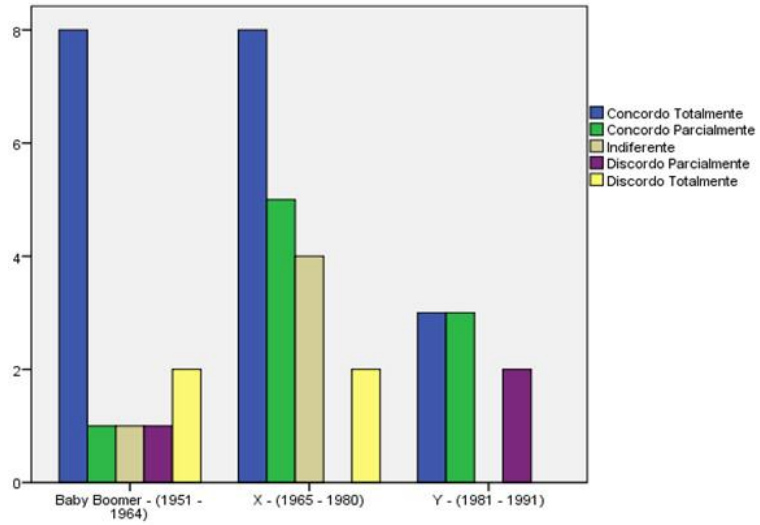
Dentre os participantes, apenas dois afirmaram que compartilhariam seu dado com qualquer outro pesquisador, conforme trechos de entrevista relatados a seguir.

> **Sim, compartilharia com qualquer pesquisador**. Não se deve limitar o acesso aos dados de uma pesquisa, a menos que, seja necessário, por algum motivo, como, por exemplo, o desenvolvimento tecnológico ou mesmo a solicitação de uma patente, aí nesse caso eu não compartilharia. Apenas iria sugerir que fosse feita a citação dos dados originais. Seria educado e ético reportar-se aos dados originais (grifo nosso).

> **O meu compartilhamento seria feito com qualquer outro pesquisador**. Não privilegiaria meus colegas de afinidade de pesquisa, grupos etc. Bom eu acho que eu colocaria apenas uma condição. Ela está mais relacionada com o direito moral, do que com o direito autoral. Na minha visão eu acho que seria correto e íntegro que essa pessoa citasse a origem desses dados. Se uma pessoa (pesquisador) está fazendo referência sobre um trabalho meu, eu preciso ser citado, esse é o processo ético. Por outro lado, se essa pessoa está utilizando meus dados brutos eu não vejo necessidade que essa pessoa me cite, mas é necessário que ela diga de onde aqueles dados vieram, qual a fonte daqueles dados. Se não o que fica parecendo é que a pessoa produziu aqueles dados sozinha (grifo nosso).

Outra questão que se mostra interessante, é que apesar de apenas 37,5% ter declarado que já teve acesso a dados de outras pesquisas, mais que a maioria, (70%) dos respondentes, concorda total ou parcialmente que compartilharia seus dados com outros pesquisadores. Quando agrupados por ano de nascimento, os dados revelam que o pesquisador compartilha seu dado independentemente da geração à qual pertence, conforme demonstram os dados do Gráfico 11.

Gráfico 11 – Acesso aos dados brutos de outras pesquisas *versus* geração do pesquisador – Brasil.




| | Concordo totalmente | Concordo parcialmente | Indiferente | Discordo parcialmente | Discordo totalmente | Total |
|---|---|---|---|---|---|---|
| **Baby Boomer** | 8 | 1 | 1 | 1 | 2 | 13 |
| **X (1965-1980)** | 8 | 5 | 4 | 0 | 2 | 19 |
| **Y (1981-1991)** | 3 | 3 | 0 | 2 | 0 | 8 |
| **Total** | 19 | 9 | 5 | 3 | 4 | **40** |

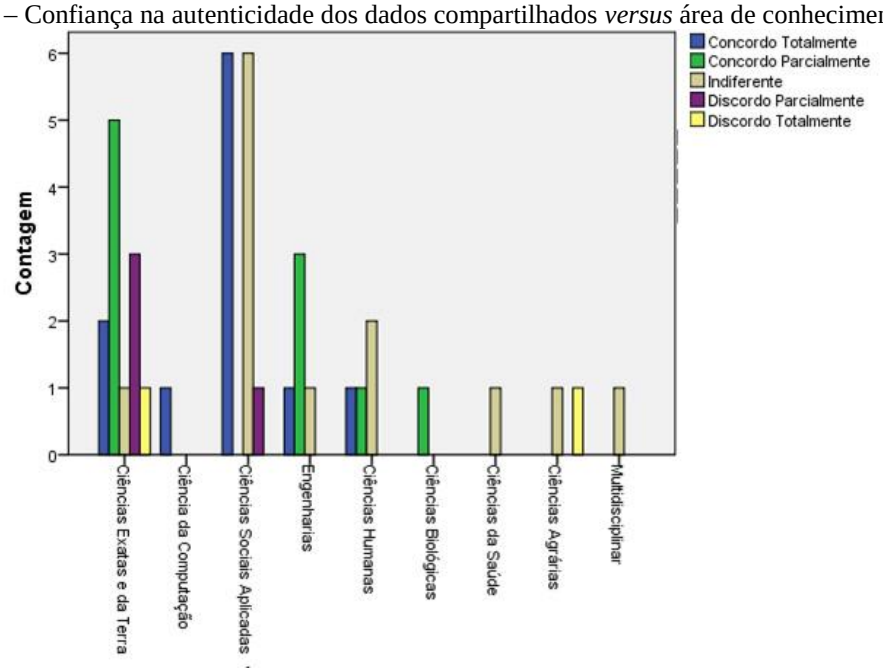
Fonte: a autora.

Dentre os participantes, 69%[109] dos componentes *baby boomer* concordam em compartilhar seus dados. Da Geração X, 68%[110] concordam em compartilhar seus dados. Já na Geração Y, 75%[111] dos participantes concordam em compartilhar seus dados. Esses números revelam que não há uma diferença de comportamento quanto ao compartilhamento de dados em relação à geração à qual o pesquisador pertence.

Interessante observar que nenhum participante afirmou que não compartilharia seu dado. Sobre as condições de compartilhamento, 55% dos participantes afirmou que compartilharia seus dados de forma total ou que possuiria regras para compartilhar seus dados; sendo que 37,5% concordam totalmente. Vinte por cento dos participantes não têm certeza se possuiria regras (*colocaram como sendo indiferente*) e apenas 25% discordam parcialmente ou totalmente do fato de se ter regras para compartilhamento. Dentre as regras citadas pelos participantes, merece ser destacado: a citação da fonte do dado primário; não permitir identificar o respondente ou a instituição da qual o dado provém; compreender como os dados serão reutilizados.

Quando questionados sobre a autenticidade dos dados a que se tem acesso, 32% manifestaram-se como indiferentes à questão de autenticidade dos dados de outras pesquisas; enquanto mais da metade dos respondentes (52,5%) concordam parcial ou totalmente que os dados compartilhados têm credibilidade, ou seja, são confiáveis, conforme Gráfico 12.

Gráfico 12 – Confiança na autenticidade dos dados compartilhados *versus* área de conhecimento – Brasil.




[109] X=(9/13)*100
[110] X=(13/19)*100
[111] X=(6/8)*100

| | | Concordo totalmente | Concordo parcialmente | Indiferente | Discordo parcialmente | Discordo totalmente | Total |
|---|---|---|---|---|---|---|---|
| Ciências Exatas e da Terra | Contagem | 2 | 5 | 1 | 3 | 1 | 12 |
| | % do Total | 5,0% | 12,5% | 2,5% | 7,5% | 2,5% | 30,0% |
| Ciência da Computação | Contagem | 1 | 0 | 0 | 0 | 0 | 1 |
| | % do Total | 2,5% | 0,0% | 0,0% | 0,0% | 0,0% | 2,5% |
| Ciências Sociais Aplicadas | Contagem | 6 | 0 | 6 | 1 | 0 | 13 |
| | % do Total | 15,0% | 0,0% | 15,0% | 2,5% | 0,0% | 32,5% |
| Engenharias | Contagem | 1 | 3 | 1 | 0 | 0 | 5 |
| | % do Total | 2,5% | 7,5% | 2,5% | 0,0% | 0,0% | 12,5% |
| Ciências Humanas | Contagem | 1 | 1 | 2 | 0 | 0 | 4 |
| | % do Total | 2,5% | 2,5% | 5,0% | 0,0% | 0,0% | 10,0% |
| Ciências Biológicas | Contagem | 0 | 1 | 0 | 0 | 0 | 1 |
| | % do Total | 0,0% | 2,5% | 0,0% | 0,0% | 0,0% | 2,5% |
| Ciências da Saúde | Contagem | 0 | 0 | 1 | 0 | 0 | 1 |
| | % do Total | 0,0% | 0,0% | 2,5% | 0,0% | 0,0% | 2,5% |
| Ciências Agrárias | Contagem | 0 | 0 | 1 | 0 | 1 | 2 |
| | % do Total | 0,0% | 0,0% | 2,5% | 0,0% | 2,5% | 5,0% |
| Multidisciplinar | Contagem | 0 | 0 | 1 | 0 | 0 | 1 |
| | % do Total | 0,0% | 0,0% | 2,5% | 0,0% | 0,0% | 2,5% |
| **TOTAL** | **Contagem** | **11** | **10** | **13** | **4** | **2** | **40** |
| | **% do Total** | **27,5%** | **25,0%** | **32,5%** | **10,0%** | **5,0%** | **100,0%** |

Fonte: a autora.

Os dados revelam uma tendência de que os pesquisadores das Ciências Sociais Aplicadas, bem como das Ciências Exatas e da Terra acreditam na autenticidade dos dados compartilhados. Por outro lado, os pesquisadores das Ciências Agrárias tendem a não confiar nesta autenticidade, pelo menos na amostra desta pesquisa. Ressalta-se que por tratar-se de uma amostra não probabilística, formada pelo critério de intencionalidade, esses dados não devem ser extrapolados. Eles servem apenas como um alerta para se observar o comportamento desses pesquisadores.

Dentre os que confiam nos dados brutos de outras pesquisas,

> Confiaria plenamente nos dados de outra pesquisa se ela fosse institucional. A chancela da instituição dá credibilidade à pesquisa. Por outro lado, também se conhece a reputação do pesquisador que está disponibilizando os dados.
>
> Quando acessei dados de outras pesquisas sequer conhecia o pesquisador, mas **conhecia a reputação da instituição onde esses dados são disponibilizados**. São dados de pesquisa públicos e que servem como referência para se testar o desenvolvimento próprio. De modo geral **tendo a confiar nos dados disponibilizados para consulta**. O que me impediria de confiar na autenticidade dos dados disponibilizados seria o fato de que minha experiência já permite perceber alguns comportamentos não éticos de pesquisadores (grifo nosso).

Dentre os que <u>não confiam</u> nos dados brutos de outras pesquisas,

> Em geral **estes dados não possuem informações suficientes para possibilitar uma avaliação da sua autenticidade, confiança e validade**. Para utilizar estes dados, entendo que é preciso entrar em contato com o pesquisador para saber detalhes da pesquisa (grifo nosso).

> Aqui no Brasil o pesquisador não tem essa obrigatoriedade. [...] **Não tem uma diretriz, ou mesmo um selo de qualidade para os dados brutos**. Ao mesmo tempo, se tivéssemos, surgem outras questões, tais como - você também vai ter que ter um grupo que vai precisar validar os dados brutos 'daquela' pesquisa. Nesse âmbito, torna-se cada vez mais complexo o processo de publicação científica, tornando-o mais lento. – Quem vai certificar os dados que certificam a qualidade da publicação. - O ideal é que os pesquisadores sejam os certificadores da qualidade de dados de outras pesquisas quando utilizarem esses dados (grifo nosso).

> Sobre o acesso a dados brutos de outras pesquisas - **Esses dados sem a proveniência, ou seja, sem tudo o que o qualifica não têm nenhuma utilidade para mim.** Isto só vai me atrapalhar, inclusive me desviar do meu objetivo (grifo nosso).

> Nunca tive acesso a dados brutos de pesquisa, apenas a resultados. No que diz respeito à autenticidade e transparência dos dados, **se esses estiverem disponíveis em um repositório público, tratado e auditado [...] sim, eu confiaria na autenticidade e transparência** (grifo nosso).

Apesar desta tese não trabalhar os dados abertos do governo, alguns participantes mencionaram na entrevista esse tipo de dado. Em todos os casos, ainda há dúvidas quanto à qualidade do dado, conforme trechos abaixo descritos.

> [...] Agora, sinceramente, eu tenho certa desconfiança do que o governo tem publicado nos repositórios de dados abertos, não posso afirmar que não acredito, mas tenho uma certa desconfiança.

> Quando se trata de dados abertos ainda não vejo a realidade brasileira apresentar características de fidedignidade e veracidade. A transparência atende aos requisitos legais e obrigatórios dos princípios do governo eletrônico.

> Quanto aos dados abertos do governo, em minha opinião, são dados disponibilizados com pouco cuidado. É preciso lapidar o dado para conseguir trabalhar com ele [...] O que se observa é que eles disponibilizam o que pode ficar totalmente transparente mesmo e não vai apresentar problema. Em função desses pequenos casos, observo que as instituições públicas não disponibilizam tudo que poderiam e/ou deveriam disponibilizar.

No que diz respeito à infraestrutura para gestão de dados científicos, 45% dos respondentes afirmaram que sua instituição não possui essa infraestrutura. Outros 25% afirmaram que a instituição possui infraestrutura para a gestão de dados. Os outros 30% não têm certeza (marcaram indiferente), conforme revelam os dados da Tabela 11.

Tabela 11 – Infraestrutura para a gestão de dados científicos na instituição – Brasil.

| | Concordo totalmente | Concordo parcialmente | Indiferente | Discordo parcialmente | Discordo totalmente | Total |
|---|---|---|---|---|---|---|
| ANTAQ | 0 | 0 | 0 | 0 | 1 | 1 |
| CGEE | 0 | 1 | 0 | 0 | 0 | 1 |
| **ELSEVIER** | **1** | 0 | 0 | 0 | 0 | 1 |
| EMBRAPA | 0 | 0 | 2 | 0 | 0 | 2 |
| EMILIO GOELDI | 0 | 0 | 0 | 1 | 0 | 1 |
| **IBICT** | **1** | 1 | 1 | 0 | 0 | 3 |
| ICMBIO | 0 | 0 | 1 | 0 | 0 | 1 |
| **IEN** | **2** | 0 | 1 | 1 | 0 | 4 |
| INPA | 0 | 1 | 0 | 0 | 0 | 1 |
| SENADO FEDERAL | 0 | 0 | 1 | 0 | 0 | 1 |
| UCB | 0 | 0 | 1 | 0 | 2 | 3 |
| UFF | 0 | 1 | 0 | 0 | 1 | 2 |
| UFG | 0 | 0 | 0 | 0 | 1 | 1 |
| **UFJF** | **1** | 0 | 1 | 0 | 0 | 2 |
| **UFMG** | **1** | 0 | 0 | 0 | 0 | 1 |
| UFPB | 0 | 0 | 0 | 0 | 3 | 3 |
| UFRGS | 0 | 0 | 1 | 0 | 0 | 1 |
| UFSC | 0 | 0 | 0 | 0 | 1 | 1 |
| UNB | 0 | 0 | 1 | 0 | 4 | 5 |
| USP | 0 | 0 | 2 | 2 | 1 | 5 |
| **Total** | **6** | **4** | **12** | **4** | **14** | **40** |
| **% do Total** | **15,0%** | **10,0%** | **30,0%** | **10,0%** | **35,0%** | **100,0%** |

Fonte: a autora.

Dentre os participantes, apenas 15% concordam plenamente que a sua instituição possui tal infraestrutura. As instituições que aparecem em destaque são UFMG, IBICT, UFJF, IEN e ELSEVIER. Causou certa surpresa a USP não figurar entre as instituições que possuem uma infraestrutura para dados científicos, até mesmo em função do Programa *e-Science* da FAPESP, pelo fato deste exigir do pesquisador um plano de gestão de dados. Assim, infere-se que os participantes desta pesquisa podem não conhecer completamente a infraestrutura tecnológica da USP.

No que diz respeito à infraestrutura para gestão de dados científicos, vale a pena retomar, na visão de Corrêa (2016), os fatores-chave, para o êxito da infraestrutura, quais sejam: a) o

incentivo, b) a formação de pesquisadores, tanto em seu papel de produtores, como de usuários de infraestrutura de informação de dados, c) a infraestrutura técnica e de organização, d) o financiamento da infraestrutura para novos desenvolvimentos e logística de dados.

Em se tratando de curadoria de dados, 62,5% dos participantes discordam que sua instituição tenha um departamento dedicado a realizar a curadoria dos dados, sendo que 10% discordam parcialmente e 52,5% totalmente. Apenas 15% concordam que sua instituição possui um departamento com essa finalidade, sendo que 7,5% concordam parcialmente e 7,5% totalmente, conforme demonstram os dados da Tabela 12.

Tabela 12 – Departamento dedicado à curadoria de dados – Brasil.

| | Frequência | Porcentual | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| Concordo totalmente | 3 | 7,5 | 7,5 | 7,5 |
| Concordo parcialmente | 3 | 7,5 | 7,5 | 15,0 |
| Indiferente | 9 | 22,5 | 22,5 | 37,5 |
| Discordo parcialmente | 4 | 10,0 | 10,0 | 47,5 |
| Discordo totalmente | 21 | 52,5 | 52,5 | 100,0 |
| **Total** | **40** | **100,0** | **100,0** | |

Fonte: a autora.

Quando analisamos institucionalmente a questão, as únicas instituições que concordam totalmente ou parcialmente que possuem um departamento que realiza a atividade de curadoria de dados são: a UFMG, o INPA e o IEN, conforme demonstra a Tabela 13.

Tabela 13 – Departamento dedicado a curadoria de dados *versus* instituição – Brasil.

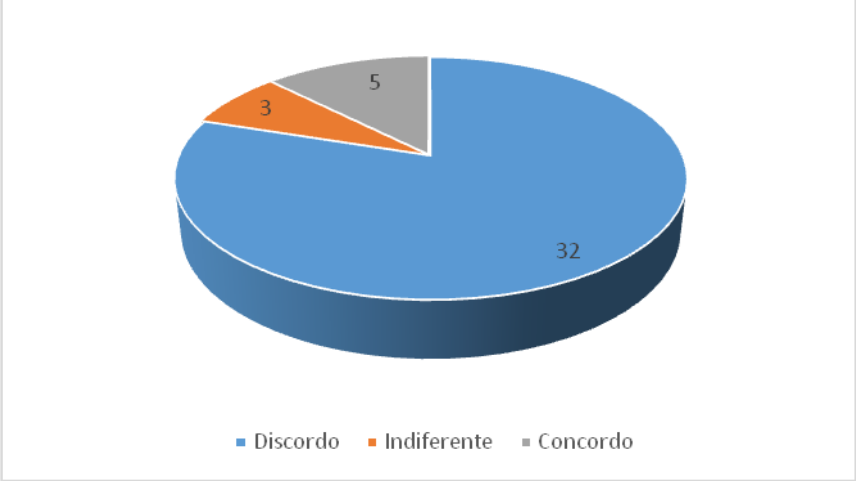
| | Concordo totalmente | Concordo parcialmente | Indiferente | Discordo parcialmente | Discordo totalmente | Total |
|---|---|---|---|---|---|---|
| ANTAQ | 0 | 0 | 0 | 0 | 1 | 1 |
| CGEE | 0 | 0 | 1 | 0 | 0 | 1 |
| ELSEVIER | 0 | 0 | 0 | 0 | 1 | 1 |
| EMBRAPA | 0 | 0 | 1 | 1 | 0 | 2 |
| EMILIO GOELDI | 0 | 0 | 0 | 1 | 0 | 1 |
| **IBICT** | **0** | **0** | **2** | **1** | **0** | **3** |
| ICMBIO | 0 | 0 | 0 | 1 | 0 | 1 |
| IEN | 3 | 0 | 1 | 0 | 0 | 4 |
| INPA | 0 | 1 | 0 | 0 | 0 | 1 |
| SENADO FEDERAL | 0 | 0 | 1 | 0 | 0 | 1 |
| UCB | 0 | 0 | 1 | 0 | 2 | 3 |
| UFF | 0 | 0 | 0 | 0 | 2 | 2 |
| UFG | 0 | 0 | 0 | 0 | 1 | 1 |

| | Concordo totalmente | Concordo parcialmente | Indiferente | Discordo parcialmente | Discordo totalmente | Total |
|---|---|---|---|---|---|---|
| UFJF | 0 | 0 | 0 | 0 | 2 | 2 |
| UFMG | 0 | 1 | 0 | 0 | 0 | 1 |
| UFPB | 0 | 0 | 0 | 0 | 3 | 3 |
| UFRGS | 0 | 1 | 0 | 0 | 0 | 1 |
| UFSC | 0 | 0 | 0 | 0 | 1 | 1 |
| UNB | 0 | 0 | 1 | 0 | 4 | 5 |
| USP | 0 | 0 | 1 | 0 | 4 | 5 |
| **Total** | **3** | **3** | **9** | **4** | **21** | **40** |

Fonte: a autora.

Quando questionados se a sua instituição possui uma política para a gestão de dados científicos – mais da metade dos respondentes, 80% (=32), acreditam que a instituição não possui uma política com tal finalidade (sendo 22,5 parcialmente e 57,5% totalmente). Apenas 12,5% dos participantes (=5) afirmaram que a sua instituição possui uma política para a gestão de dados científicos, conforme demonstra o Gráfico 13.

Gráfico 13 – Política institucional para a gestão de dados científicos – Brasil.




| | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| Discordo parcialmente | 9 | 22,5 | 22,5 | 42,5 |
| Discordo totalmente | 23 | 57,5 | 57,5 | 100,0 |
| Indiferente | 3 | 7,5 | 7,5 | 20,0 |
| Concordo totalmente | 3 | 7,5 | 7,5 | 7,5 |
| Concordo parcialmente | 2 | 5,0 | 5,0 | 12,5 |
| **Total** | **40** | **100,0** | **100,0** | |

Fonte: a autora.

As instituições cujos participantes citaram possuir uma política para a gestão de dados científicos foram: INPA, IBICT, IEN e ICMBIO.

O ICMBIO de fato possui uma política explícita para tratar da gestão de dados científicos. Essa política está registrada na Instrução Normativa nº 03, de 1º de setembro de 2014 que, dentre outros, "[...] regulamenta a disponibilização, o acesso e o uso de dados e informações recebidos pelo Instituto de Informações Chico Mendes de Conservação e Biodiversidade por meio do SISBio", bem como na Instrução Normativa nº 2 de 25 de novembro de 2015 que "Institui a política de dados e informações sobre biodiversidade do Instituto Chico Mendes de Conservação da Biodiversidade e dispõe sobre sua disponibilização, acesso e uso".

O IEN, por sua vez, até a análise dos dados desta pesquisa, não havia publicado uma política explícita. Mas há que se considerar que o instituto já possui, na visão de Herrera (1995) uma política implícita sobre o tema. Como norte para os pesquisadores da instituição, Sayão e Sales (2015) publicaram o Guia de Gestão de Dados de Pesquisa pela CNEN, porém há que se observar que esse documento não caracteriza a existência de uma política institucional explícita, a exemplo do ICMBIO. A respeito do assunto, merece ser ressaltado o trecho de uma entrevista com um participante do IEN.

> No âmbito da nossa instituição, nós estamos tentando implementar. Temos isso como premissa até formalizar a portaria que cria no papel essa política. Agora uma coisa é a nossa visão e enuncia-la. Já a outra, é entrar no dia a dia da instituição. A gente quer desenvolver mais essa cultura e procurar oferecer instrumentos nesse sentido. A própria CarpeDiem para que toda produção científica do instituto seja armazenada, seja oferecido um espaço para os pesquisadores também guardarem suas informações. Nós temos essa visão, mas ainda é incipiente.

No que diz respeito ao INPA, o participante da pesquisa refere-se a um *draft* de política institucional que se encontra pendente de aprovação pela diretoria, bem como a política de dados do Programa de Pesquisa em Biodiversidade (PPBIO) que foi publicada no Diário Oficial da União por meio da Portaria 693, de 20 de agosto de 2009, conforme trecho de entrevista a seguir:

> Por vezes existe uma política específica para determinado projeto, como PPBIO. Atualmente há um *draft* de uma política na Diretoria do INPA. Há dois anos tenta-se aprovar esse *draft*. Esse *draft* foi produzido pelo Comitê de Tecnologia da Informação. O draft representa a consolidação das diversas políticas que tiveram na instituição em função da exigência de determinados projetos.

A exemplo da situação do IEN, a política do PPBIO não representa uma política explícita institucional do instituto, mas sim uma política explícita para o programa em questão. Porém, há que se ressaltar que a política do PPBIO disciplina o que são dados, o que são dados sensíveis e, ostensivos. Além disso, regulamenta sobre a gestão e auditoria dos dados, sobre o uso e acesso às bases de dados e orienta sobre questões de propriedade intelectual. Essa política reflete o alto grau de internacionalização da pesquisa em biodiversidade produzida no Brasil, bem como a necessidade de o pesquisador brasileiro acompanhar as tendências internacionais.

Já sobre o IBICT, o instituto ainda não possui uma política implícita, tão pouco explícita sobre os dados de pesquisa produzida pelo instituto, tão pouco diretrizes que sinalizem uma política para determinada área do conhecimento. Porém, os dados desta pesquisa revelam que o IBICT tem se dedicado a pesquisar sobre o tema de forma a construir uma metodologia de gestão de dados. Há que se ressaltar que a situação no IBICT é mais complexa, pois além de ter que desenvolver uma política institucional para os dados de pesquisa produzidas pelo Instituto, o órgão, enquanto instituição responsável por promover a competência, o desenvolvimento de recursos e a infraestrutura de informação em ciência e tecnologia, precisa desenvolver um conjunto de diretrizes que atendam às diferentes áreas conhecimento.

No que diz respeito à elaboração de uma política para a gestão de dados científicos no Brasil, ou seja, uma política nacional, a maioria dos participantes (87,5%) afirmou que o Brasil precisa de uma política para a gestão de dados científicos, sendo 80% totalmente e 7,5% parcialmente. Quando questionados sobre quem seriam os interlocutores dessa política, os respondentes dividiram-se entre agência de fomento, institutos de pesquisa do MCTIC, pesquisadores das universidades, o INPI – que tem como foco a propriedade intelectual e questões vinculadas à concessão de patentes e o IBICT em conjunto com o MCTIC, que por sua vez lideram o número de respostas qualitativas como agente interlocutor da política. Merecem ser destacados os trechos de entrevistas abaixo relacionados:

> CNPQ, INPI, juntamente com instituições de ensino superior e fomento por meio de um comitê aos moldes do marco civil.
>
> **Ministérios da Educação e Ciência e Tecnologia** e órgãos vinculados. Vejo também a necessidade de envolvimento do próprio gabinete da Presidência para garantir a institucionalização dessa iniciativa (grifo nosso).
>
> As agências de fomento, ABC, SBPC, o Conselho Técnico-Científico da Capes, os programas de Pós-Graduação, as sociedades científicas.
>
> A política deve contemplar diferentes dimensões (política, normativa, legal, cultural, técnica, tecnológica e ética)... deve ser capitaneada por uma

> instituição federal e direcionada para a aplicação e normalização nas universidades públicas, instituições de ensino. Além de ser uma norma mandatária de regras para financiamento de projetos de pesquisa.
>
> Deveria contar com a participação das universidades, instituições de fomento, pesquisadores, **órgãos de governo responsáveis por políticas de C&T** (grifo nosso).
> Seriam necessários padrões de formatação e segurança de acesso. O principal interlocutor deveria ser o **MCTIC**, com o apoio do **IBICT** e das agências de fomento (grifo nosso).
>
> Membros do **MCTI**, **IBICT** e agências de fomento (grifo nosso).
>
> Ministérios, **IBICT**, Universidades, Institutos de Pesquisa (grifo nosso).

Quando questionados sobre o acesso a documento que contivesse diretrizes para armazenamento e gestão de dados científicos, dentre os participantes, 60% afirmaram que não tiveram acesso a tais documentos, sendo que 55% totalmente e 5% parcialmente. Dos que concordam, 25% concordam totalmente e 10% parcialmente e apenas 5% marcaram ser indiferente, conforme demonstram os dados da Tabela 14.

Tabela 14 – Política institucional para a gestão de dados científicos – Brasil.

| | | Concordo totalmente | Concordo parcialmente | Indiferente | Discordo parcialmente | Discordo totalmente | Total |
|---|---|---|---|---|---|---|---|
| Ciências Exatas e da Terra | Contagem | **2** | **1** | 1 | 1 | **7** | 12 |
| | % do Total | 5,0% | 2,5% | 2,5% | 2,5% | 17,5% | 30,0% |
| Ciência da Computação | Contagem | 0 | 0 | 0 | 0 | 1 | 1 |
| | % do Total | 0,0% | 0,0% | 0,0% | 0,0% | 2,5% | 2,5% |
| Ciências Sociais Aplicadas | Contagem | **4** | **1** | 0 | 0 | **8** | 13 |
| | % do Total | 10,0% | 2,5% | 0,0% | 0,0% | 20,0% | 32,5% |
| Engenharias | Contagem | **3** | 0 | 0 | 0 | 2 | 5 |
| | % do Total | 7,5% | 0,0% | 0,0% | 0,0% | 5,0% | 12,5% |
| Ciências Humanas | Contagem | **1** | 0 | 0 | 1 | 2 | 4 |
| | % do Total | 2,5% | 0,0% | 0,0% | 2,5% | 5,0% | 10,0% |
| Ciências Biológicas | Contagem | 0 | **1** | 0 | 0 | 0 | 1 |
| | % do Total | 0,0% | 2,5% | 0,0% | 0,0% | 0,0% | 2,5% |
| Ciências da Saúde | Contagem | 0 | 0 | 0 | 0 | 1 | 1 |
| | % do Total | 0,0% | 0,0% | 0,0% | 0,0% | 2,5% | 2,5% |
| Ciências Agrárias | Contagem | 0 | **1** | 0 | 0 | 1 | 2 |
| | % do Total | 0,0% | 2,5% | 0,0% | 0,0% | 2,5% | 5,0% |

| | | Concordo totalmente | Concordo parcialmente | Indiferente | Discordo parcialmente | Discordo totalmente | Total |
|---|---|---|---|---|---|---|---|
| Multidisciplinar | Contagem | 0 | 0 | 1 | 0 | 0 | 1 |
| | % do Total | 0,0% | 0,0% | 2,5% | 0,0% | 0,0% | 2,5% |
| **Total** | **Contagem** | **10** | **4** | **2** | **2** | **22** | **40** |
| | **% do Total** | **25,0%** | **10,0%** | **5,0%** | **5,0%** | **55,0%** | **100,0%** |

Fonte: a autora.

Os dados da Tabela 14 quando analisados em conjunto com a área de formação do pesquisador não permitem inferências do tipo – *dentre os participantes, o que tiveram acesso a documentos de diretrizes, há uma predominância da área de exatas e da terra,* por exemplo.

Dentre os participantes que responderam qualitativamente à questão, foram citados como documentos com diretrizes para gestão de dados científicos o material produzido pela NSF, Dataone, Research Data Aliance, Digital Curation Center e Rainforts. No âmbito nacional, foram citados o Portal da Biodiversidade, o PPBIO, PELD, SISBIO, o Guia de Gestão de Dados de Sayão e Sales (2015), a Política do Museu Emílio Goeldi e a Política do Jardim Botânico do Rio de Janeiro. A respeito dessas duas últimas, a pesquisadora não conseguiu contato com as instituições de forma a explorar nesta tese o conteúdo de tais políticas. Interessante ressaltar que a política do Museu Emílio Goeldi não foi mencionada pelo pesquisador da instituição que participou desta pesquisa.

No que diz respeito à capacidade para elaborar um plano de gestão de dados, dentre os participantes, 77,5% afirmaram ter condições de elaborá-lo caso a agência de fomento solicitasse um. Dentre os que pediriam ajuda, foi citado, dentre os possíveis profissionais para apoiar a elaboração do plano, o bibliotecário, em primeiro lugar, seguido por profissional da área de tecnologia da informação e, em terceiro lugar, o estatístico. A respeito do destaque da profissão de bibliotecário, a literatura internacional, incluso o material produzido pela ARL, já expõe essa nova habilidade do profissional e a emergência desse perfil profissional. Além disso, o próprio material produzido e disponibilizado pelo DataONE já fornece diretrizes básicas para a construção de um plano de gestão de dados científicos, o que facilita os primeiros passos do pesquisador na gestão de seus dados.

## 4.2 ANÁLISE DOS DADOS REFERENTES ÀS AGÊNCIAS DE FOMENTO

O instrumento de coleta de dados aplicado aos funcionários de agências de fomento, funcionários e/ou membros de conselhos das fundações de amparo à pesquisa, bem como funcionários do MCTIC, foi enviado para 169 pessoas, sendo respondido por 22 funcionários. A Tabela 15 apresenta a composição da amostra.

Tabela 15 - Número de respondentes por instituição.

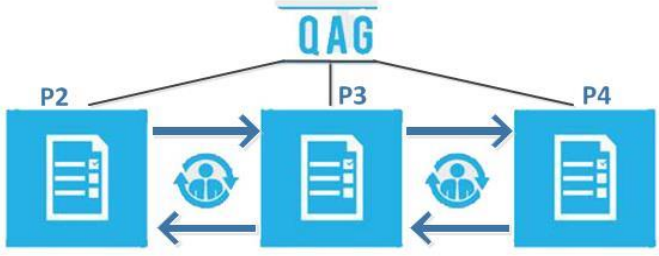
| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| Válido | **CAPES** | 8 | 36,4 | 36,4 | 36,4 |
| | **CNPq** | 3 | 13,6 | 13,6 | 50,0 |
| | **FAP** | 9 | 40,9 | 40,9 | 90,9 |
| | **MCTIC** | 2 | 9,1 | 9,1 | 100,0 |
| | **Total** | 22 | 100,0 | 100,0 | |

Fonte: a autora.

Os dados revelam que o questionário teve a participação de 13% dos respondentes, com predominante participação da CAPES, baixa participação do CNPq, assim como das FAP, ao se considerar que cada estado da Federação possui uma FAP.

A compreensão de como as Agências de Fomento e Fundações de Amparo à Pesquisa têm observado a necessidade do pesquisador quanto à necessidade de gestão de dados científicos e sua atuação no tema se deu por meio de três perguntas diferentes (P2[112], P3[113], P4[114]), mas complementares, conforme ilustrado na Figura 28.

Figura 28 – Relacionamento entre P2, P3 e P4 – questionários agências de fomento.



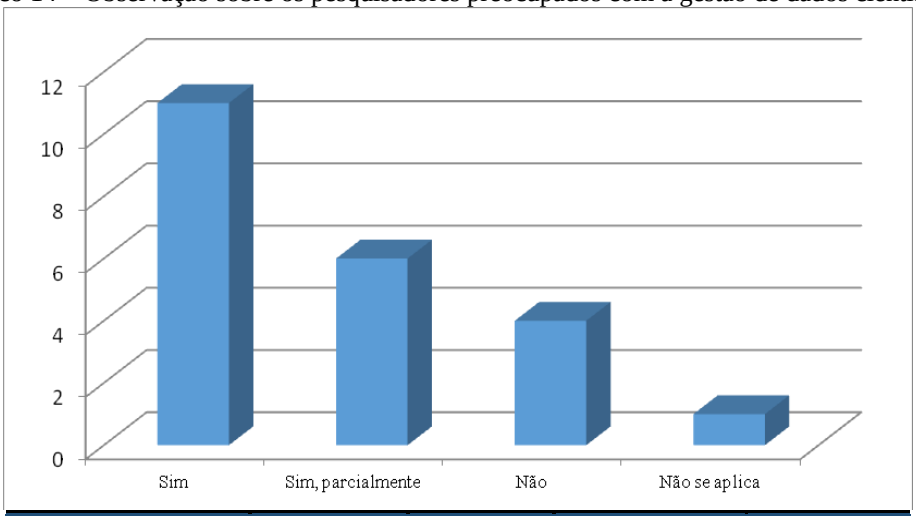

Fonte: a autora.

[112] P2 – Durante a sua trajetória em agências de fomento para pesquisa, observou se havia pesquisadores preocupados com a gestão e a preservação dos dados produzidos por suas respectivas pesquisas?
[113] P3 – Enquanto servidor de uma agência de fomento, o (a) senhor (a) conhece e/ou conheceu projetos desenvolvidos pelas universidades e/ou instituições de pesquisa brasileiras que têm/tiveram a necessidade de uma política que norteasse a gestão de dados científicos? Se possível, dê exemplos.
[114] P4 – Na sua opinião, as agências de fomento estão atentas à necessidade de tratamento, armazenamento e à preservação digital de dados científicos brutos que estão sendo produzidos pelas instituições brasileiras? As agências precisam fomentar essa discussão?

O pressuposto de entendimento é – ***Em termos de coerência administrativa****, se as agências de fomento têm observado que os pesquisadores brasileiros estão preocupados com a gestão e preservação dos dados produzidos pelas suas pesquisas, elas estarão atentas à necessidade de tratamento, armazenamento e preservação digital de dados científicos e fomentarão essa discussão e há uma chance de que elas conheçam projetos que necessitaram de uma política que norteasse a gestão desses dados.*

Quando questionados se durante a sua trajetória foi observado se havia pesquisadores preocupados com a gestão e a preservação dos dados produzidos por suas respectivas pesquisas, a pergunta (P2) foi feita de forma qualitativa e quando analisadas geraram as seguintes categorias de informação: <sim, plenamente>, <sim, parcialmente>, <não>, <não se aplica>. Dentre os participantes, pelo menos 17 respondentes manifestaram que sim, sendo que 7 informaram que sim, mas parcialmente, conforme ilustra o Gráfico 14.

Gráfico 14 – Observação sobre os pesquisadores preocupados com a gestão de dados científicos.




| | Frequência | P% | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| Sim | 11 | 50 | 50 | 50 |
| Sim, parcialmente | 6 | 27,3 | 27,3 | 95,5 |
| Não | 4 | 18,2 | 18,2 | 68,2 |
| Não se aplica | 1 | 4,5 | 4,5 | 100 |
| **Total** | 22 | 100 | 100 | |

Fonte: a autora.

Ao analisar-se as respostas qualitativamente, observa-se que os funcionários das agências de fomento que responderam <**não**> (4 respondentes[115]) acreditam que o pesquisador

[115] Ressalta-se que nem sempre todos os participantes responderam as questões qualitativas, uma vez que essas foram classificadas no instrumento como optativas.

está *mais preocupado* com a produção de artigos e patentes, e que observaram que o pesquisador não atribui essa responsabilidade às agências de fomento, conforme trechos transcritos no Quadro 15.

Quadro 15 – A preocupação com artigos e patentes: respostas qualitativas da pergunta 2.

| Perfil entrevistado | Texto | Códigos |
|---|---|---|
|  **FAP** | *"A preocupação dos pesquisadores no geral é **produzir artigo** para passar em concurso ou obter um elevado grau no CNPq. Não encontrei esse perfil de pesquisador interessado em guardar seus dados em agências de fomento".* | ***Gestão e a preservação dos dados (preocupação)*** • Produção de artigos |
|  **FAP** | *"Nem sempre eles se preocupam. Apenas a **produção de artigos** ou artefatos relacionados aos projetos de pesquisa ficam como legado".* | ***Gestão e a preservação dos dados (preocupação)*** • Produção de artigos |
|  **CAPES** | *"Não em relação à gestão e preservação, mas sim à respeito de **domínio dos dados e patentes**".* | ***Gestão e a preservação dos dados (preocupação)*** • Produção de patentes |
|  **CAPES** | *Não percebi por parte dos pesquisadores expectativas de que a Capes tivesse a responsabilidade com gestão e preservação dos dados produzidos por suas pesquisas.* | ***Gestão e a preservação dos dados (preocupação)*** • Não é responsabilidade da agência |

Fonte: a autora.

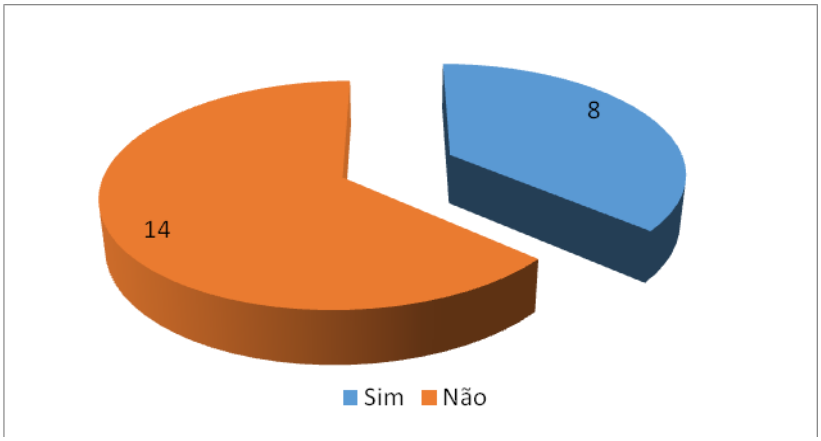
Merece destaque o comentário de um respondente do CNPq que atribui à agência a responsabilidade de gestão e preservação dos dados produzidos por pesquisas fomentadas pelo órgão, conforme transcrição a seguir.

**CNPq**

*"Sim, muitas vezes observei esta preocupação e sempre correlacionei o papel do CNPq em ser responsável por este tipo de dados".*

Quando questionados se o funcionário conhece e/ou conheceu *projetos desenvolvidos pelas universidades e/ou instituições de pesquisa brasileiras que têm/tiveram a necessidade de uma política que norteasse a gestão de dados científicos*, 14 respondentes afirmaram que <**não**>, o que causa certa preocupação se a agência de fomento não está atenta ao fenômeno, ou o que seria pior, o pesquisador brasileiro não sente a necessidade de gerir e preservar seus dados. O Gráfico 15 sintetiza o entendimento dos respondentes sobre o tema.

Gráfico 15 – Visão dos participantes sobre a necessidade de uma política para a gestão de dados científicos.




| | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| Sim | 8 | 36,4 | 36,4 | 36,4 |
| Não | 14 | 63,6 | 63,6 | 100 |
| **Total** | **22** | **100** | **100** | |

Fonte: a autora.

Merece ser comentado que dentre os poucos que responderam <**sim**>, a avaliação da resposta qualitativa revelou um desconhecimento para com a pergunta, pois os exemplos citados não representam gestão de dados científicos e em outros casos a resposta foi desconexa, conforme revelam as respostas descritas no Quadro 16.

Quadro 16 – Exemplos de projetos com necessidade de política para a gestão de dados científicos: o desconhecimento do tema.

| Perfil entrevistado | Texto | Códigos |
|---|---|---|
|  **CAPES** | *Sim, por exemplo, o banco de teses da CAPES e iniciativas do IPEA em copilar de modo global info sobre C&T.* | ***Desconhecimento do tema*** • Banco de teses da CAPES • Modo global info sobre C&T |
|  **FAP** | *Sim. Mapeamento de redes de colaborações cientificas.* | ***Desconhecimento do tema*** • Mapeamento de redes de colaborações cientificas |
|  **FAP** | *Sim, a FAPESP gera um quadro de notícias todo dia a respeito das pesquisas realizadas, principalmente em sua instituição.* | ***Desconhecimento do tema*** • Quadro de notícias FAPESP |

Fonte: a autora.

São exemplos de resposta desconexa para com o que foi questionado o conteúdo apresentado no Quadro 17.

Quadro 17 – Exemplos de projetos com necessidade de política para a gestão de dados científicos: respostas desconexas com a pergunta.

| Perfil entrevistado | Texto | Códigos |
|---|---|---|
|  **FAP** | *Sim. Alguns pesquisadores têm dificuldades de gestão, principalmente, administrativo financeiros o que acaba atrapalhando as pesquisas propriamente ditas.* | ***Respostas desconexas com a pergunta*** • Dificuldades de gestão, principalmente, administrativo financeiros |
|  **FAP** | *Existem as leis que dão ao pesquisador a segurança do plágio e as leis de patentes que asseguram a autenticidade e mérito do desenvolvedor.* | ***Respostas desconexas com a pergunta*** • Segurança do plágio • Leis de patentes |
|  **CAPES** | *Observamos que nas IES esse problema também é recorrente. Talvez seja mesmo um viés cultural brasileiro, mas que precisa ser mudado.* | ***Respostas desconexas com a pergunta*** • Problema recorrente • Viés cultural brasileiro |

Fonte: a autora.

Apenas dois respondentes que colocaram <**sim**> como opção descreveram um exemplo que de fato corresponde ao tema, conforme revela o Quadro 18.

Quadro 18 – Exemplos de projetos com necessidade de política para a gestão de dados científicos: respostas pertinentes com o questionamento.

| Perfil entrevistado | Texto | Códigos |
|---|---|---|
|  **CNPq** | *Sim. Por exemplo, o projeto Genoma (Xylella Fastidiosa) e a Rede Nacional de Sequenciamento de DNA.* | ***Política para a gestão de dados científicos*** • Projeto Genoma (Xylella Fastidiosa) • Rede Nacional de Sequenciamento de DNA |
|  **FAP** | *Sim. Por exemplo, os dados gerados a partir do mapeamento de polimorfismos importantes para a anemia falciforme na região Nordeste sem o devido impacto nas políticas de serviços de saúde.* | ***Política para a gestão de dados científicos*** • Mapeamento de polimorfismos importantes para a anemia falciforme na região Nordeste |

Fonte: a autora.

Os respondentes que responderam <**não**>, quando se manifestaram de forma qualitativa, foram mais contundentes e revelaram conhecimento sobre o tema, conforme demostra o depoimento a seguir.

> Não. Entretanto, com respeito a dados sobre pesquisa e desenvolvimento, atualmente há um contato com o IBCT para a **implementação do BR-CRIS. [...]**

Contrapondo a visão dos pesquisadores em relação à postura das agências de fomento sobre o tema, segue a transcrição de trecho de entrevista realizada com uma pesquisadora, pertencente à Geração X, da área de Ciências Sociais Aplicadas com mais de dez anos de

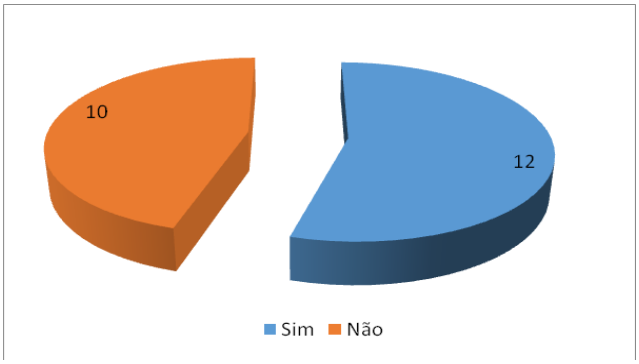
atuação, que sintetiza a percepção dos pesquisadores de que as agências de fomento desconhecem questões inerentes à gestão de dados científicos

> Entendo que o acesso a dados brutos de pesquisa é de **suma importância para a realidade científica do Brasil e do mundo**. Especial atenção às pesquisas que recebem fomento governamental. Neste caso, disponibilizar estes dados deveria ser uma exigência. **No Brasil as agências de fomento desconhecem mecanismos e modelos de fluxos para a gestão destes dados**. [grifos do respondente]

A realidade entre os participantes da pesquisa mostra-se preocupante quando se retoma a literatura de dados científicos, afinal o Reino Unido e os Estados Unidos têm programas de incentivo ao compartilhamento e gestão de dados desde 2001, ou seja, há pelo menos dezesseis anos. Percebe-se na própria literatura a especificação de diretrizes para tratamento de dados pelas próprias agências de fomento, a exemplo da NSF que desde 2011 exige que todas as propostas tenham um documento complementar de duas páginas com o título Plano de Gestão de Dados.

Quando questionados se as agências de fomento estão *atentas à necessidade de tratamento, armazenamento e à preservação digital de dados científicos brutos que estão sendo produzidos pelas instituições brasileiras*, os respondentes ficaram bem divididos, sendo que 12 responderam <**sim**> e 10 responderam <**não**>, conforme demonstra o Gráfico 16.

Gráfico 16 – Necessidade de tratamento, armazenamento e preservação digital de dados científicos.




| | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|
| Sim | 12 | 54,5 | 54,5 | 54,5 |
| Não | 10 | 45,5 | 45,5 | 100 |
| **Total** | **22** | **100** | **100** | |

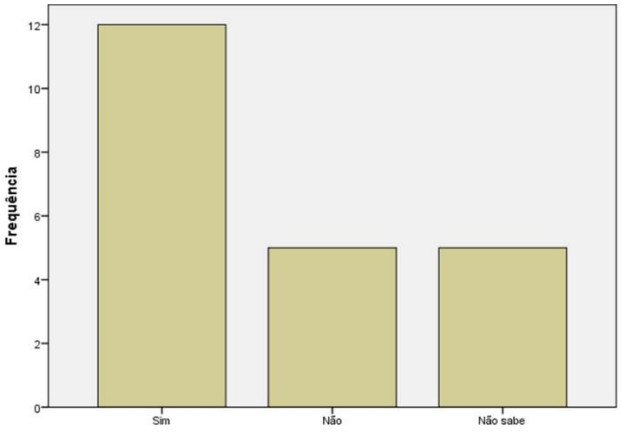
Fonte: a autora.

Sabe-se que o fato de estar atenta não significa ter iniciativas para o tratamento, armazenamento e a preservação digital. Mas merece ser ressaltado que durante esta pesquisa, apenas os editais do Programa *e-Science* da FAPESP exigiam um plano de gestão de dados dos

pesquisadores, ou seja, foi a única agência de fomento que manifestou de forma explícita a sua atenção para as questões de tratamento, armazenamento e preservação de dados científicos brutos. As demais agências, apesar de terem participantes que responderam <**sim**> para a questão, não demonstraram como essa atenção é manifestada.

Quando questionados se as agências precisam fomentar essa discussão, os respondentes novamente ficaram divididos. Sendo que 12 responderam que <**sim**>, cinco responderam que <**não**> e os demais cinco responderam que <**não sabem**>, conforme demonstra o Gráfico 17.

Gráfico 17 – As agências precisam fomentar a discussão sobre tratamento, armazenamento e preservação de dados científicos.




Fonte: a autora.

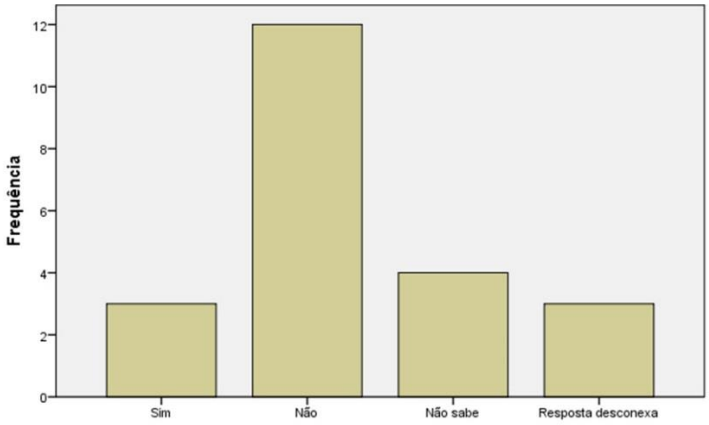
A questão que fica para a reflexão é que as agências não apenas devem fomentar a discussão, como devem incentivar a elaboração de planos de gestão de dados científicos e cobrar do pesquisador uma postura proativa na administração de seus dados, conforme já indicado na literatura internacional revisada sobre o tema.

Novamente, em termos de uma gestão coerente na administração pública, espera-se que, se uma agência de fomento está atenta ao fenômeno de gestão de dados científicos, que ela forneça diretrizes de coleta, tratamento técnico, armazenamento e preservação de dados científicos gerados para aquele pesquisador que tem apoio (financiamento) da agência para com a sua pesquisa. Nesse sentido, foi formulada a **P5** do instrumento de coleta de dados, que pergunta aos entrevistados *se na sua respectiva agência existia alguma diretriz de coleta, tratamento técnico, armazenamento e preservação de dados científicos gerados pela pesquisa financiada pela agência.*

Os dados revelam que 12 participantes da pesquisa responderam que a agência não possui diretrizes de *coleta, tratamento técnico, armazenamento e preservação de dados*

*científicos gerados pela pesquisa financiada pela agência.* Porém, merece ser ressaltado que em uma amostra de 22 pessoas, 10 pessoas dividiram-se entre as opções de resposta <sim>, <não sabe> ou <resposta desconexa>, conforme demonstra o Gráfico 18.

Gráfico 18 – Agência de fomento com diretriz de coleta, tratamento técnico, armazenamento e preservação de dados científicos.




Fonte: a autora.

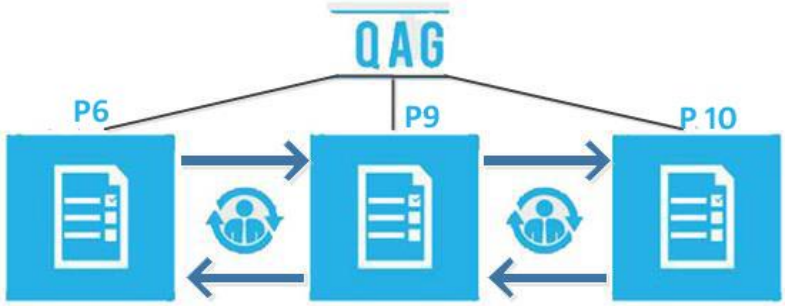
Uma vez que os respondentes também se dividiram quanto ao fato das agências *estarem atentas à necessidade de tratamento, armazenamento e a preservação digital de dados científicos brutos*, parece lógico que esses também se mantenham divididos quanto ao fato das agências terem *diretriz para coleta, tratamento técnico, armazenamento e preservação de dados*, mas o fato é que isso revela um desconhecimento das diretrizes da própria agência.

A respeito do assunto, é importante ressaltar as colocações de Hey e Trefhten (2002) que já alertavam para o fato de que "talvez as agências de financiamento precisem acrescentar algum incentivo para encorajar essa abordagem de compartilhamento de dados científicos". Assim, torna-se fundamental que essas agências estejam atentas ao fenômeno de explosão do volume de dados científicos, bem como da necessidade de diretrizes para a gestão, uso, compartilhamento e preservação desses dados em longo prazo.

Ao aprofundar-se a compreensão do tema (*e-science*), foi questionado se as agências de fomento possuíam algum sistema que recuperasse os dados brutos da pesquisa por ela financiada (P6), e se os funcionários dessas agências acreditavam que a gestão dos dados da pesquisa financiada era um dever da respectiva agência (P9). Ambas as perguntas foram feitas em escala Likert, com a opção de se realizar comentários qualitativos na P 10. Assim, a

compreensão desse ponto foi realizada por meio da análise entre as perguntas P6[116], P9[117] e P10[118], conforme ilustra a Figura 29.

Figura 29 – Relacionamento entre P6, P9 e P10 – questionários das agências de fomento.




Fonte: a autora.

Os resultados quantitativos da P6 mostram-se preocupantes, pois, considerando a amostra de 22 participantes, quatro marcaram a opção de resposta <indiferente>, dois marcaram <concordam totalmente> e outros três marcaram <concordo parcialmente>, conforme demonstra a Tabela 16. Ou seja, cinco respondentes indicaram como resposta que as agências têm um sistema de recuperação para esses dados. Ressalta-se que durante a revisão de literatura não foi encontrado nenhum relato na literatura brasileira e internacional que comentasse sobre um sistema desenvolvido, em nível nacional, com apoio de alguma agência. O único relato na literatura brasileira é o Portal Brasileiro de Dados Espaciais, criado por meio do Decreto 6.666 de 27/11/2008. A utilização da *web* como fonte de dados secundários para a pesquisa (vista aos *sites* das respectivas agências de fomento) também não revelou que alguma agência possuísse tal sistema. A visita em instituições e a segunda coleta de dados desta pesquisa, realizada com doutores envolvidos com o tema, revelou apenas duas plataformas de dados brutos de pesquisa – o SISBIO desenvolvido pelo Ministério do Meio Ambiente, Instituto Chico Mendes, USP e com apoio do MCTIC por meio da Secretaria de Pesquisa e Desenvolvimento (SEPED). A outra iniciativa de destaque é a realizada pela CNEN no âmbito de preservação dos dados de pesquisas financiadas apoiadas ela própria CNEN.

[116] A sua Agencia de Fomento (CAPES, CNPq, FAP) possui algum sistema que recupere os dados brutos (*raw data*) das pesquisas por ela financiada/apoiada?
[117] Na sua opinião, a gestão dos dados científicos, gerados por pesquisas financiadas por agências de fomento, é um dever da respectiva agência?
[118] Por favor, se possível, comente sobre a questão anterior.

Tabela 16 – A agência de fomento possui um sistema de recuperação de dados científicos da pesquisa por ela financiada.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| Válido | Concordo totalmente | 2 | 9,1 | 9,1 | 9,1 |
| | Concordo parcialmente | 3 | 13,6 | 13,6 | 22,7 |
| | Indiferente | 4 | 18,2 | 18,2 | 40,9 |
| | Discordo parcialmente | 3 | 13,6 | 13,6 | 54,5 |
| | Discordo totalmente | 10 | 45,5 | 45,5 | 100,0 |
| **Total** | | **22** | **100,0** | **100,0** | |

Fonte: a autora.

Ao serem questionados sobre ser de responsabilidade da agência financiadora ser a responsável pela gestão dos dados científicos coletados em pesquisas por ela financiada, as respostas configuraram-se conforme demonstra a Tabela 17.

Tabela 17 – A agência é responsável pela gestão dos dados científicos coletados por pesquisa que teve seu apoio financeiro.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| Válido | Concordo totalmente | 10 | 45,5 | 45,5 | 45,5 |
| | Concordo parcialmente | 2 | 9,1 | 9,1 | 54,5 |
| | Indiferente | 6 | 27,3 | 27,3 | 81,8 |
| | Discordo parcialmente | 1 | 4,5 | 4,5 | 86,4 |
| | Discordo totalmente | 3 | 13,6 | 13,6 | 100,0 |
| **Total** | | **22** | **100,0** | **100,0** | |

Fonte: a autora.

A análise das respostas qualitativas parece revelar um consenso quanto ao fato de ser demasiado para as agências assumirem a função de gerenciar os dados científicos da pesquisa por elas financiadas. Da mesma forma, há um consenso de que pesquisa financiada com dinheiro público deve ter seus dados de acesso público. Nas próprias respostas, foram apresentadas alternativas administrativas, a exemplo das relacionadas no Quadro 19.

Quadro 19 – Dados qualitativos: a agência é responsável pela gestão dos dados científicos coletadas por pesquisa que teve seu apoio financeiro.

| Perfil Entrevistado | Texto | Códigos |
|---|---|---|
|  **CAPES** | *As pesquisas realizadas com financiamento público devem se constituir em* ***informações de acesso público****.* | ***Responsabilidade da agência na gestão de dados*** • O dado deve ser público |

| Perfil entrevistado | Texto | Códigos |
|---|---|---|
|  **CAPES** | *Acredito que seja um **dever compartilhado** pela agência e pela instituição à qual o pesquisador está vinculado. Entretanto, me pergunto se não seria mais adequado ter um órgão que centralizasse isso. Teoricamente deveria ser o IBICT, mas ele está tão sucateado que mesmo com projetos interessantes há pouca colaboração por parte das instituições.* | ***Responsabilidade da agência na gestão de dados***<br>• Deve ser compartilhado entre agência e instituição de pesquisa.<br>• IBICT |
|  **CAPES** | *Considero importante a agência ter clareza do universo dos dados produzidos pelas pesquisas que são apoiadas pela agência, principalmente para facilitar a rede de articulação entre pesquisadores e grupos de pesquisa. No entanto, me parece um **desafio a agência assumir a responsabilidade da gestão e armazenamento do grande volume de dados** produzidos pelas pesquisas.* | ***Responsabilidade da agência na gestão de dados***<br>• Instituição específica |
|  **CAPES** | *Entendo que deveria haver uma **instituição para esse fim** pois essa gestão é um mundo de informações a serem tratadas. As agências de fomento deveriam se beneficiar dessas informações para nortear suas políticas, mas não gerar essas informações.* | ***Responsabilidade da agência na gestão de dados***<br>• Instituição específica |
|  **ANÔNIMO** | *Não acredito que seja um dever, mas pode ser uma **discussão encabeçada pelas agências**.* | ***Responsabilidade da agência na gestão de dados***<br>• Discussão liderada pelas agências |
|  **FAP** | *Isso vai depender do tipo de pesquisa e do sigilo dos dados. Além disso, **repositórios para todos os dados produzidos em todas as pesquisas podem se tornar inviáveis**. Uma solução seria obrigar que as pesquisas feitas deixassem os dados (não sigilosos) disponíveis em algum **repositório global** pelos autores das pesquisas, assim como é feito com teses e dissertações. Esse também seria um papel dos avaliadores dos trabalhos: cobrar que os dados usados nas avaliações sejam divulgados publicamente, junto com os artigos produzidos com esses mesmos dados.* | ***Responsabilidade da agência na gestão de dados***<br>• Instituição específica<br>• Repositório global |
|  **FAP** | *Na própria **universidade** é que a ciência ocorre, e **nessas instituições é que deveriam haver portais de informação científica**. Agências de fomento servem apenas para analisar, firmar convênios e fornecer as condições para que a pesquisa e sua divulgação ocorram.* | ***Responsabilidade da Agência na gestão de dados***<br>• Instituição de pesquisa. |

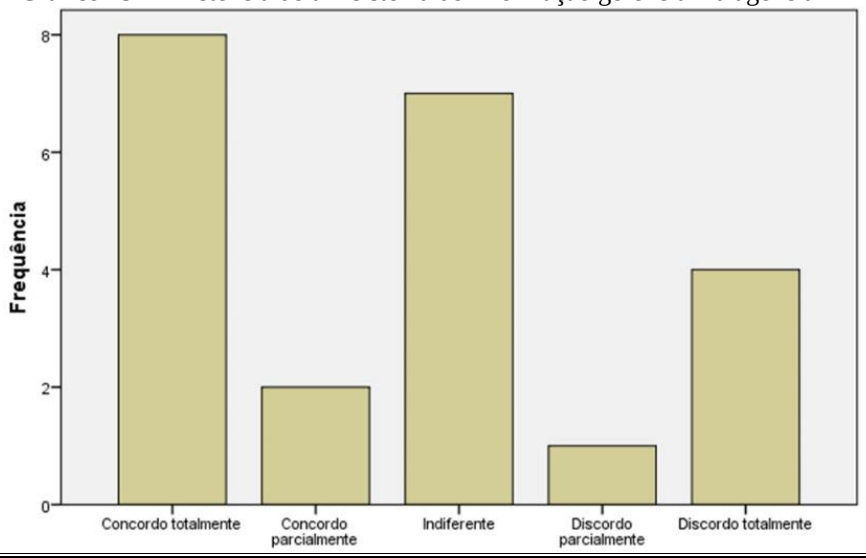
Fonte: a autora.

De fato, a missão institucional do IBICT vai ao encontro da necessidade de se ter um órgão central que administre os dados científicos. O próprio instituto já tem lastro na realização de tais atividades no âmbito da informação bibliográfica. A diferença, agora, é que o instituto,

frente aos desafios da pesquisa do Século XXI e da *e-Science*, precisa assumir uma postura proativa na gestão de dados científicos com iniciativas semelhantes à BDTD e às já realizadas pela Rede Cariniana.

Do ponto de vista da avaliação da política pública e do redirecionamento da mesma quando necessário, espera-se que os órgãos financiadores de pesquisa tenham informações gerenciais sobre as pesquisas por ele financiadas. Nesse sentido foram realizadas as perguntas 7 e 8 – que tinham como objetivo *identificar se a agência de fomento possuía um sistema de informação gerencial e se esse sistema registrava o tipo de dado produzido pelo pesquisador*. As respostas de P7 são apresentadas no Gráfico 19.

Gráfico 19 – Existência de um sistema de informação gerencial na agência




| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| Válido | Concordo totalmente | 8 | 36,4 | 36,4 | 36,4 |
| | Concordo parcialmente | 2 | 9,1 | 9,1 | 45,5 |
| | Indiferente | 7 | 31,8 | 31,8 | 77,3 |
| | Discordo parcialmente | 1 | 4,5 | 4,5 | 81,8 |
| | Discordo totalmente | 4 | 18,2 | 18,2 | 100,0 |
| **Total** | | **22** | **100,0** | **100,0** | |

Fonte: a autora.

No que diz respeito ao registro do tipo de dado produzido pelo pesquisador no sistema de informações gerenciais (P8), foi realizada uma análise qualitativa das respostas categorizando-as em <sim>, <não> e <resposta desconexa>. As repostas configuraram-se conforme demonstrado na Tabela 18. Ou seja, é possível concluir que os sistemas de informações gerenciais das agências não registram o tipo de dado produzido pelo pesquisador.

Tabela 18 – O sistema de informações gerenciais registra o tipo de dado produzido pelo pesquisador.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Sim | 3 | 13,6 | 13,6 | 13,6 |
| | Não | 16 | 72,7 | 72,7 | 86,4 |
| | Resposta desconexa | 3 | 13,6 | 13,6 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

Dentre as respostas qualitativas destaca-se a resposta de um funcionário do CNPq por citar a *Política de Dados do Programa de Pesquisa Ecológica de Longa Duração* (PELD), com o objetivo de regulamentar as formas de disponibilização, acesso e uso dos dados gerados pelos pesquisadores da rede PELD.

Ao se considerar a experiência internacional de repositórios de dados, a exemplo do Repositório Datacite, questionou-se os funcionários das agências de fomento sobre *a pertinência do Brasil ter um repositório central de dados brutos de pesquisa (P11), a exemplo da BDTD no caso de repositório de teses e dissertações*. Essa pergunta foi realizada de forma qualitativa e depois houve uma categorização das respostas. Foi observada praticamente uma unanimidade nas respostas, pois 21 participantes responderam que <**sim**> e apenas uma resposta foi categorizada como desconexa.

As respostas desta pergunta corroboram os comentários qualitativos feitos na Tabela 8 e no Quadro 17 ao citarem o IBICT como uma instituição que apresenta missão e expertise compatível com a atividade. Interessante observar que os respondentes que citaram o IBICT são servidores da CAPES e do CNPq – o que revela conhecimento da máquina pública federal. Por outro lado, servidores das FAP acreditam que a iniciativa de gerenciar um repositório central seria da CAPES e do CNPq, conforme revelam os dados do Quadro 20.

Quadro 20 – O Brasil precisa de um repositório central de dados de pesquisa – instituição responsável.

| Perfil entrevistado | Texto | Códigos |
|---|---|---|
|  **CNPq** | *Sim, desde que a iniciativa seja no sentido de tentar produzir conhecimento e não apenas ser um "empilhador" de dados desconectados. Acredito que o **IBICT**, caso bem gerenciado, poderia assumir esse papel.* | ***Repositório central de dados de pesquisa*** • IBICT |
|  **CAPES** | *Acho muito pertinente a discussão sobre um repositório para armazenar a produção no âmbito da ciência e tecnologia. No entanto, a Capes na forma como está estruturada hoje não teria condições de assumir esta responsabilidade. Considero que o **IBICT** poderia alavancar esta discussão.* | ***Repositório central de dados de pesquisa*** • IBICT |

| Perfil entrevistado | Texto | Códigos |
|---|---|---|
|  **CAPES** | *Sim. Seria uma ferramenta estratégica, de alta relevância e interesse político, social e econômico. A BDTD não é compulsória, o que acaba por desmerecer a ferramenta, que não contempla tudo. O Banco de Teses da Capes, entretanto, possui todas as informações porque os dados são alimentados pelo Sucupira que é obrigatório para as instituições que desejam ter boas avaliações e mais fomento para suas atividades. Como disse anteriormente, o **IBICT** deveria ser este órgão, mas sucateado, sem pessoal e sem apoio da comunidade científica é praticamente impossível. Desconheço instituição que atualmente tenha capacidade técnica para isso.* | ***Repositório central de dados de pesquisa*** • IBICT |
|  **CAPES** | *Sim, um repositório central seria de extrema relevância para o desenvolvimento da ciência nacional.* | ***Repositório central de dados de pesquisa*** • Relevância para o desenvolvimento da ciência nacional |
|  **CAPES** | *Sim. Centralizar informações permitem ter uma visão geral do investimento em pesquisa e seu retorno para o país.* | ***Repositório central de dados de pesquisa*** • Visão geral do investimento em pesquisa |
|  **CNPq** | *Sim deveria. Grau: importantíssimo. As teses são um primeiro passo já consolidado. Sugeriria o uso da INDA – Infraestrutura de Dados Abertos do Ministério do Planejamento.* | ***Repositório central de dados de pesquisa*** • INDA |
|  **FAP** | *Sim. Um banco central de informações de todas as pesquisas realizadas no país. Creio que talvez o CNPq.* | ***Repositório central de dados de pesquisa*** • CNPq |
|  **FAP** | *Essa ação é importantíssima, muito relevante. Concordo que a condução desse projeto deveria ficar a cargo do CNPq.* | ***Repositório central de dados de pesquisa*** • CNPq |
|  **FAP** | *Certamente, o Brasil necessita de mais iniciativas semelhantes à Biblioteca Digital de Teses e Dissertações, o que deveria assumir proporções de um dos principais meios de busca de dados científicos aos pesquisadores e gestores, norteando tomada de decisões em CT&I. Acredito que o CNPq e a CAPES sejam os órgãos mais adequados a encabeçar esse tipo de atividade.* | ***Repositório central de dados de pesquisa*** • CNPq • CAPES |

Fonte: a autora.

Quando questionados sobre se faz parte do planejamento da agência de fomento desenvolver *softwares* de acesso aos dados brutos (*raw data*) da pesquisa por ela financiada (P12), as respostas mostraram-se bem divididas, conforme Tabela 19.

Tabela 19 – A agência de fomento planeja desenvolver *softwares* de acesso a dados brutos.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Sim | 7 | 31,8 | 31,8 | 31,8 |
| | Não | 8 | 36,4 | 36,4 | 68,2 |
| | Não sabe | 6 | 27,3 | 27,3 | 95,5 |
| | Resposta desconexa | 1 | 4,5 | 4,5 | 100,0 |
| **Total** | | **22** | **100,0** | **100,0** | |

Fonte: a autora.

Analisar P12 em relação aos dados apresentados no Quadro 17, referente à necessidade de um repositório central de dados de pesquisa, causa certa surpresa. Pois, se as agências desenvolvessem tal *software* seria uma inciativa duplicada em relação ao repositório central de dados.

Sobre o fato das agências de fomento precisarem de uma política para a gestão dos dados científicos, as perguntas foram feitas em P13 em escala Likert, obtendo comentários qualitativos em P14, conforme resultados a seguir apresentados (Tabela 20).

Tabela 20 – A necessidade de uma política institucional para a gestão de dados científicos.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Concordo totalmente | 14 | 63,6 | 63,6 | 63,6 |
| | Concordo parcialmente | 2 | 9,1 | 9,1 | 72,7 |
| | Indiferente | 3 | 13,6 | 13,6 | 86,4 |
| | Discordo totalmente | 3 | 13,6 | 13,6 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

A pergunta 14 foi classificada no instrumento de pesquisa como opcional e sua opção de resposta foi livre (texto corrido). Dentre os respondentes, 14 a deixaram em branco. Três respostas mostraram-se desconexas e as demais cinco estão descritas no Quadro 21.

Quadro 21 – Necessidade de uma política institucional para a gestão de dados de pesquisa na agência de fomento.

| Perfil Entrevistado | Texto | Códigos |
|---|---|---|
|  **CAPES** | *A exemplo do Ministério das Relações Exteriores, é preciso que a administração pública como um todo tenha um* ***sistema de coleta de dados e memória****. Hoje, infelizmente, observamos o* ***esforço individual de alguns servidores para essa preservação****, contudo, isso fica centrado em uma pessoa.* | ***Necessidade de uma política para a gestão desses dados***<br>• Sistema de coleta de dados e memória<br>• Esforço individual para preservação |
|  **CNPq** | *É fundamental desenvolver essa política, mas acredito que não deva ser apenas institucional, mas sim nacional ou mundial. É algo que impacta no desenvolvimento da ciência como um todo.* | ***Necessidade de uma política para a gestão desses dados***<br>• *Nacional*<br>• *Mundial* |
|  **FAP** | *As próprias universidades detêm os dados digitalizados de teses e dissertações e poderiam adicionar artigos relevantes para comunidade acadêmica, avaliá-las e premiar os melhores pesquisadores. Sinto falta de uma maior divulgação das pesquisas realizadas dentro da universidade tanto que a maioria esmagadora dos alunos não detêm o conhecimento das pesquisas realizadas nos laboratórios, inclusive do próprio laboratório que realiza sua pesquisa.* | ***Necessidade de uma política para a gestão desses dados***<br>• *Iniciativa da instituição responsável pela pesquisa.* |

| Perfil Entrevistado | Texto | Códigos |
|---|---|---|
|  CAPES | *Avalio que uma política institucional para gestão de dados científicos deva incluir a observação de práticas de produção, planeamento e armazenamento de dados científicos e estudo dos dados.* | ***Necessidade de uma política para a gestão desses dados*** • Iniciativa institucional |
|  FAP | *As informações científicas resultantes de projetos de cooperação, teses, parcerias, e pesquisas individuais são um patrimônio intelectual riquíssimo para subsidiar políticas públicas de fomento ao desenvolvimento do conhecimento, em todas as áreas e temas.* | ***Necessidade de uma política para a gestão desses dados*** • Políticas públicas de fomento ao desenvolvimento do conhecimento |

Fonte: a autora.

No que diz respeito à formulação de uma *política nacional de dados científicos* o questionamento foi realizado em escala Likert na P15 com comentários qualitativos em P16. Dentre os participantes, dezoito respondentes concordam totalmente com essa necessidade, conforme demonstra a Tabela 21.

Tabela 21 – Necessidade de uma política para a gestão de dados científicos.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| Válido | Concordo totalmente | 18 | 81,8 | 81,8 | 81,8 |
| | Concordo parcialmente | 1 | 4,5 | 4,5 | 86,4 |
| | Discordo parcialmente | 2 | 9,1 | 9,1 | 95,5 |
| | Discordo totalmente | 1 | 4,5 | 4,5 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

Dezesseis (16) respondentes limitaram-se a responder sobre a necessidade de uma política nacional para a gestão de dados científicos de forma quantitativa na P15. Ou seja, apenas seis respondentes teceram comentários qualitativos na P16, conforme trechos e entrevista a seguir relacionados.

Quadro 22 – Comentários qualitativos sobre a necessidade de uma política para a gestão de dados científicos.

| | |
|---|---|
|  CAPES | *Isso é uma meta imprescindível. Infelizmente, o que temos visto é uma ênfase apenas em números e mesmo assim, temos acompanhado cortes e mais cortes no orçamento de C&T do país.* |
|  FAP | *Mais do que a necessidade de existir uma política, talvez uma lei ou protocolo.* |
|  FAP | *Sem a existência de uma política inexistirão ações concretas por parte das instituições e tão somente ações isoladas.* |

| | |
|---|---|
|  **CNPq** | *(...) uma política nacional voltada para a gestão de dados científicos traria enorme benefício para as pesquisas em andamento e para alimentar pesquisas e projetos futuros de interesse do país.* |

Fonte: a autora.

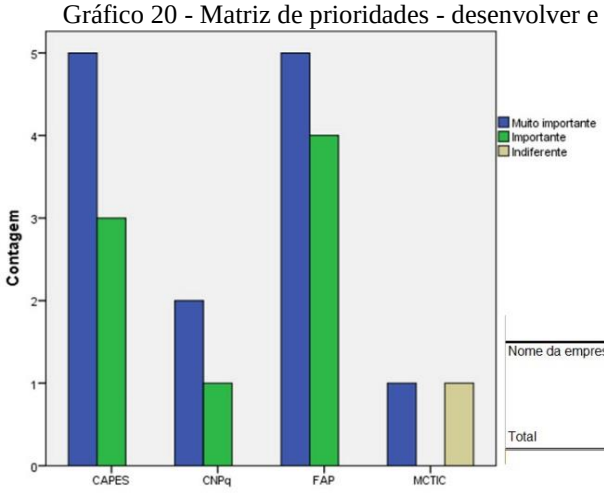
Analisando a questão do ponto de vista proposto por Herrera (1995), ou seja, a formulação de política explícita ou a existência de uma política implícita, os participantes da pesquisa entendem que é importante que o governo dê um direcionamento sobre a gestão de dados científicos. Apenas um participante manifestou apreço pela formulação de uma política explícita, ao colocar que poderia haver a "*necessidade de uma lei ou protocolo"*.

Ao se considerar a publicação do Decreto 6.666 de 2008, referente a INDE, cumpre relembrar que as iniciativas internacionais para tratamento da informação geoespacial iniciaram em 1994, culminando com a publicação da Diretiva 2007/EC no ano de 2007, ou seja, 13 anos após o movimento inicial em nível mundial e 14 anos para a publicação de uma política explícita no Brasil. Em razão do exposto, é emergente que cientistas brasileiros e profissionais da informação envolvidos com a produção e a organização de dados científicos se articulem de forma a dar celeridade ao processo de formulação de uma política nacional para a gestão dos dados científicos.

A pergunta 17 (P17) do instrumento de pesquisa procurou formar a *matriz de atividades prioritárias para se implementar a gestão de dados científicos no Brasil*. Para tanto, foi solicitado que o respondente classificasse o grau de prioridade de iniciativas identificadas como importantes na revisão de literatura. Os dados da matriz são apresentados, de forma individualizada, nos Gráficos 20 a 27. A análise de cada afirmativa da matriz de prioridades está a seguir descrita.

**Afirmativa 1 –** Desenvolver e disponibilizar um repositório de dados científicos.

Gráfico 20 - Matriz de prioridades - desenvolver e disponibilizar um repositório de dados científicos.

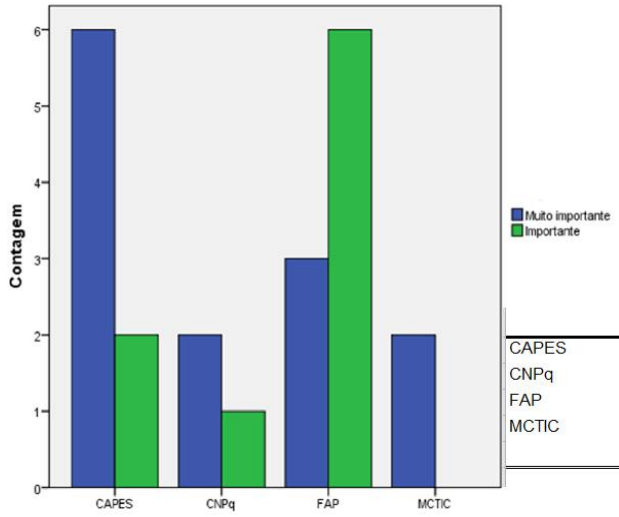



| | | Muito importante | Importante | Indiferente | Total |
|---|---|---|---|---|---|
| Nome da empresa | CAPES | 5 | 3 | 0 | 8 |
| | CNPq | 2 | 1 | 0 | 3 |
| | FAP | 5 | 4 | 0 | 9 |
| | MCTIC | 1 | 0 | 1 | 2 |
| Total | | 13 | 8 | 1 | 22 |

Fonte: a autora.

**Afirmativa 2 –** Desenvolver diretrizes para a coleta, tratamento técnico, armazenamento e preservação dos dados científicos gerados por pesquisas financiadas pelo governo.

Gráfico 21 - Matriz de prioridades – diretrizes para a coleta, tratamento técnico, armazenamento e preservação dos dados científicos.

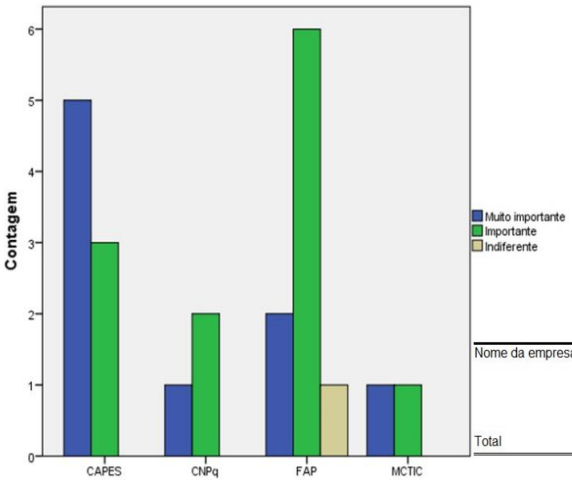



| | Muito importante | Importante | Total |
|---|---|---|---|
| CAPES | 6 | 2 | 8 |
| CNPq | 2 | 1 | 3 |
| FAP | 3 | 6 | 9 |
| MCTIC | 2 | 0 | 2 |
| | 13 | 9 | 22 |

Fonte: a autora.

**Afirmativa 3 –** Desenvolver diretrizes para a reutilização dos dados, para além do contexto inicial em que foram criados, com o objetivo de poupar recursos públicos de financiamento.

Gráfico 22 - Matriz de prioridades – diretrizes para a reutilização dos dados, para além do contexto inicial em que foram criados.

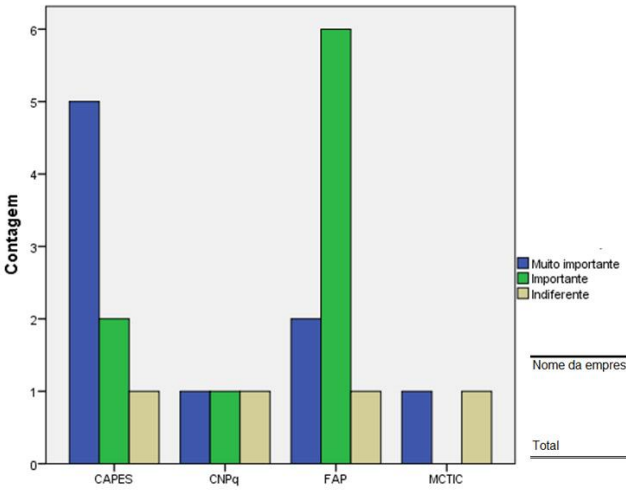



| | | Muito importante | Importante | Indiferente | Total |
|---|---|---|---|---|---|
| Nome da empresa | CAPES | 5 | 3 | 0 | 8 |
| | CNPq | 1 | 2 | 0 | 3 |
| | FAP | 2 | 6 | 1 | 9 |
| | MCTIC | 1 | 1 | 0 | 2 |
| Total | | 9 | 12 | 1 | 22 |

Fonte: a autora.

**Afirmativa 4** – Discutir com a comunidade acadêmica questões relacionadas à propriedade do dado (quem coletou é um técnico de coleta ou é autor do dado? Ou ainda, o dado é de propriedade do governo brasileiro?).

Gráfico 23 - Matriz de prioridades - questões relacionadas à propriedade do dado.

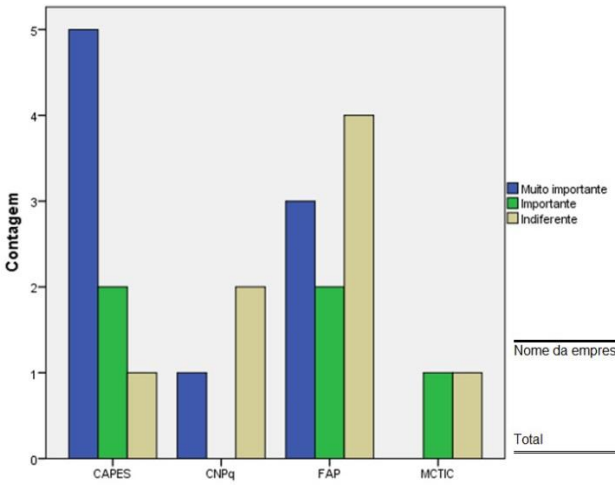



| | | Muito importante | Importante | Indiferente | Total |
|---|---|---|---|---|---|
| Nome da empresa | CAPES | 5 | 2 | 1 | 8 |
| | CNPq | 1 | 1 | 1 | 3 |
| | FAP | 2 | 6 | 1 | 9 |
| | MCTIC | 1 | 0 | 1 | 2 |
| Total | | 9 | 9 | 4 | 22 |

Fonte: a autora.

**Afirmativa 5** – Desenvolver uma tabela de temporalidade para o prazo de carência dos dados (quando o dado pode ser divulgado) por área de conhecimento.

Gráfico 24 - Matriz de prioridades - Desenvolver uma tabela de temporalidade.

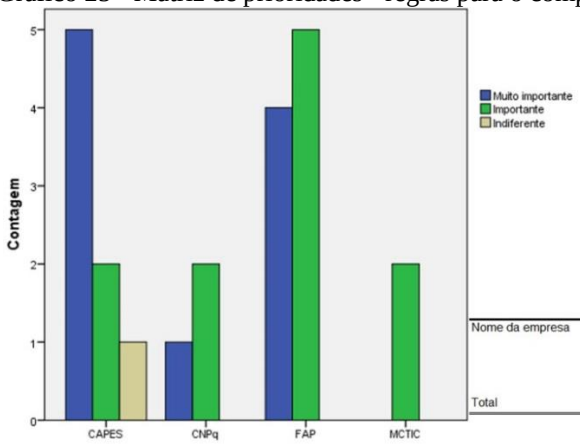



| | | Muito importante | Importante | Indiferente | Total |
|---|---|---|---|---|---|
| Nome da empresa | CAPES | 5 | 2 | 1 | 8 |
| | CNPq | 1 | 0 | 2 | 3 |
| | FAP | 3 | 2 | 4 | 9 |
| | MCTIC | 0 | 1 | 1 | 2 |
| Total | | 9 | 5 | 8 | 22 |

Fonte: a autora.

**Afirmativa 6** – Desenvolver regras para o compartilhamento de dados em nível nacional.

Gráfico 25 - Matriz de prioridades - regras para o compartilhamento de dados em nível nacional.

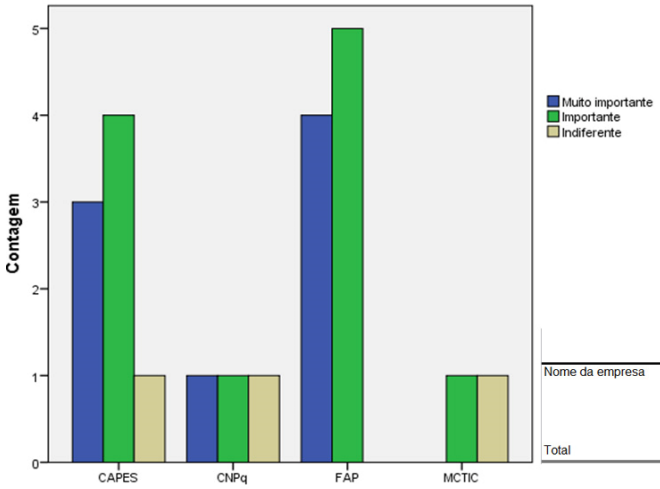



| | | Muito importante | Importante | Indiferente | Total |
|---|---|---|---|---|---|
| Nome da empresa | CAPES | 5 | 2 | 1 | 8 |
| | CNPq | 1 | 2 | 0 | 3 |
| | FAP | 4 | 5 | 0 | 9 |
| | MCTIC | 0 | 2 | 0 | 2 |
| Total | | 10 | 11 | 1 | 22 |

Fonte: a autora.

**Afirmativa 7 -** Desenvolver regras para o compartilhamento de dados em nível internacional.

Gráfico 26 - Matriz de prioridades – regras para o compartilhamento de dados em nível internacional.

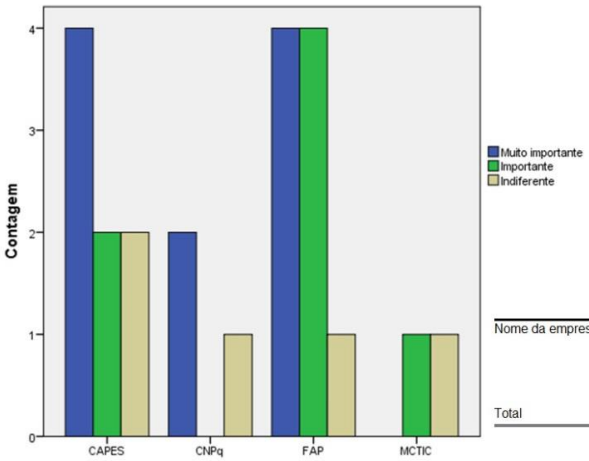



| | | Muito importante | Importante | Indiferente | Total |
|---|---|---|---|---|---|
| Nome da empresa | CAPES | 3 | 4 | 1 | 8 |
| | CNPq | 1 | 1 | 1 | 3 |
| | FAP | 4 | 5 | 0 | 9 |
| | MCTIC | 0 | 1 | 1 | 2 |
| Total | | 8 | 11 | 3 | 22 |

Fonte: a autora.

**Afirmativa 8 –** Desenvolver mecanismos de reconhecimento ao pesquisador que coleta dados (a exemplo do pesquisador que publica artigos).

Gráfico 27 - Matriz de prioridades – mecanismos de reconhecimento ao pesquisador que coleta dados.




| | | Muito importante | Importante | Indiferente | Total |
|---|---|---|---|---|---|
| Nome da empresa | CAPES | 4 | 2 | 2 | 8 |
| | CNPq | 2 | 0 | 1 | 3 |
| | FAP | 4 | 4 | 1 | 9 |
| | MCTIC | 0 | 1 | 1 | 2 |
| Total | | 10 | 7 | 5 | 22 |

Fonte: a autora.

Em função dos dados apresentados nos Gráficos 20 a 27 referentes às afirmativas da matriz de prioridades, a nova configuração da matriz (ordem de prioridade de implementação de acordo com os participantes da pesquisa) configura-se conforme descrito no Quadro 23.

Quadro 23 – Matriz de atividades para a gestão de dados científicos no Brasil.

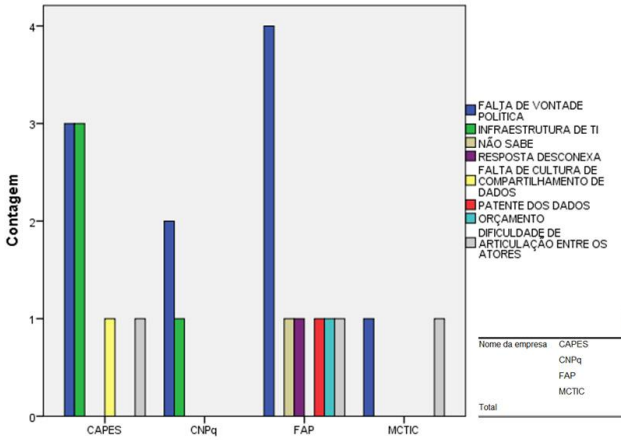
| Nível de Prioridade | Atividade |
|---|---|
| 1 | Desenvolver e disponibilizar um repositório de dados científicos. |
| 2 | Desenvolver diretrizes para a coleta, tratamento técnico, armazenamento, preservação dos dados científicos gerados por pesquisas financiadas pelo governo. |
| 3 | Desenvolver regras para o compartilhamento de dados em nível nacional. |
| 4 | Desenvolver mecanismos de reconhecimento ao pesquisador que coleta dados (a exemplo do pesquisador que publica artigos. |
| 5 | Desenvolver diretrizes para a reutilização dos dados, para além do contexto inicial em que foram criados, com o objetivo de poupar recursos públicos de financiamento. |
| 6 | Discutir com a comunidade acadêmica questões relacionadas à propriedade do dado (quem coletou é um técnico de coleta ou é autor do dado? Ou, ainda, o dado é de propriedade do governo brasileiro? |
| 7 | Desenvolver uma tabela de temporalidade para o prazo de carência dos dados (quando o dado pode ser divulgado) por área de conhecimento. |
| 8 | Desenvolver regras para o compartilhamento de dados em nível internacional. |

Fonte: a autora.

Na América do Norte (EUA e Canadá), observa-se que já há repositórios temáticos para armazenar dados científicos. Consequentemente, entende-se que já existem regras de compartilhamento e citação de dados. O que não foi possível identificar foram as questões de reconhecimento ao pesquisador que produziu o dado primário. As questões inerentes à reutilização dos dados para além do contexto inicial que foram criados também não foram aprofundadas nesta pesquisa. Considerando o exposto, a matriz proposta pelos funcionários das agências de fomento no Brasil mostra-se alinhada ao que já está em andamento no exterior.

Esta pesquisa procurou mapear quais seriam as dificuldades para se implementar uma política nacional dos dados científicos (P18). A pergunta foi feita de forma qualitativa e posteriormente as respostas foram agrupadas em categorias de informação, seguindo, assim, os preceitos da Teoria Fundamentada em Dados. Os funcionários das agências de fomento identificaram como principal motivo a <falta de vontade política>, os demais motivos abaixo se encontram relacionados no Gráfico 28.

Gráfico 28 – Motivos identificados como dificultadores para a implementação de uma política nacional para a gestão de dados científicos.

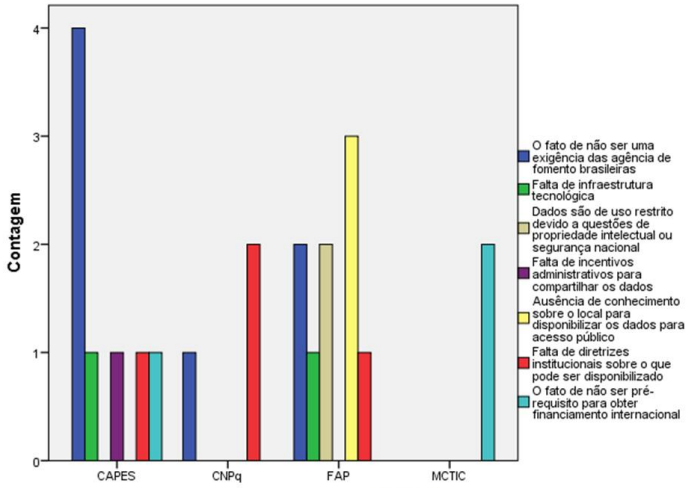



| | | FALTA DE VONTADE POLÍTICA | INFRAESTRUTURA DE TI | NÃO SABE | RESPOSTA DESCONEXA | FALTA DE CULTURA DE COMPARTILHAMENTO DE DADOS | PATENTE DOS DADOS | ORÇAMENTO | DIFICULDADE DE ARTICULAÇÃO ENTRE OS ATORES | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| Nome da empresa | CAPES | 3 | 3 | 0 | 0 | 1 | 0 | 0 | 1 | 8 |
| | CNPq | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | FAP | 4 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 9 |
| | MCTIC | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| Total | | 10 | 4 | 1 | 1 | 1 | 1 | 1 | 3 | 22 |

Fonte: a autora.

Os funcionários de agências de fomento que participaram desta pesquisa identificaram como principal motivo pelo fato do dado científico não estar disponibilizado para a consulta *online* em um repositório de acesso público (P21): o fato de não ser uma exigência das agências de fomento (7), a falta de diretrizes institucionais sobre o que pode ser disponibilizado (4) – em ambos os casos há um reconhecimento da importância de as agências fornecerem uma diretriz a respeito do tema. Na sequência, aparecem como motivo o fato de não ser um pré-requisito para se obter financiamento internacional (3) e a ausência de conhecimento sobre o local onde disponibilizar os dados (3).

Gráfico 29 - Motivos que justificam o dado científico não estar disponibilizado *online*.




| | | O fato de não ser uma exigência das agência de fomento brasileiras | Falta de infraestrutura tecnológica | Dados são de uso restrito devido a questões de propriedade intelectual ou segurança nacional | Falta de incentivos administrativos para compartilhar os dados | Ausência de conhecimento sobre o local para disponibilizar os dados para acesso público | Falta de diretrizes institucionais sobre o que pode ser disponibilizado | O fato de não ser pré-requisito para obter financiamento internacional | Total |
|---|---|---|---|---|---|---|---|---|---|
| Nome da empresa | CAPES | 4 | 1 | 0 | 1 | 0 | 1 | 1 | 8 |
| | CNPq | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 3 |
| | FAP | 2 | 1 | 2 | 0 | 3 | 1 | 0 | 9 |
| | MCTIC | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| Total | | 7 | 2 | 2 | 1 | 3 | 4 | 3 | 22 |

Fonte: a autora.

As respostas a esta pergunta vão ao encontro do que prega a literatura científica sobre o tema. Interessante observar que a literatura revela que a menor dificuldade para se disponibilizar os dados de pesquisa é a questão referente à infraestrutura tecnológica, afinal a capacidade de *hardware* está cada vez maior, assim como a velocidade de processamento, essa é uma das características do *big data*. Em contrapartida, questões inerentes ao aspecto político e humano são as relacionadas pelos autores como as que geram dificuldades para se disponibilizar o dado. Hey e Thefethen (2002) sintetizam essa questão ao afirmarem que o sucesso para os projetos de *e-Science* não envolvem apenas questões técnicas de infraestrutura tecnológica tais como escalabilidade, confiabilidade, interoperabilidade, tolerância a falhas, gerenciamento de recursos, desempenho e segurança. É preciso atenção para questões inerentes às pessoas envolvidas nos projetos tais como a vontade de trabalhar de forma colaborativa, aceitando o compartilhamento de recursos e dados.

Já no que diz respeito aos motivos que incentivam o depósito de dados em um repositório de acesso público, os respondentes identificaram como principal motivo o fato desse ser um requisito da agência de fomento ao apoiar uma pesquisa (sete), na sequência aparece o fato do repositório oferecer tal funcionalidade (seis), em seguida o reconhecimento acadêmico, conforme demonstra a Tabela 22.

Tabela 22 – Motivos que incentivam o depósito de dados em um repositório de acesso público.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Auxílio institucional para o gerenciamento de dados | 2 | 9,1 | 9,1 | 9,1 |
| | Reconhecimento acadêmico para o pesquisador que coletou o dado | 4 | 18,2 | 18,2 | 27,3 |
| | Ser um requisito da agência de fomento | 7 | 31,8 | 31,8 | 59,1 |
| | O repositório oferecer a funcionalidade | 6 | 27,3 | 27,3 | 86,4 |
| | O repositório oferecer o serviço de armazenamento de longo prazo | 1 | 4,5 | 4,5 | 90,9 |
| | O fato de haver um incremento no valor de financiamento para preparar os dados para serem compartilhados | 2 | 9,1 | 9,1 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

O reconhecimento acadêmico foi comentado durante as entrevistas com os pesquisadores (n=40). Nesse aspecto, merece ser comentado o fato de que dois pesquisadores, um da área de engenharia florestal, outro da agricultura manifestaram expressamente o desejo

de que o fato de se disponibilizar um dado para acesso público fosse valorizado tal como um artigo publicado. Pois, o dado coletado também tem um alto valor.

Com o objetivo de exemplificar a preocupação dos pesquisadores brasileiros com a questão, a seguir está transcrito o trecho de uma entrevista com uma pesquisadora da Geração X que atua na área de Engenharia Florestal.

> Na ecologia os dados são muito custosos para serem obtidos. Por exemplo, medir o teor de carbono em folhas para depois encaminhar o material para o laboratório para análise - estimar o grau de carbono é algo muito interessante para ecologia **é difícil de ser feito é valorizado por quem fez mas é difícil compartilhar.** [...] E até o momento não existe nada que ajude esse pesquisador, ou que proteja esse pesquisador que passou muito tempo dentro do campo para coletar esses dados ecológicos mesmo eu sendo uma pessoa super militante no que diz respeito a dar acesso aos dados científicos eu preciso ter consciência desse outro lado. **É preciso valorizar o trabalho realizado por esse pesquisador porque não são dados fáceis de serem coletados** [...] Vou dar um exemplo – a Plataforma Lattes**, o currículo Lattes ele só se preocupa com o artigo, com o livro – mas ele não se preocupa com o dado que foi coletado e esse dado pode ser citado em outra pesquisa**. Então, a **Plataforma Lattes deveria ter um campo de dados produzidos pelo pesquisador, isso já é uma forma de reconhecimento**. Afinal quem coleta o dado é um pesquisador doutor, ou é apenas um peão na coleta de dados? E é justamente o contrário, são pessoas altamente qualificadas, extremamente preocupadas com a qualidade do dado produzido. [grifos do respondente.]

Considerando o processo de publicação científica no Brasil, bem como o processo de publicação científica internacional (a janela de tempo entre a submissão do artigo, o aceite e a publicação do mesmo), foi questionado aos servidores das agências de fomento a opinião sobre qual seria o prazo razoável para o embargo dos dados (período de carência para acesso a totalidade dos dados). A pergunta 25 foi feita em escala Likert, as opções de resposta encontram-se relacionadas na Tabela 23.

Tabela 23 – Prazo para embargo dos dados na visão dos servidores das agências de fomento.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Menos de 1 ano | 4 | 18,2 | 18,2 | 18,2 |
| | Entre 1 e 2 anos | 5 | 22,7 | 22,7 | 40,9 |
| | Entre 3 e 5 anos | 4 | 18,2 | 18,2 | 59,1 |
| | Entre 5 e 8 anos | 2 | 9,1 | 9,1 | 68,2 |
| | Entre 8 e 10 anos | 3 | 13,6 | 13,6 | 81,8 |
| | Não sabe | 4 | 18,2 | 18,2 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

As respostas revelam que os servidores das agências de fomento têm uma expectativa de que o embargo seja de no máximo cinco anos (13 respostas). Porém, quatro respondentes

desejam que o embargo desses dados seja de menos de um ano. Merece ser comentado que tal meta torna-se inexequível, pois o INPI leva aproximadamente oito anos para conceder uma patente e tem como meta reduzir esse prazo para até seis.

Por outro lado, há que se ressaltar que a propriedade intelectual nos EUA é concedida em uma média de dois anos, conforme dados do Escritório de Patente dos Estados Unidos (USPTO, 2017).

O questionamento feito sobre o prazo de embargo dos dados em função de solicitação de registro de patente (P27) foi realizado de forma qualitativa e não obrigatória. Essa pergunta foi desprezada na análise dos dados, pois apenas dois participantes a responderam e a análise de conteúdo das mesmas revelou pouco conhecimento sobre a questão de concessão de patentes. Por outro lado, um participante da amostra de doutores comentou durante a entrevista sobre o tema, em função da relevância do comentário o trecho da entrevista segue transcrito abaixo.

> [...] Sinceramente eu acho muito diferente um embargo de dados de dois anos para um pesquisador de uma faculdade americana em comparação com pesquisador brasileiro. Afinal, o professor e pesquisador brasileiro trabalham em uma universidade sucateada tendo que exercer atividades de pesquisa, bem como atividades administrativas. Na maioria das vezes os NITS não funcionam no auxílio ao registro da propriedade intelectual. Por esse motivo, eu acho que os americanos conseguem ter um embargo de dados de apenas dois anos, mas para nós brasileiros fica praticamente inviável. / Eu acho que o embargo de dois anos é pouco para o pesquisador brasileiro por todos os motivos colocados acima.

Quando questionados sobre o *perfil e as características do cientista de dados* (P28), as respostas sobre o *perfil* se apresentaram dispersas, com uma leve concentração no perfil multidisciplinar (quatro) e no de tecnologia da informação (quatro), conforme demonstra a Tabela 24.

Tabela 24 – Perfil do cientista de dados na percepção dos servidores das agências de fomento.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Política de C&T | 1 | 4,5 | 4,5 | 4,5 |
| | Multidisciplinar | 4 | 18,2 | 18,2 | 22,7 |
| | Não sabe | 4 | 18,2 | 18,2 | 40,9 |
| | Resposta desconexa | 2 | 9,1 | 9,1 | 50,0 |
| | Estatística | 3 | 13,6 | 13,6 | 63,6 |
| | Ciência da Informação | 1 | 4,5 | 4,5 | 68,2 |
| | Conhecimento em C&T | 3 | 13,6 | 13,6 | 81,8 |
| | Tecnologia da Informação | 4 | 18,2 | 18,2 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

Já no que diz respeito às características, as respostas novamente mantiveram-se dispersas, conforme demonstra a Tabela 25.

Tabela 25 – Características do cientista de dados na percepção dos servidores das agências de fomento.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Tratamento de dados científicos | 4 | 18,2 | 18,2 | 18,2 |
| | Não sabe | 4 | 18,2 | 18,2 | 36,4 |
| | Resposta desconexa | 3 | 13,6 | 13,6 | 50,0 |
| | Ciência da Informação | 1 | 4,5 | 4,5 | 54,5 |
| | Raciocínio lógico | 1 | 4,5 | 4,5 | 59,1 |
| | Conhecimento em C&T | 3 | 13,6 | 13,6 | 72,7 |
| | Pragmático | 4 | 18,2 | 18,2 | 90,9 |
| | Curiosidade | 2 | 9,1 | 9,1 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

Quando questionados sobre quem seria o *profissional capacitado* para tratar o dado científico, ou seja, realizar a curadoria de dados (P29), os dados revelam que cinco participantes declararam que não sabem, outros quatro participantes responderam de forma desconexa, logo essas respostas não foram aproveitadas e outros cinco identificaram o profissional como *multidisciplinar* – o que já configura mais da metade das respostas (14 respostas). O profissional de *Ciência da Informação* aparece empatado com o profissional de *Tecnologia da Informação* (dois respondentes para cada categoria), conforme demonstra a Tabela 26.

Tabela 26 – Perfil do profissional capacitado para tratar o dado científico.

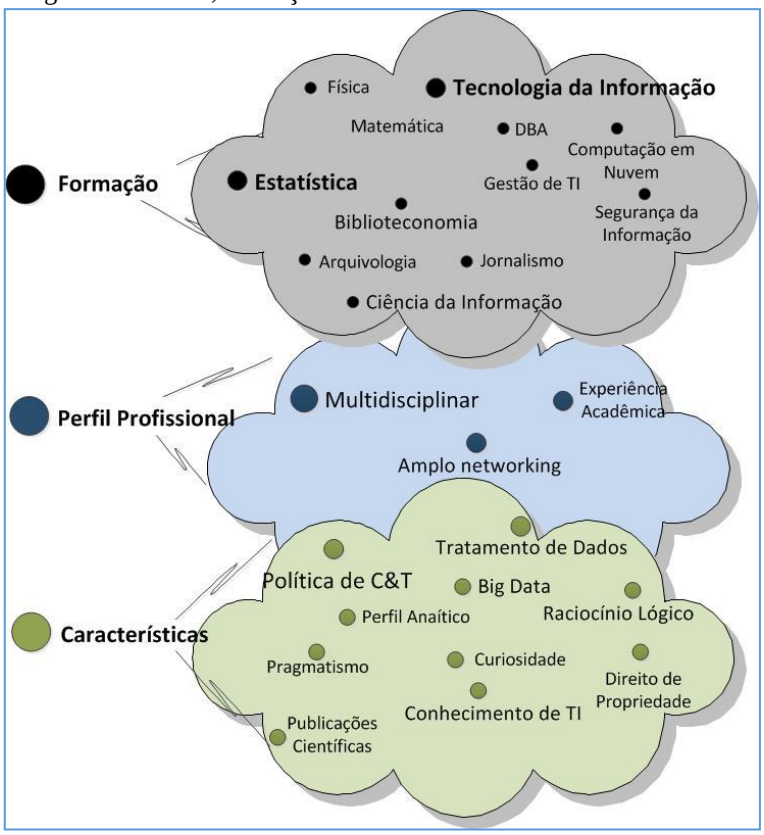
| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Multidisciplinar | 5 | 22,7 | 22,7 | 22,7 |
| | Não sabe | 5 | 22,7 | 22,7 | 45,5 |
| | Resposta desconexa | 4 | 18,2 | 18,2 | 63,6 |
| | Estatística | 1 | 4,5 | 4,5 | 68,2 |
| | Ciência da Informação | 2 | 9,1 | 9,1 | 77,3 |
| | Conhecimento em C&T | 3 | 13,6 | 13,6 | 90,9 |
| | Tecnologia da Informação | 2 | 9,1 | 9,1 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

O que os dados revelam é uma zona de intersecção entre o perfil profissional, as características do cientista de dados e sua formação. Percebe-se uma ênfase na formação na área de tecnologia da informação e estatística, assim como um perfil multidisciplinar com

características voltadas para o *big data*, tratamento de dados e política de C&T, conforme ilustra a Figura 30.

Figura 30 – Perfil, formação e características do cientista de dados.




Fonte: a autora.

Sobre o fato de as universidades brasileiras estarem contribuindo para a formação do profissional de gestão e curadoria de dados científicos (P30), a pergunta foi feita em escala Likert, portanto, sua análise é predominantemente quantitativa. Os dados revelam que nove respondentes marcaram a opção <indiferente> na escala, o que representa praticamente metade das respostas. Outros dez participantes discordam de que a universidade esteja contribuindo para a formação desse profissional. Os resultados coletados estão apresentados na Tabela 27.

Tabela 27 – Universidades brasileiras *versus* formação do profissional de gestão e curadoria de dados.

| | | Frequência | % | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| **Válido** | Concordo totalmente | 1 | 4,5 | 4,5 | 4,5 |
| | Concordo parcialmente | 2 | 9,1 | 9,1 | 13,6 |
| | Indiferente | 9 | 40,9 | 40,9 | 54,5 |
| | Discordo parcialmente | 3 | 13,6 | 13,6 | 68,2 |
| | Discordo totalmente | 7 | 31,8 | 31,8 | 100,0 |
| | **Total** | **22** | **100,0** | **100,0** | |

Fonte: a autora.

## 4.3 ANÁLISE CONJUNTA DOS TERMOS CURADORIA DE DADOS, GESTÃO DE DADOS CIENTÍFICOS E SOBRE A POLÍTICA NACIONAL PARA DADOS CIENTÍFICOS

Um ponto que se mostrou relevante durante a realização desta pesquisa foi compreender o significado dos termos curadoria de dados (P32) e gestão de dados científicos (P33), no questionário aplicado aos servidores das agências de fomento, correspondentes às perguntas P33 e P34 nas entrevistas com os doutores envolvidos com o tema.

As perguntas foram feitas de forma qualitativa e as respostas foram analisadas de forma a extrair categorias de informação. As categorias de informação e a frequência das mesmas foi identificada com o apoio do Nvivo. Após essa etapa, elaborou-se nuvem de *tags* com o objetivo de representar as categorias de informação e a frequência dessas categorias, conforme ilustram os Gráficos 30 e 31.

Gráfico 30 – Nível de entendimento sobre curadoria de dados.


Fonte: a autora.

Os dados revelam que dentre o total de participantes da pesquisa (agências de fomento e doutores/ n=62), 30% acreditam que o termo curadoria de dados corresponde à gestão dos dados. Outros 22% acreditam que se refere a tratamento e armazenamento de dados. Enquanto 11% acreditam que se refere a organização e acesso aos dados. Outros 16% tiveram as respostas divididas entre as categorias <*não sabe*>, <*em branco*> e <*resposta desconexa*>.

Analisando a literatura sobre o tema, o termo curadoria de dados é entendido como a *gestão e a preservação de dados em longo prazo*, incluindo-se nesse contexto o fato de agregar valor aos dados digitais, bem como viabilizar a criação de novos dados, de forma colaborativa, a partir dos já existentes. Além disso, a atividade de curadoria também pode propiciar a redução dos riscos de obsolescência digital (ABOTT, 2008; DIGITAL CURATION CENTRE, 2016; GIARETTA, 2004; HEY; TANSLEY; TOLLE, 2011). Para o Digital Curation Centre (2016), a curadoria digital "envolve a manutenção, a preservação e a agregação de valor aos dados da pesquisa digital em toda sua vida útil.

A análise dos dados desta pesquisa permite afirmar que os termos curadoria de dados e gestão de dados foram utilizados como sinônimo pelos respondentes, o que reflete a literatura internacional sobre o tema.

Gráfico 31 – Nível de entendimento sobre gestão de dados.


Fonte: a autora.

Os dados revelam que dentre o total de participantes da pesquisa (agências de fomento e doutores/ n=62), 38% acreditam que o termo gestão de dados corresponde ao gerenciamento do ciclo de vida do dado. Outros 30% dos participantes entendem que o termo se refere à organização e ao acesso dos dados. Merece ser ressaltado que 20% das respostas dos participantes dividiram-se entre as categorias *<não sabe>*, *<em branco>* e *<resposta desconexa>*.

Em termos de análise qualitativa, vale a pena recordar as práticas de gestão de dados, bem como a visão geral do ciclo de vida do dado científico do DataONE. Em ambos os casos, são citadas as atividades de planejar, coletar, assegurar, descrever, preservar, descobrir, integrar e analisar. Assim, entende-se que os termos "gestão de dados" e "ciclo de vida do dado" são utilizados como sinônimos no Projeto DataONE. Esse entendimento foi o expresso por 38% dos participantes desta pesquisa, o que permite inferir-se que os participantes conhecem a literatura sobre o tema e suas respostas refletem a literatura da área.

Bell (2011), por sua vez, defende que o processo de gestão dos dados consiste em três atividades básicas: captura, curadoria e análise. Este autor, portanto, confirma que curadoria é uma parte da gestão de dados científicos.

Do ponto de vista da administração, Peter Drucker defende que gestão é representada por um conjunto de atividades que procuram atingir objetivos e metas de forma eficiente e eficaz. Para o autor, as atividades de gestão são organizar, planejar, liderar e controlar. Em síntese, para a teoria da administração, a atividade de gestão engloba coordenação e planejamento. Sendo que a atividade de coordenação é representada por um grupo de pessoas e o que elas vão fazer. Já a atividade de planejamento significa o que vai ser feito, quando e como.

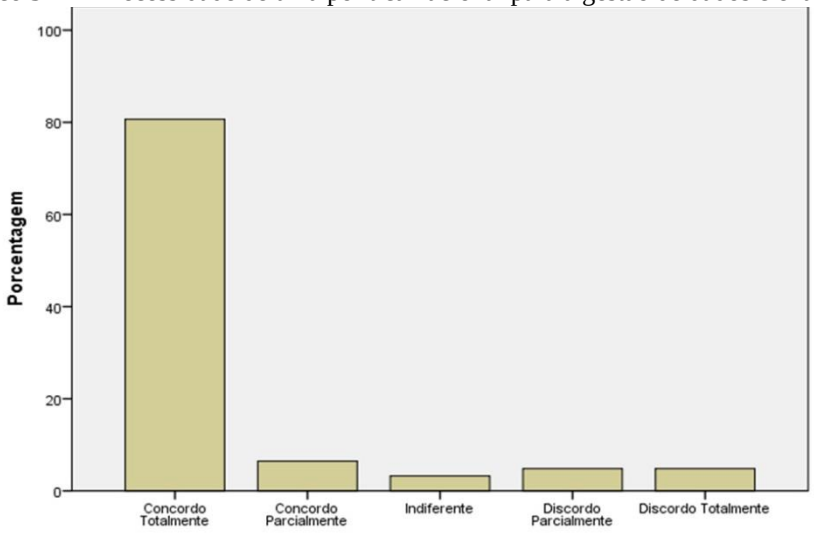
Parece claro que no âmbito dos dados científicos os termos gestão e ciclo de vida do dado estão sendo utilizados como sinônimos na própria literatura internacional e consequentemente entre os cientistas da área no Brasil. Porém, analisando o termo gestão e seu significado de origem, parece coerente afirmar que o gerenciamento do ciclo de vida do dado é apenas uma parte do seu processo de gestão.

Infere-se que em ambos os termos (curadoria de dados, gestão de dados) no âmbito dos dados científicos da *e-Science* a terminologia ainda é incipiente e carece de estudos que procurem padronizar o uso de termos de forma a não se sobreporem em significado, bem como, respeitar a origem (etimologia) dos termos oriundo de outras áreas, a exemplo do termo gestão.

A análise conjunta dos instrumentos sobre o fato do Brasil precisar de uma política nacional para a gestão de dados científicas, revela que 87% dos participantes concordam que tal política é relevante para o país, sendo que 80,6% concordam totalmente e 6,5% concordam

parcialmente. Apenas 9,6% dos participantes discordam que a elaboração de uma política nacional seja relevante, conforme demonstram os dados do Gráfico 32.

Gráfico 32 – A necessidade de uma política nacional para a gestão de dados científicos.




**O Brasil precisa de uma politica nacional de gestão de dados científicos**

| | | Frequência | Porcentual | Porcentagem válida | Porcentagem acumulativa |
|---|---|---|---|---|---|
| Válido | Concordo Totalmente | 50 | 80,6 | 80,6 | 80,6 |
| | Concordo Parcialmente | 4 | 6,5 | 6,5 | 87,1 |
| | Indiferente | 2 | 3,2 | 3,2 | 90,3 |
| | Discordo Parcialmente | 3 | 4,8 | 4,8 | 95,2 |
| | Discordo Totalmente | 3 | 4,8 | 4,8 | 100,0 |
| | Total | 62 | 100,0 | 100,0 | |

Fonte: a autora.

Dentre os participantes da amostra de doutores que responderam qualitativamente a questão, foram citados como documentos com diretrizes para gestão de dados científicos o material produzido pela NSF, Dataone, Research Data Aliance, Digital Curation Center e Rainforts. No âmbito nacional, foram citados o Portal da Biodiversidade, o PPBIO, PELD, SISBIO, o Guia de Gestão de Dados de Sayão e Sales (2015), a Política do Museu Emílio Goeldi e a Política do Jardim Botânico do Rio de Janeiro. Por outro lado, apenas um participante de agência de fomento citou qualitativamente a política do PPBIO, o que permite inferir que os participantes da amostra de fomento desconhecem as poucas políticas para tratamento de dados científicos que existem no Brasil.

Em face do exposto, é possível afirmar que as áreas de ponta no Brasil em gestão de dados científicos são a geoespacial, a área de meio ambiente, ecologia e biodiversidade. As políticas explícitas, desenvolvidas ainda que de forma incipiente e identificadas durante a realização da tese, foram as relacionadas na Figura 31 com o intuito de apresentar um marco legal em política de dados científicos no Brasil.

Figura 31 – Políticas de Dados Científicos vigentes no Brasil em 2017.


Fonte: a autora com fundamento na literatura revisada e coleta de dados por meio de entrevista e questionário.

Ao retomar-se a literatura sobre gestão de dados científicos, denominada por Shearer de política de RDM (*Research Data Management*), vale a pena destacar que para a autora os detalhes das políticas variam entre regiões, agências e domínios, mas também têm uma série de coisas em comum. Os componentes de políticas mais frequentes são requisitos em torno de padrões e metadados, compartilhamento de dados e retenção de dados e /ou preservação de longo prazo. Geralmente, os planos de gerenciamento de dados são exigidos no contexto dessas políticas, pois obrigam os pesquisadores a pensar sobre como eles administrarão seus dados antes do projeto, um requisito fundamental para boas práticas de gerenciamento de dados. As políticas também contêm consistentemente disposições para a proteção da confidencialidade, propriedade intelectual e dados sensíveis.

Estudos conduzidos por Shearer demonstram que o cenário mundial para a RDM é de que os pesquisadores estão preocupados com o tempo, conhecimento e recursos envolvidos na preparação dos dados. As instituições de pesquisa, por sua vez, estão preocupadas com a forma como irão financiar serviços de apoio à gestão de dados e infraestruturas. Nesse aspecto, é

válido ressaltar que as preocupações dos pesquisadores brasileiros e dos funcionários das agências de fomento parecem estar alinhadas à percepção de Shearer.

Durante a pesquisa, foi possível perceber que as poucas políticas de gestão de dados científicos no Brasil foram desenvolvidas em função da maturidade internacional da área de pesquisa no tema, bem como em função da necessidade dos pesquisadores brasileiros se alinharem às diretrizes internacionais desse tipo de informação, tanto para obterem financiamento internacional, como para compartilharem seus dados em repositórios internacionais – dando visibilidade à sua pesquisa, ou até mesmo para viabilizar a publicação de um artigo em periódico internacional, fato a ser comentado com mais profundidade no capítulo 4.4 referente à Teoria Fundamentada em Dados.

## 4.4 A TEORIA FUNDAMENTADA EM DADOS: PROPOSTA DE *FRAMEWORK* DE DIRETRIZES PARA A ELABORAÇÃO DE UMA POLÍTICA DE GESTÃO DE DADOS CIENTÍFICOS

Com fundamento na literatura pesquisada, bem como, na coleta de dados entre os 62 participantes desta pesquisa, abaixo segue uma síntese qualitativa da análise dos dados e um *framework* com itens considerados de extrema relevância para a elaboração de um conjunto de diretrizes que venham a servir de elementos norteadores para a elaboração de uma política para a gestão de dados científicos no Brasil.

Uma análise integrada das respostas concedidas pelos diferentes instrumentos de coleta de dados permite afirmar-se que os respondentes acham a atividade de gestão de dados científicos relevante para o país. Além disso, informaram que a criação de um repositório de dados é uma atividade de extrema importância, mas não há um consenso se esse repositório deve ser centralizado, a exemplo da BDTD, ou se deve ser assumido pelas instituições responsáveis pelas pesquisas. É possível afirmar que os respondentes próximos ao Governo Federal (CAPES, CNPq, unidades de pesquisa do MCTI, universidades etc.) acreditam que o IBICT é a instituição que tem capacidade técnica de gerenciar dados de pesquisa. Apesar dessa opinião, os respondentes informaram que a condição de sucateamento que o instituto tem enfrentado é prejudicial para assumir a gestão de dados científicos. Por outro lado, respondentes das FAP acreditam que a gestão de dados no país deve ser assumida pela CAPES ou pelo CNPq.

As respostas dos participantes das agências de fomento tendem a considerar que a principal dificuldade para se implementar a gestão de dados refere-se à infraestrutura tecnológica. Em contrapartida, os doutores envolvidos com o tema afirmaram que a principal

dificuldade é intrínseca à natureza humana (querer disponibilizar e compartilhar o dado). Assim, é possível afirmar-se, pela literatura revisada, que as dificuldades para implementar uma política de gestão de dados científicos estão realmente atreladas ao aspecto humano do compartilhamento de dados, portando em consonância com as respostas da amostra de doutores desta pesquisa. Infere-se que os participantes da amostra de doutores já têm certa maturidade no compartilhamento de dados e suas dificuldades. Com fundamento no exposto, conclui-se que a política que deve estabelecer um modelo de gestão de dados para a infraestrutura. Ou seja, a integração de dados deve ser refletida à luz da política de gestão de dados. A respeito do assunto, merece ser destacado o trecho de uma entrevista com um pesquisador brasileiro, por sintetizar bem a questão do aspecto humano versus o tecnológico na política para a gestão de dados.

> No que diz respeito à política para a gestão de dados na instituição, como disse anteriormente, temos apenas iniciativas isoladas. Mas, de qualquer forma, a política não tem que estar preocupada com aspectos tecnológicos, mas sim com aspectos conceituais de como o Brasil vai conduzir suas pesquisas, produzir seus dados, quais os critérios de armazenamento de dados, temporalidade dentre outros aspectos.

Ainda no que diz respeito às questões inerentes ao compartilhamento de dados, os participantes da amostra de doutores entendem que as agências de fomentos ainda não compreendem o fenômeno de ciência colaborativa presente na pesquisa do Século XXI. Eles desejam que as agências estimulem as atividades de citação e compartilhamento do dado, a exemplo de como são valorizadas as citações de artigo.

Um fator que foi percebido na coleta de dados desta pesquisa e que não foi constatado na revisão de literatura refere-se ao orçamento e à infraestrutura tecnológica necessária ao registro do DOI para dados. Infere-se que essa é uma realidade de países de Terceiro Mundo, onde, assim como os cortes de governo afetam as coleções das bibliotecas, a ausência de uma rubrica específica para viabilizar o registro de DOI para dados científicos impacta no desejo de algumas instituições iniciarem tal atividade. Dentre as instituições que manifestaram preocupação com o fato de não terem um orçamento específico para tal finalidade estão o IEN e o IBICT. A partir da revisão de literatura, do contexto histórico no país e da análise documental das instituições envolvidas nesta pesquisa, entende-se que o IBICT deve centralizar a atividade de registro de DOI para dados no Brasil. Como atividade semelhante do instituto cita-se a centralização do registro de ISSN para todos os periódicos nacionais, ou mesmo a atividade de registro no LOCKSS de artigos de periódicos nacionais.

As instituições, que de alguma forma já iniciaram atividades que envolvem a gestão de dados científicos, revelam que o processo não é *top down*, pelo contrário, ele se inicia a partir da necessidade real de pesquisadores, que por sua vez vão se articulando e movimentando as instâncias superiores. De certa forma, o *Relatório Atkins* (ATKINS *et al*., 2003) revela o mesmo processo nos EUA, ou seja, um movimento de pesquisadores frente à National Science Foundation para buscar verbas que auxiliem no desenvolvimento de uma infraestrutura para novas práticas de pesquisa, a grande questão é que esse movimento nos EUA começou em 2003, portanto há quatorze anos.

A respeito do assunto, os dados da pesquisa revelam que no Brasil ainda há muita incerteza sobre de quem é o papel de legislar e de quem é o papel de executar a lei no âmbito da gestão de dados científicos. Não foram poucas as respostas que indicavam o IBICT como formulador de política pública, ou mesmo a RNP. Porém, em termos de estrutura de funcionamento do Governo Federal é importante ressaltar que esses institutos executam políticas públicas e podem movimentar a sociedade acadêmica por meio de manifestos, cartas abertas, mas não legislar. O processo de legislar cabe ao Congresso Nacional. Em face do exposto, em função da matéria objeto deste estudo, seria prudente um movimento dos cientistas da Sociedade Brasileira de Progresso da Ciência, ou mesmo da Comissão Nacional de Ciência e Tecnologia junto aos parlamentares que levassem a questão à discussão. Esse é um processo comum em ambas as casas legislativas. A exemplo, cita-se o caso do Manifesto de Extensão Tecnológica, ocorrido em 16 de agosto de 2011, liderado pelo Deputado Ariosto Holanda[119], que tinha como objetivo discutir a necessidade de "*aplicar nas empresas o conhecimento gerado nas universidades e nas escolas técnicas do país*" (PSB-40, 2016). Dentre os frutos do Seminário e do Manifesto sobre Extensão Tecnológica, foi criado em 2013 o Fórum Nacional de Defesa da Extensão Tecnológica (FNET), liderado pela deputada Luciana Santos[120] que defenderá a consolidação de uma Rede Nacional de Extensão Tecnológica em até 180 dias. Outro exemplo pertinente foi a realização da Audiência Pública sobre a Empresa Brasileira de Pesquisa e Inovação Industrial, ocorrida em 21/08/2012, com a presença do então Ministro Aluísio Mercadante, quando foram debatidos temas como risco tecnológico, *scale up* (escalonamento de produtos), e até mesmo o posicionamento expresso do ex-Ministro de Ciência e Tecnologia – Sérgio Rezende, que firmou apoio ao SIBRATEC[121] e que a EMBRAPII

[119] Deputado pelo PSB-CE.
[120] Deputada pelo PCdoB-PE, ex-prefeita de Olinda e ex-secretária de Ciência, Tecnologia e Meio Ambiente de Pernambuco.
[121] SIBRATEC – Sistema Brasileiro de Tecnologia.

deve ser uma continuidade do outro programa, dentre outros aspectos vitais para colocar no mercado produtos oriundos de pesquisas científicas.

Os exemplos acima citados revelam a necessidade de as casas legislativas promoverem seminários, manifestos, audiências públicas e outros instrumentos que permitam a discussão e o amadurecimento das questões inerentes à gestão de dados científicos no país em conjunto com a Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática. A partir desse amadurecimento, os parlamentares terão subsídios para desenvolver, por meio de uma lei, um norte para pesquisadores, funcionários de agências de fomento e para a comunidade científica como um todo. Assim, entende-se que as bases para a elaboração de diretrizes de uma política para gestão de dados científicos, seja ela explícita ou implícita, devem partir das instâncias legisladoras. Porém, se o Poder Legislativo não se manifestar de forma emergente, caberá ao Poder Executivo se manifestar por meio de um decreto, a exemplo do Decreto n. 6.666 de 2008, que trata da informação geoespacial no Brasil – INDE.

No que diz respeito a elaboração de um manifesto em prol da necessidade de gestão dos dados científicos, merece ser comentado que, em 28 de setembro de 2016, o IBICT lançou o *Manifesto de Acesso Aberto a Dados da Pesquisa Brasileira para Ciência Cidadã*[122]. Ressalta-se que nesse documento o instituto "estende a sua visão sobre o acesso aberto, e reconhece os dados de pesquisa como um recurso imprescindível para as ações de Ciência Aberta, Ciência para todos, Ciência Cidadã" (IBICT, 2017). O documento em questão trata de dados de pesquisa para uma ciência aberta, ou seja, uma visão mais ampla do que os dados coletados em grande escala oriundos da *e-science*. Por outro lado, há que se ressaltar que o manifesto tem um caráter tão extenso que pouco elucida as questões inerentes ao dilúvio de dados, da ciberinfraestrutura necessária ao armazenamento e preservação dos dados e, tão pouco, as questões intrínsecas ao compartilhamento de dados primários. Ao se dirigir para um público tão amplo (pesquisadores, universidades, institutos de pesquisa, agências de fomento, sociedade científica, editores de periódicos científicos, cursos de pós-graduação em Ciência da Informação, gestores e executores de programas e projetos de dados de pesquisa) o documento parece ter perdido a sua eficácia argumentativa e de persuasão ao público ao qual foi dirigido.

A exemplo do *Manifesto de Extensão Tecnológica* acima citado, um manifesto, em sua essência, é um instrumento democrático, que tem como objetivo convencer alguma coisa a alguém, caracterizando assim um texto do gênero argumentativo. Dentre as suas características, há uma análise da situação-problema, uma argumentação a respeito da situação, bem como as

[122] Disponível para consulta em - http://www.ibict.br/Sala-de-Imprensa/noticias/2016/ibict-lanca-manifesto-de-acesso-aberto-a-dados-da-pesquisa-brasileira-para-ciencia-cidada.

possíveis soluções. Além disso, o documento é datado e assinado por um conjunto de pessoas ou representantes de instituições. Nesse sentido, o documento lançado pelo IBICT parece não deixar claro qual a situação-problema, não apresenta possíveis soluções e tão pouco é direcionado há uma instancia que tenha capacidade de solucionar a questão.

No âmbito da necessidade de diretrizes para a gestão de dados científicos no Brasil, o recomendável seria um manifesto que elucidasse as questões inerentes ao dilúvio de dados e à infraestrutura tecnológica necessária ao tratamento, organização, preservação de longo prazo e compartilhamento de dados. Além disso, entende-se que no âmbito da informação científica e tecnológica, para ganhar força política, o documento deveria ter sido assinado também por representantes, entre outros, da Academia Brasileira de Ciências, Associação Brasileira das Instituições de Pesquisa Tecnológica e Inovação, Associação Brasileira dos Reitores das Universidades Estaduais e Municipais, Associação Nacional dos Dirigentes de Instituições Federais de Ensino Superior, Conselho Nacional das Fundações Estaduais de Amparo à Pesquisa, Conselho Nacional de Secretários Estaduais para Assuntos de Ciência e Tecnologia, Fórum Nacional de Secretários Municipais da Área de Ciência e Tecnologia, Sociedade Brasileira para o Progresso da Ciência (SBPC) e, finalmente, o próprio IBICT.

Ademais, entende-se que o documento deveria ter sido direcionado ao Ministro de Ciência, Tecnologia, Inovações e Comunicações, ou, pelo menos, ao Conselho Nacional de Ciência e Tecnologia, ou à Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática da Câmara dos Deputados e do Senado Federal.

Para iniciar a discussão sobre a elaboração de diretrizes para uma política de gestão de dados científicos, recomenda-se que sejam envolvidos os diferentes atores de Ciência e Tecnologia em nível altamente estratégico, como, por exemplo, os ministérios, as universidades, os institutos de pesquisa, trazendo para a discussão a necessidade real de pesquisadores, como, por exemplo, os que trabalham com dados espaciais (AEB, INPE), ou dados sobre a biodiversidade (Mar e Antártica, Museu Emílio Goeldi, INPA), ou ainda as necessidade inerentes ao tratamento de dados de energia nuclear (CNEN, IEN). Além disso, a discussão também deve abordar aspectos de infraestrutura computacional para transmissão e armazenamento de dados (RNP, LNCC), infraestrutura para tratar o ciclo documental do dado de pesquisa e suas particularidades (IBICT).

Os dados revelam que o sucesso de implementação de uma política interministerial para a gestão de dados científicos dependerá das agências de fomento (CAPES, CNPq, FAP) ao exigir o depósito de dados coletados em um repositório e ao criarem mecanismos de estímulo ao depósito do dado, como, por exemplo, o peso de reconhecimento da citação do dado. Nas

palavras de um pesquisador entrevistado – "*as universidades só vão se mexer quando uma instância superior disser faça*".

O cenário exposto nesta tese coloca em evidência o tema dados científicos e sua complexidade. Considerando que o Reino Unido iniciou os programas de *e-Science* em 2001 e, os EUA em 2003, portanto, há pelo menos 15 anos, o Brasil encontra-se extremamente atrasado no que diz respeito tanto ao entendimento das agências de fomento sobre o tema, bem como quanto ao envolvimento dos pesquisadores para com a complexidade da questão. O Brasil até o início de 2017 carecia de uma política explícita, em nível federal, que norteasse as ações do Estado em termos de gestão e preservação dos dados científicos, bem como diretrizes para reutilização dos dados em questão. O fato é que o Brasil precisa se posicionar quanto à necessidade de acesso aberto aos dados de pesquisas financiadas por agências de fomento brasileiras.

Apesar de haver apenas uma política explícita em nível federal – referente à informação geoespacial, estabelecida no Decreto nº6.666 de 2008, algumas instituições de pesquisa, ainda que de forma embrionária, têm usado sua autonomia para desenvolver políticas locais que atendam aos editais de fomento internacionais, bem como às necessidades de diretrizes quanto ao armazenamento, à preservação e reutilização de dados, a exemplo, cita-se as instituições ICMBIO, Museu Emilio Goeldi e Jardim Botânico do Rio de Janeiro. Além dessas, há politicas desenvolvidas especificamente para alguns programas de biodiversidade, em função de regras impostas ao recebimento de apoio internacional. São exemplos dessa situação a política do Portal Brasileiro da Biodiversidade, as dos programas PELD e PPBIO.

Entende-se que a gestão de dados científicos é um projeto de grande e complexa envergadura nacional. À exemplo do ocorrido nos EUA e no Reino Unido, é necessária a destinação de um orçamento para viabilizar o desenvolvimento de uma infraestrutura tecnológica de compartilhamento e armazenamento de dados. Sugere-se que essas atividades inerentes à tecnologia da informação sejam assumidas, ou pelo menos realizadas com o apoio da RNP e do LNCC por já possuírem um *background* nessas atividades. Por outro lado, em função da natureza de sua missão institucional, recomenda-se que sejam atribuídas ao IBICT as atividades vinculadas à organização e gestão da informação.

A execução da política de armazenamento de dados e preservação, inexoravelmente, deve estar ligada às agências de fomento que apoiam as pesquisas. Para tanto, recomenda-se que as respectivas agências exijam um plano de gestão de dados a qualquer pesquisador que almejar recurso público. Por outro lado, o papel das universidades não pode ser subestimado. A exemplo do processo de organização e preservação das teses e dissertações produzidas no

Brasil, as universidades e institutos de pesquisa brasileiros precisam dividir a responsabilidade de organização desses dados em parceria com o IBICT e as próprias agências de fomento. Enquanto isso, RNP e LNCC atuam em parceria no fornecimento da infraestrutura tecnológica necessária para esse processo.

Entende-se que uma instituição sozinha não terá condições de abraçar todo o processo de gestão de dados científicos no Brasil em função da sua complexidade, dimensão territorial, dificuldades orçamentárias, escassez de recursos humanos, dentre tantos outros fatores. Nesse cenário, a formação de uma espécie de consórcio para a gestão de dados científicos parece profícua no sentido de aproveitar o que cada instituição tem de melhor nessas áreas. Assim, o desafio que se apresenta é a articulação de um consórcio com um mínimo de três grandes instituições (IBICT, RNP, LNCC), que atuam em diferentes estados, para que funcionem de forma sinérgica.

Uma atividade essencial desse processo parece ser a necessidade de decisão quanto ao formato dos repositórios, se seria institucional ou temático. Essa pesquisa não consegue dar uma resposta satisfatória para a questão em razão da diversidade das respostas e complexidade da questão. Se por um lado os repositórios de teses e dissertações foram desenvolvidos nas universidades, tendo o IBICT à frente da coleta de metadados e gestão de uma interface de busca, o cenário para os dados científicos se apresenta um pouco mais complexo. Um dos motivos é a própria infraestrutura tecnológica, pois, enquanto as teses e dissertações são armazenadas em sistema de gerenciamento de banco de dados relacional, os dados científicos não obedecem a essas regras. Os próprios *workflows* científicos também possuem uma estrutura diferenciada para armazenar esses dados. Em termos de *big data*, a literatura já indica um novo tipo de banco de dados (o Hadoop) com novas linguagens de programação, capacidade de armazenamento etc. Nesse cenário, como as universidades brasileiras, sucateadas em termos de recursos humanos, infraestrutura física e tecnológica, terão capacidade de atender essa demanda? Por outro lado, o desenvolvimento de um repositório temático também não se mostra de fácil escolha. Pois, vamos supor que a RNP e o LNCC compartilhem essa responsabilidade em termos de armazenamento e suporte ao volume de troca de dados, essas instituições certamente precisarão do apoio do IBICT no desenvolvimento de uma metodologia para definição de metadados de forma a viabilizar um *harvesting* eficiente dos mesmos à determinada área do conhecimento. Ao centralizar-se o armazenamento por área temática infere-se que serão minimizados os problemas de recursos humanos, por outro lado, há que se investir de forma mais severa em termos de infraestrutura tecnológica. Além disso, um *software* de coleta de metadados deverá ser disponibilizado para as universidades.

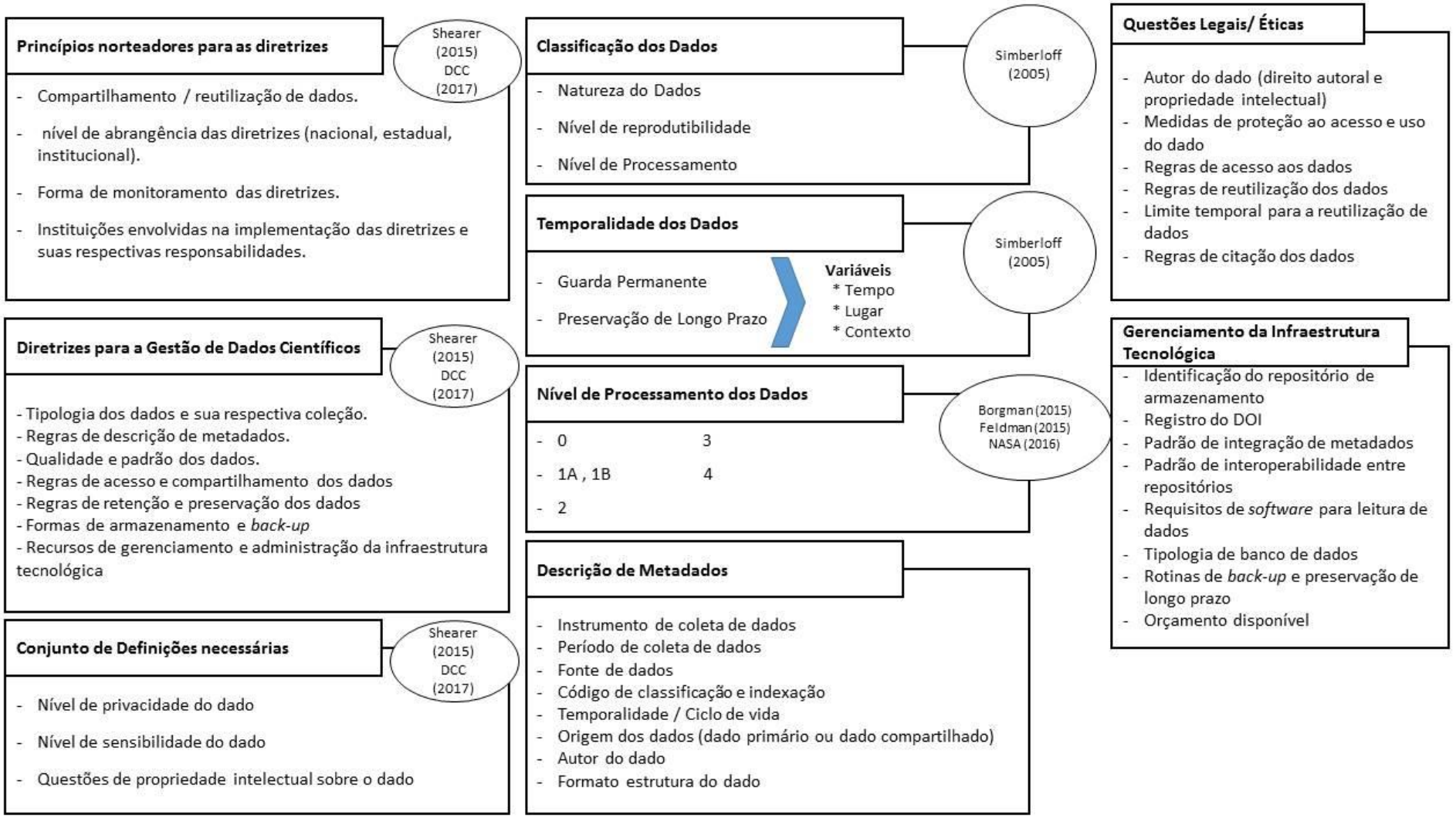
Todos os cenários fatalmente carecem de investimento de recurso público, o que torna temerário o sucesso do projeto, tendo em vista que de acordo com informações publicadas no *Diário Oficial da União* de 23/03/2017, a verba disponível para o empenho do Ministério da Ciência, Tecnologia, Inovações e Comunicações caiu 44% na comparação com o recurso inicialmente previsto na LOA, de R$ 5,049 bilhões, em 2017. Em resumo, os recursos, que já eram praticamente a metade dos cerca de R$ 10 bilhões registrados em 2013, agora se apresentam como um dos piores da história da ciência, tecnologia e inovação.

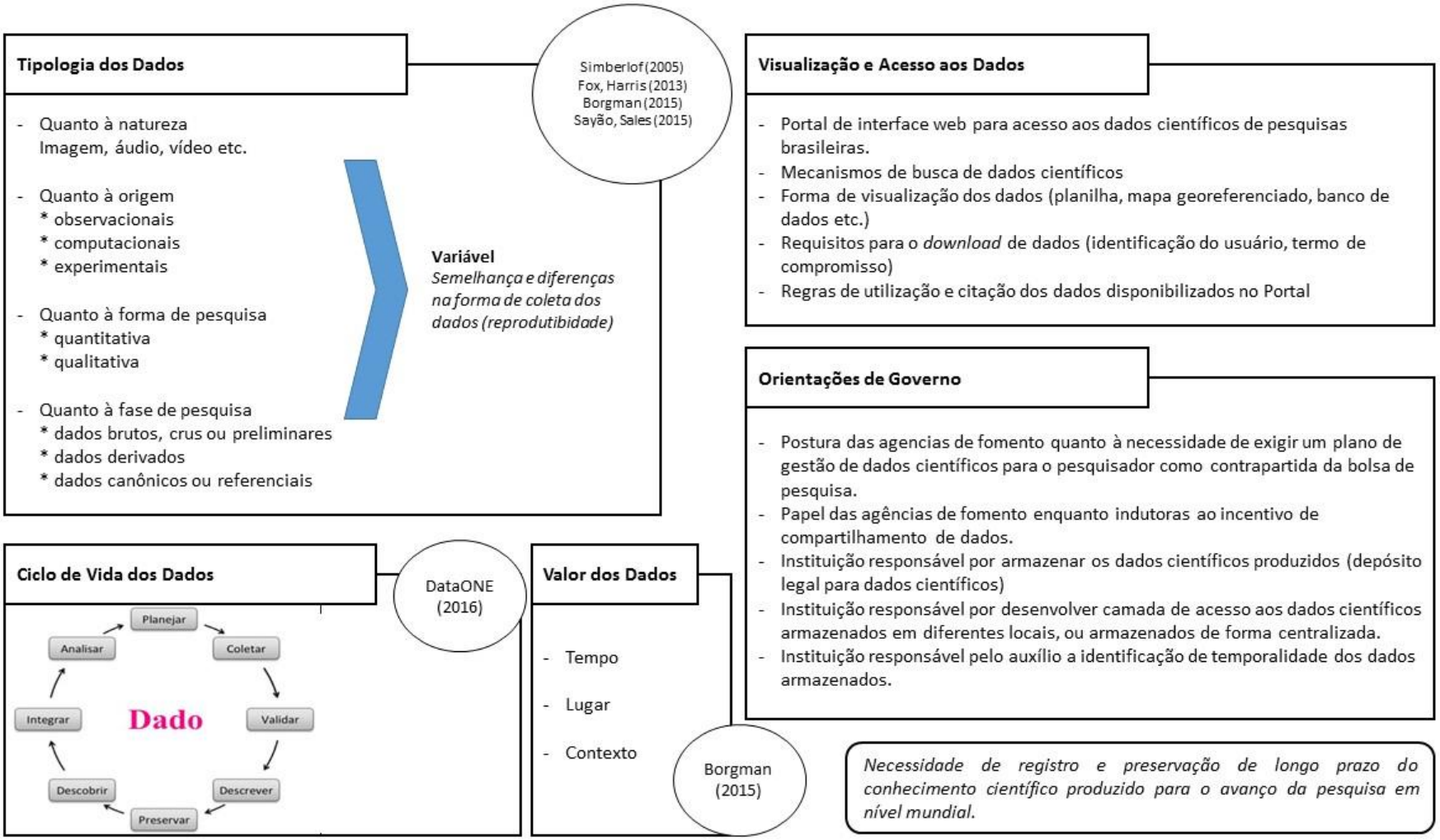
Além do corte no orçamento, não se pode deixar de afirmar que os poucos recursos disponíveis serão duramente disputados para a execução de novas pesquisas e conclusão de outras anteriormente iniciadas. Mas, em um país onde pouco se valoriza a memória e suas bibliotecas, o que esperar desse orçamento para investimento em infraestrutura de armazenamento e compartilhamento de dados? Apenas uma ação em nível altamente estratégico poderá modificar o cenário obscuro em que a gestão de dados científicos no Brasil se encontra. Nesse sentido, entende-se que o Centro de Gestão e Estudos Estratégicos (CGEE) vinculado ao MCTIC pode contribuir para mapear as áreas temáticas que têm prioridade para o Governo Federal no tratamento de dados científicos e, em conjunto com o IBICT, RNP e LNCC, apresentar uma agenda positiva de ações necessárias para os próximos dez anos. Os resultados e diretrizes apontadas nesse estudo poderiam se fazer contar na próxima Estratégia Nacional de Ciência e Tecnologia (antigamente publicado como Plano de Ação de Ciência e Tecnologia). Esta tese já apresenta pequenas contribuições ao estudo ao propor uma matriz de ações prioritárias na percepção dos próprios servidores de agências de fomento.

Esta tese atingiu seu objetivo geral de pesquisa quando se propôs a desenvolver um conjunto de diretrizes para a elaboração de uma política nacional para a gestão de dados científicos no Brasil. Para alcançar esse objetivo geral, foram delimitados cinco objetivos específicos, alcançados por meio de diferentes técnicas de pesquisa, já relacionadas no capítulo de metodologia e sintetizadas na Figura 25 (ver capítulo 3 – Metodologia).

Como resultado da revisão de literatura, bem como da análise de dados da pesquisa de campo, foi desenvolvido o *framework* com as diretrizes para a elaboração de uma política de gestão de dados. No próprio *framework*, constante na Figura 32, estão relacionados os autores que trouxeram contribuições relevantes para essas diretrizes.

Figura 32 - *Framework* com diretrizes para a elaboração de uma política de gestão de dados científicos






Fonte: a autora com fundamento na literatura revisada e nos dados coletados para esta pesquisa.

Àqueles que se dedicarem a formular uma política para a gestão de dados científicos, faz-se necessário esclarecer que é imprescindível iniciar esse processo com uma campanha de conscientização para os pesquisadores que sejam alvo da política. Afinal, de pouco adianta ter um normativo legal que não tenha objetivo punitivo, se o seu público alvo não estiver consciente da importância do mesmo. A literatura deixa evidente que se o pesquisador não tiver consciência sobre a importância do seu dado coletado, ele pouco contribuirá em termos de descrição dos metadados da coleção de dados; consequentemente, os profissionais da informação terão dificuldade para realizar o tratamento técnico desse dado, armazená-lo, preservá-lo e recuperá-lo, o que trará inevitavelmente dificuldades para compartilhar o dado posteriormente.

Em síntese, se não houver uma campanha de conscientização dos pesquisadores sobre a importância de os dados científicos estarem abertos em prol do próprio avanço da ciência, talvez a única maneira de viabilizar a criação de repositórios de dados científicos de pesquisa seja por meio da adoção de medidas que repreendam àquele pesquisador que não compartilhar seu dado. Porém, a questão mostra-se complexa, afinal que tipo de medida pode repreender um pesquisador doutor, líder em determinada área do conhecimento? A literatura analisada indica que apenas a suspensão de benefícios financeiros, tais como, bolsa de pesquisa, verbas de custeio, patrocínio, subsídio dentre outros. Por outro lado, há que se considerar que para aplicar uma medida tão extrema~~,~~ faz-se necessário que o pesquisador assine um termo de compromisso com o patrocinador de sua pesquisa, informando que disponibilizará seus dados em determinado período. Consequentemente, a instituição que elaborar tal termo de compromisso com o pesquisador, precisará implementar formas de avaliação e controle, sob o risco de o termo de compromisso não ser respeitado. Por fim, é vital entender que o pesquisador só disponibilizará seus dados ao final do seu processo de pesquisa. Logo, ele já terá usufruído dos benefícios concedidos para seu projeto de pesquisa. Assim, resta refletir – qual a instituição que irá se dispor a criar mecanismos rígidos de controle e avaliação de forma a subsidiar a decisão de glosar do pesquisador benefícios anteriormente concedidos? Em função dos motivos acima expostos, não há dúvidas de que o melhor caminho é o da conscientização. Importante ressaltar que as instituições envolvidas com a promoção da ciência aberta no Brasil são àquelas aptas a realizar em eventos técnicos e científicos a conscientização do pesquisador a respeito do tema.

Ao se considerar os princípios norteadores da política, um dos primeiros aspectos para se considerar é que é preciso que a política se posicione quanto ao seu nível de abrangência. Por exemplo, se for uma política institucional, faz-se necessário refletir se ela irá atingir apenas as pesquisas produzidas no âmbito institucional ou, se também contemplará pesquisas produzidas por redes de pesquisadores presentes em diferentes instituições. É essencial levar em conta que as dificuldades para a obtenção de recursos financeiros no âmbito científico e tecnológico, bem como, a própria complexidade da pesquisa culmina com a formação de redes de pesquisadores. Outro ponto

a ser considerado é se a pesquisa desenvolvida institucionalmente obteve auxílio de agência de fomento (nacional ou internacional), empresa privada e até mesmo de recursos advindos de prêmios obtidos pelo pesquisador. Assim, é fundamental que a política para a gestão de dados científicos contemple o intrincado processo de colaboração dos pesquisadores e as variadas formas de obtenção de financiamento para melhor delimitar seu escopo de atuação e nível de abrangência.

Para a melhor compreensão daqueles que precisarão seguir a política de gestão de dados científicos, é importante que seja definido pela política o que é dado científico de pesquisa, coleção de dado e até mesmo o que não se configura como dado. Da mesma forma, mostra-se pertinente contextualizar em quais circunstancias o dado poderá ser compartilhado, bem como, em qual cenário não será permitido o seu compartilhamento. Por exemplo, uma pesquisa na área de fármaco que culmine com desenvolvimento tecnológico e a produção de uma droga específica, certamente não poderá ter seus dados primários compartilhados por um determinado período. Caberá a política contextualizar quais são essas exceções e qual o prazo do embargo dos dados.

Assim como no planejamento de um sistema de informações tem-se os requisitos funcionais e os desejáveis, ao se elaborar uma política de gestão de dados científicos é preciso identificar quais são os requisitos essenciais à política e, quais são os que são desejáveis. Por exemplo, no seu escopo de abrangência é essencial, a definição do que é dado e coleção de dados também é essencial. Por outro lado, a temporalidade de armazenamento, as regras de compartilhamento ou mesmo as regras de citação dos dados são requisitos desejáveis, mas não imprescindíveis à formulação da política.

No que diz respeito à temporalidade de armazenamento, a principal variável desse fator é o grau de dificuldade para se coletar e reproduzir o dado científico. Há sentido em preservar por longo prazo dados relativamente fáceis e por vezes com baixo custo para serem reproduzidos? A partir do momento em que a política der diretrizes sobre o que deve ser armazenado por longo prazo, definindo a sua temporalidade, é que se desenhará a infraestrutura tecnológica necessária ao armazenamento e preservação dos dados científicos e, consequentemente o valor a ser investido nessa infraestrutura.

# CONCLUSÕES

Enquanto o conhecimento pode ser considerado o motor do desenvolvimento científico e tecnológico, os dados científicos são o combustível que coloca esse motor em funcionamento. Por meio dessa metáfora entende-se a relevância do tema desta tese para a sociedade brasileira. Nessa perspectiva, este trabalho apresenta uma reflexão sobre a atual estrutura dos dados científicos no Brasil e como a gestão desses tem evoluído. Para tanto, desenhou-se como objetivo geral – identificar as ações de governo para a gestão de dados científicos em países desenvolvidos, de forma a viabilizar a elaboração de um conjunto de diretrizes para a gestão de dados científicos no Brasil.

Em função dos resultados do estudo bibliométrico e da análise da literatura sobre o tema, pode-se afirmar que, incontestavelmente, os países que estão mais avançados em termos de política para a gestão de dados científicos são o Reino Unido e os Estados Unidos. Depois verificam-se ações de peso no Canada, na Austrália, Espanha, em Portugal e nos países que são signatários da "Declaração da OCDE sobre o Acesso aos Dados de Pesquisa do Financiamento Público" de 2014. Aqui já cabe a reflexão a respeito do silêncio do governo brasileiro sobre o tema e sobre a sua ausência na participação de iniciativas internacionais.

Percebe-se que as diferentes políticas variam entre os países e, consequentemente, nas organizações em termos de força, cobertura, funções, responsabilidades e requisitos, mas, de um modo geral, conforme visto no capítulo de revisão de literatura, elas estão preocupadas com: a) os tipos de dados cobertos pela política; b) as expectativas para compartilhamento de dados, incluindo acesso e prazos; c) os períodos mínimos de retenção de dados (embargo); d) o uso de metadados e padrões de documentação; e) as isenções justificadas à partilha de dados; f) os custos associados à gestão de dados que podem ser pagos por meio de subvenções; e g) o reconhecimento de criadores de dados. Em alguns casos, as políticas são adotadas com pouco ou nenhum monitoramento quanto ao seu cumprimento. Em outros casos, são anexadas a propostas e ao plano de gestão de dados e essas são submetidas a uma revisão clara por comitês de revisão por pares, mas com pouco ou nenhum acompanhamento ao final do projeto. Também há situações em que a conformidade com a política é revista no final de um projeto, uma vez que existem sanções em caso de descumprimento.

Observa-se que o Decreto nº 6.666 /2008 contempla a única iniciativa explícita em nível nacional para a gestão de dados científicos. Esta norma legal refere-se aos dados geoespaciais. Ela determina que são obrigatórios o compartilhamento e a disseminação desses dados e seus respectivos metadados. Da mesma forma, estabelece que em situações em que os dados

comprometam a segurança da sociedade e do Estado, os mesmos não devem ser compartilhados. Constata-se que o Decreto em questão não regula o prazo de embargo dos dados, tão pouco ajusta a conduta sobre o responsável pela criação do dado. Por outro lado, atribui à Comissão Nacional de Cartografia (CONCAR) a responsabilidade sobre a implementação da infraestrutura tecnológica que viabilize o armazenamento e o acesso aos dados geoespaciais, bem como o monitoramento da política.

A literatura internacional indica que as políticas de gestão de dados científicos têm como objetivo melhorar a eficiência da pesquisa, apoiar a reutilização de dados para novos *insights* e novas descobertas, promover uma maior transparência e fomentar a colaboração entre pesquisadores. Consequentemente, para atingir esses objetivos, os dados científicos devem ser adequadamente gerenciados ao longo do seu ciclo de vida. A partir do exposto, entende-se que o Brasil se quiser participar do desenvolvimento de pesquisa de ponta e compartilhar seus dados com outros pesquisadores, em nível nacional e internacional, precisa aceitar a necessidade de uma política que norteie as ações dos cientistas quanto ao gerenciamento de seus dados de pesquisa.

Para atingir o segundo objetivo específico desse estudo foi realizada a análise das ações de governo de países desenvolvidos sobre a gestão de dados científicos nos países identificados. Assim, foram priorizadas na revisão de literatura as análises dos documentos sobre EUA e Reino Unido, o que levou à identificação de importantes documentos sobre o tema, a exemplo do *Relatório Atkins* (2003) e do histórico do e-Science Core Program. Nesse aspecto, merece ser ressaltado que o movimento no exterior foi *bottom-up*, ou seja, partiu da necessidade da comunidade de cientistas em direção ao governo. A partir de então, os respectivos governos desenvolveram programas para atender a demanda da comunidade de cientistas. Acrescenta-se que esse cenário no Brasil mostra-se confuso, pois, apesar de já terem ocorrido diversos eventos relacionados ao tema não foi assinada nenhuma carta de intenções, ou mesmo um manifesto para ser entregue ao Ministro de Ciência, Tecnologia, Inovações e Comunicações, ou à Comissão Nacional de Ciência e Tecnologia, ou à Comissão de Ciência, Tecnologia, Inovação, Comunicação e Informática. Tal situação revela uma profunda desarticulação dos cientistas envolvidos com a problemática dos dados científicos coletados em grande escala e, até mesmo, do IBICT, enquanto órgão responsável pelo tratamento da informação científica e tecnológica e líder do movimento de acesso aberto à informação científica no Brasil, haja vista que o próprio manifesto elaborado pelo instituto não foi direcionado à nenhuma das instâncias acima mencionadas.

A desarticulação entre os pesquisadores e, até mesmo, a inexistência de uma instituição capitaneando as atividades para a gestão de dados científicos foi constatada por meio das entrevistas. Tal situação deve ser remediada por meio da criação de comunidades para a troca de ideias, alinhamento de esforços e expectativas. Assim, evita-se duplicidade de iniciativas com o mesmo objetivo, consequentemente, otimiza a utilização dos poucos recursos disponíveis para a execução de projetos de C&T.

Quanto ao terceiro objetivo específico, este identificou os principais problemas e as soluções inerentes à construção de uma política estruturada para a gestão de dados científicos. Nesse ponto, chegou-se à conclusão de que a maior dificuldade de implementação está relacionada ao aspecto humano, ou seja, ao interesse do pesquisador querer estabelecer um ciclo de vida para os seus dados, assim como querer compartilhar esses mesmos dados. Nessa perspectiva, Corrêa (2016) é enfático ao afirmar que poucos pesquisadores realmente se preocupam com o registro e a preservação de seus dados científicos; de um modo geral eles mantêm apenas aquilo que necessitam. A respeito do tema, a literatura revela que a menor das dificuldades se refere ao aspecto de infraestrutura tecnológica, o que se mostra coerente em tempos de *big data*.

A postura das agências de fomento no Brasil com relação ao tema foi objeto de análise do quarto objetivo específico da tese. De um modo geral, uma análise qualitativa aprofundada de todas as respostas revela o desconhecimento dos entrevistados para com a problemática da gestão dos dados científicos. A análise integrada das respostas concedidas por pesquisadores que precisam de uma política para a gestão de seus dados, bem como as respostas coletadas junto aos funcionários das agências de fomento sobre o sistema de gestão de dados, o sistema de informação gerencial, as diretrizes da agência a respeito da coleta, o tratamento técnico, o armazenamento e a preservação de dados científicos, dentre outras questões, permite afirmar que os funcionários das agências de fomento que participaram da pesquisa mostraram-se confusos entre dados de gerenciamento e controle de projetos de pesquisa financiados pelas agências *versus* dados brutos científicos coletados pelos pesquisadores. Essa situação é no mínimo preocupante, uma vez que a implementação de uma política de dados científicos precisa do envolvimento das agências de fomento no Brasil.

Outro aspecto relevante é o fato de que as agências concordam que é um peso para elas assumirem a função de gerenciar os dados científicos da pesquisa por elas financiadas. Da mesma forma, há um consenso de que a pesquisa, financiada com dinheiro público, deve ter seus dados de acesso público. Elas apontaram como principal instituição capaz para realizar a missão de tratar esses dados científicos o IBICT, sendo essa uma decisão coerente em função

do próprio instituto já ter lastro na realização de tais atividades no âmbito da informação bibliográfica. A diferença, agora, é que o Instituto, frente aos desafios da pesquisa do Século XXI e da *e-science*, precisa assumir uma postura proativa na gestão de dados científicos com iniciativas semelhantes aos projetos de informação bibliográfica da BDTD e da Rede Cariniana. Porém, as agências ressaltaram as questões inerentes ao sucateamento pelo qual o instituto tem enfrentado na última década como um fator dificultador para o IBICT assumir a liderança desse processo.

O quinto objetivo específico desta tese se propôs a identificar o posicionamento dos pesquisadores brasileiros envolvidos com a gestão de dados científicos. Para tanto, foram entrevistados quarenta pesquisadores de diversas áreas do conhecimento e de diferentes gerações. O perfil demográfico dos entrevistados permite afirmar que não há uma diferença de comportamento quanto ao compartilhamento de dados em relação à geração à qual o pesquisador pertence, a maioria manifestou interesse em compartilhar seus dados, bem como ter acesso aos dados compartilhados de outras pesquisas. Porém, os que conseguiram ter acesso a dados brutos, de outras pesquisas, enfatizaram que o acesso se deu em função da proximidade do grupo de pesquisa, ou mesmo do pesquisador líder – prática já conhecida na comunidade científica, mas que não é prática de *e-science* – dados brutos disponibilizados em um repositório de dados científicos para qualquer pesquisador.

Um fato interessante, referente ao acesso a dados de outra pesquisa, que esta tese trouxe é que os dados revelam uma tendência de que os pesquisadores das Ciências Sociais Aplicadas, bem como, das Ciências Exatas e da Terra acreditam na autenticidade dos dados compartilhados. Por outro lado, os pesquisadores das Ciências Agrárias tendem a não confiar nesta autenticidade, pelo menos na amostra desta pesquisa. Ressalta-se que por tratar-se de uma amostra não probabilística, formada pelo critério de intencionalidade, esses dados não devem ser extrapolados. Eles servem apenas como um alerta para se observar o comportamento desses pesquisadores.

Os pesquisadores não estão familiarizados com *softwares* de *workflow científico*, sendo que apenas três participantes declararam conhecer tais *softwares*, sendo eles o Kepler, o Taverna e o YAWL. Eles também ainda não implementaram o ciclo de vida para os seus dados de pesquisa, tão pouco têm um *software* que dê acesso a esses dados. Também foi constatado que há laboratórios de ponta no Brasil que ainda têm os dados de pesquisa registrados em cadernos pessoais, prática não recomendada em função da necessidade de preservação dos dados a longo prazo e do próprio compartilhamento de dados.

Uma das conclusões desta tese é a de que os termos curadoria de dados, gestão de dados e ciclo de vida dos dados foram utilizados como sinônimo pelos respondentes, o que reflete a literatura internacional sobre o tema. Porém, analisando-se o termo gestão e seu significado de origem, parece coerente afirmar que o gerenciamento do ciclo de vida do dado é apenas uma parte do seu processo de gestão. Assim, recomenda-se um estudo terminológico para esses termos-chave inerentes ao processo de gestão de dados científicos, de forma a contribuir para o amadurecimento teórico da área que futuramente será refletida em políticas de indexação e recuperação de documentos em bases de dados.

Como vimos, todos os objetivos específicos propostos por esta pesquisa foram atingidos, viabilizando-se dessa forma o alcance do objetivo geral que visou elaborar um esboço de diretrizes para a gestão de dados científicos no Brasil. Nesse sentido, uma análise integrada das respostas concedidas pelos diferentes instrumentos de coleta de dados permite afirmar-se que os respondentes acham a atividade de gestão de dados científicos relevante para o país. Além disso, a criação de um repositório de dados é uma atividade de extrema importância, mas as respostas obtidas revelam que não há um consenso se esse repositório deve ser centralizado, a exemplo da iniciativa de tratamento da informação bibliográfica realizada pela BDTD/IBICT, ou se o repositório deve ser desenvolvido e assumido pelas instituições responsáveis pelas respectivas pesquisas. É possível registrar que os respondentes próximos ao Governo Federal (CAPES, CNPq, unidades de pesquisa do MCTI, universidades etc.) acreditam que o IBICT é a instituição que tem capacidade técnica de gerenciar dados de pesquisa. Por outro lado, respondentes das FAP acreditam que a gestão de dados no país deve ser assumida pela CAPES ou pelo CNPq.

As respostas dos participantes das agências de fomento tendem a considerar que a principal dificuldade para se implementar a gestão de dados refere-se à infraestrutura tecnológica. Em contrapartida, os doutores envolvidos com o tema afirmaram que a principal dificuldade é intrínseca à natureza humana (querer disponibilizar e compartilhar o dado). Como já comentado na conclusão desta tese, é possível reiterar, pela literatura revisada, que as dificuldades para implementar uma política de gestão de dados científicos estão realmente atreladas ao aspecto humano do compartilhamento de dados, portando em consonância com as respostas da amostra de doutores desta pesquisa. Assim, infere-se que os participantes da amostra de doutores já têm certa maturidade no compartilhamento de dados e suas dificuldades. Por conseguinte, as respostas dos funcionários das agências de fomento revelam pouco conhecimento sobre os dados científicos e os aspectos relacionados à sua gestão, fato inquietante, posto que os dados desta pesquisa revelam que o sucesso da implementação de uma

política interministerial para a gestão de dados científicos dependerá das agências de fomento (CAPES, CNPq, FAP) ao exigir o depósito dos dados coletados em um repositório e ao criarem mecanismos de estímulo ao depósito do dado, como, por exemplo, o peso de reconhecimento da citação do dado.

Apesar do considerável avanço em relação ao Brasil, conforme já visto na revisão de literatura, há estudos que relatam algumas dificuldades enfrentadas por cientistas de países de Primeiro Mundo no compartilhamento de dados. Da mesma forma há estudos que indicam que esses mesmos cientistas não receberam treinamento em gestão de dados e possuem receio de compartilhar seus dados em função da disputa de recursos para financiamento, além de argumentarem que não há incentivo para compartilhar seu respectivo dado. Essas são dificuldades intrínsecas ao gerenciamento de dados científicos presentes em países de primeiro mundo, que iniciaram esse processo há aproximadamente quinze anos. Assim, é perfeitamente compreensível que os cientistas brasileiros ainda enfrentem tanta dificuldade para gerenciar e para compartilhar, visto que eles sequer possuem uma diretriz norteadora do Governo Federal, da sua instituição, ou mesmo da agência de fomento que apoia a sua pesquisa.

Dessa forma, é possível afirmar que as áreas de ponta no Brasil em gestão de dados científicos são a geoespacial, a do meio ambiente, a de ecologia e a de biodiversidade, pois elas possuem uma política, ainda que incipiente, para a gestão de dados. Essas poucas políticas de gestão de dados científicos foram desenvolvidas em função da maturidade internacional da área de pesquisa no tema, bem como em função da necessidade dos pesquisadores brasileiros se alinharem às diretrizes internacionais desse tipo de informação, tanto para obterem financiamento internacional, como para compartilharem seus dados em repositórios internacionais – dando visibilidade à sua pesquisa – ou até mesmo para viabilizar a publicação de um artigo em periódico internacional.

A literatura revisada e os dados desta tese revelam, inquestionavelmente, que as políticas não deveriam ser adotadas isoladamente; pouco adianta uma instituição ou programa ter uma política e outra não. Tem-se como fator preponderante para esta anomalia, a insegurança jurídica a qual o pesquisador fica sujeito, pois há situações em que, dentro de uma mesma instituição, há diretrizes explicitas para as pesquisas conduzidas com o apoio de determinados programas de financiamento e há total omissão em relação aos dados das pesquisas conduzidas com outras formas de financiamento. Tal fato pode ser observado, por exemplo, em instituições que fazem parte do programa PELD e PPBIO. As pesquisas conduzidas por instituições que fazem parte desses programas possuem clareza quanto à forma de depósito e compartilhamento dos dados. Entretanto, as pesquisas realizadas dentro desse

mesmo órgão que não se inserem nos programas citados, coexiste uma área cinzenta, sem clareza jurídica quanto aos direitos e deveres dos pesquisadores no âmbito da gestão de dados científicos. Em função disso, o Brasil carece de um amplo e uniformizado posicionamento para a gestão de dados científicos em todas as suas instituições de pesquisa.

Outra faceta dessa insegurança jurídica se relaciona com os direitos autorais dos resultados da pesquisa e dos seus dados primários. Sem adentrar na densa discussão de autoria dos dados científicos, aspecto não contemplado nos objetivos desta tese, é vital que se clarifique os direitos dos dados coletados e de sua análise. Serão esses direitos similares aos contemplados pela legislação dos direitos autorais ou seguirão os ditames da legislação sobre a propriedade intelectual? A resposta para essa vital indagação demanda maiores estudos por parte dos órgãos governamentais, das sociedades científicas, agências de fomento e, claro, dos pesquisadores das variadas áreas temáticas. De posse de um documento conclusivo das partes envolvidas é mister que seja elaborado um projeto de lei que vise dar a necessária segurança jurídica para a coleta, depósito, compartilhamento, preservação e divulgação dos dados gerados por pesquisas científicas no contexto brasileiro.

As boas práticas de gestão de dados de investigação dependerão de múltiplos fatores contribuintes, incluindo incentivos, conhecimentos especializados, serviços e infraestruturas, bem como mecanismos de financiamento adequados. Os países que optaram por investir na gestão e no compartilhamento de dados científicos descobriram que, embora a conformidade total de uma política não possa ser imediatamente implementada, uma diretriz governamental ajuda muito na conscientização da gestão de dados científicos. Assim, é preciso reconhecer que uma política nacional levará alguns anos para ser plenamente implementada em nosso país. Afinal, a dimensão territorial e as profundas diferenças entre as regiões do país dificultam um alinhamento de ações e evidenciam que as ações de gestão de dados científicos devem ser implementadas de forma constante e gradual.

Em se retomando o desenvolvimento da política de desenvolvimento científico e tecnológico no Brasil, é imperativo relembrar que o setor sempre sofreu com a descontinuidade de iniciativas e com o corte de orçamentos. Durante o governo de Fernando Henrique Cardoso foram criados os Fundos Setoriais com o objetivo de se desenvolver condições mais estáveis para o financiamento público de iniciativas para a ciência e tecnologia e evitar assim as restrições fiscais que o setor enfrentou durante as décadas de 1970, 1980 e 1990 que culminaram com o sucateamento das instituições de pesquisa. Porém, a realidade da ciência e tecnologia no governo Michel Temer é que a grave crise política e econômica que assola o país provocou, em 2017, o contingenciamento de 44% dos recursos de C&T, o que causou espanto e preocupação

na comunidade científica, pois tal medida provocou dentre tantos outros prejuízos, situações alarmantes tais como: a notícia de que a falta de recursos levaria a RNP a suspender o serviço de internet para 20 universidades, a notícia de que o CNPq não teria recursos em 2017 para cumprir com seus compromissos já assumidos, dentre eles o pagamento de quase 100 mil bolsistas de Iniciação Científica, de Pós-Graduação e de Pesquisa, a notícia de que o Museu Emílio Goeldi deixaria de funcionar, dentre tantas outras. Essa situação apenas corrobora o que a revisão de literatura afirma sobre o setor de C&T – a constante descontinuidade de aplicação de recursos. Ou seja, nem a criação dos Fundos Setoriais melhorou a situação.

Diante do cenário exposto, não é exagero afirmar que pouco será feito no âmbito de investimento para a gestão de dados científicos no Brasil nos próximos quatro anos, afinal, se cientistas estão batalhando para a obtenção de recursos que viabilizem a conclusão de suas pesquisas, se as associações de C&T têm lutado para que recursos sejam repassados para viabilizar o funcionamento de institutos de pesquisa, o que esperar no âmbito de atitudes que envolvem gestão e preservação de dados em um país que não valoriza sua informação bibliográfica, sua memória institucional etc.?

Se na esfera da política de desenvolvimento científico e tecnológico o futuro próximo não se mostra promissor para a gestão de dados científicos, o que esperar no âmbito da política de informação? A revisão de literatura sobre o tema já revelou que há propostas para políticas de informação no Brasil, porém elas não têm sido prioridade para o governo e na maioria das vezes não saem do papel.

Nesse aspecto, é mister relembrar que a política de informação no Brasil está imbricada com a política de desenvolvimento científico e tecnológico em função da criação na década de 1950 do IBBD, hoje IBICT. Ao se considerar os dados obtidos nesta pesquisa sobre a relevância do IBICT, em função do histórico de sua trajetória no tratamento da informação científica e tecnológica, é preciso advertir que o Instituto, enquanto instituição de pesquisa vinculada ao MCTIC, também tem enfrentado cortes orçamentários que culminaram com o seu sucateamento. Consequentemente, suas iniciativas para a gestão de dados científicos ainda são tímidas em face das necessidades impostas à envergadura de uma política nacional. Desse modo, infere-se que apenas pequenas[123] iniciativas institucionais e a postura individual dos pesquisadores, ao procurarem melhor gerenciar seus dados de forma a viabilizar sua preservação e compartilhamento, podem contribuir para um cenário mais otimista na gestão de dados científicos no Brasil pelos próximos quatro anos.

[123] Pequenas em função da restrição orçamentária pela qual as instituições de pesquisa têm passado, bem como, por serem inciativas de âmbito institucional.

Dentre as recomendações para estudos futuros, é importante retomar as iniciativas do Portal Brasileiro de Dados Espaciais e o Portal Brasileiro da Biodiversidade. Recomenda-se a realização de um estudo de caso sobre ambos os portais, incluindo a necessidade de investigar as dificuldades encontradas para disponibilizar os dados de forma *online* (tratamento técnico, padronização dos dados disponibilizados e questões de infraestrutura tecnológica), bem como fazer uma avaliação sobre a utilização desses portais, identificando o seu usuário, às suas necessidades de informação e as dificuldades de uso dos portais. Acredita-se que esse estudo poderá contribuir tanto para o aprimoramento dos sistemas hoje disponíveis, quanto como para servir de elemento norteador ao desenvolvimento de novas iniciativas no Brasil.

Além do exposto, recomenda-se que seja realizada a avaliação do plano de gestão de dados desenvolvido pelos pesquisadores que foram aprovados nos editais de fomento para projetos *e-science* da FAPESP. O que esse pesquisador entende como plano de gestão de dados está em consonância com a literatura sobre o tema? O plano de gestão de dados foi implementado? Como a FAPESP administrou esses dados? Os resultados dessa avaliação servirão de elementos norteadores para futuros editais de pesquisa que envolvam a gestão de dados científicos no Brasil.

Nas palavras de Shearer (2015)

> gerenciar dados é muito mais do que apoiar a excelência em pesquisa. Os dados digitais são a matéria prima da economia do conhecimento e estão se tornando cada vez mais importantes para todas as áreas da sociedade. Políticas, serviços e infraestrutura devem estar em vigor se quisermos aproveitar esta crescente maré de dados.

# REFERÊNCIAS


ABEL, Ricardo. **Sem ciência e tecnologia não haverá erradicação da pobreza, diz Dilma**. dez. 2012. Disponível em: <http://www.ebiotecnologia.org/2012/12/sem-ciencia-e-tecnologia-nao-havera.html>. Acesso em: 25 nov. 2016.

AGUIAR, Afrânio Carvalho. Coordenação de uma rede nacional de informação em ciência e tecnologia: um plano prioritário do IBICT. **Ciência da Informação**, v. 9, n. 1/2, p. 83-88, 1980.

ALFONSO-GOLDFARB, Ana Maria. **O que é história da ciência**. São Paulo: Brasiliense, 1994 (Coleção primeiros passos; 286).

ALISSON, Elton. FAPESP apresenta programa de pesquisas e-Science. **Agencia Notícias**, 31 mar. 2014. Disponível em: <http://agencia.fapesp.br/fapesp_apresenta_programa_de_pesquisas_em_escience/18833/>. Acesso em: 6 mar. 2017.

ALVARO, Elsa *et al.* E-science librarianship: field undefined. **Issues in Science & Technology Librarianship**, n. 66, p. 28-43, Summer 2011.

AMARAL, Ana Maria Barros Mais do. O cenário da política nacional de informação no Brasil. **Informação & Sociedade**, v. 1, n. 1, p. 47-53, jan./dez. 1991.

ANDRADE, Maria Eugenia Albino; OLIVEIRA, Marlene. A ciência da informação no Brasil. In: OLIVEIRA, Marlene (Coord.). **Ciência da informação e biblioteconomia**: novos conteúdos e espaços de atualização. Belo Horizonte: UFMG, 2005.

ARAÚJO, Bruno Cesar *et al*. **Impacto dos fundos setoriais nas empresas**. Brasília: IPEA, 2012. (Textos para discussão, n. 1737).

ATKINS, Daniel E. *et al*. **Revolutionizing science and engineering through cyberinfrastructure**: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington, DC, Jan. 2003. Disponível em: <http://www.nsf.gov/cise/sci/reports/atkins.pdf>. Acesso em: 10 abr. 2015.

AUN, Marta Pinheiro. A construção de políticas nacional e supranacional de informação: desafio para os Estados nacionais e blocos regionais. **Ciência da Informação**, v. 28, n. 2, 1999.

BARROS, Fernando Antônio Ferreira de. Os avanços da tecnociência, seus efeitos na sociedade contemporânea e repercussões no contexto brasileiro. In: BAUMGARTEN, Maíra (Org.). **A era do conhecimento**: matrix ou ágora. Brasília: UnB, 2001. p. 73-87.

BAUMGARTEN, Maíra. Globalização e ciência & tecnologia no limiar do século XXI: os anos 90 no Brasil. In: BAUMGARTEN, Maíra (Org.). **A era do conhecimento**: matrix ou ágora. Brasília: UnB, 2001. p. 89-119.

BELKIN, N. J. Progress in documentation: information concepts for information science. **Journal of Documentation**, v. 34, n. 1, p. 55-85, Mar. 1978.

BELL, Gordon. Prefácio. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011. p. 11- 15.

BLUE RIBBON TASK FORCE. **Sustainable economics for a digital planet**: ensuring long-term access to digital information. Final Report. OCLC, February, 2010. Disponível em < http://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf >

BOLLIER, David. **The promise and peril of big data**. Washington: Aspen Institute, 2010. Disponível em <http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf>. Acesso em: 15 maio 2014.

BORGMAN, Christine. Why are the attribution and citation of scientific data important? In: UHLIR, P. E. **For attribution**: developing data attribution and citation practices and standards: summary of an international workshop. Washington: The National Academic Press, 2013.

BORGMAN, Christine L. **Big data, little data, no data**: scholarship in the networked world. Cambridge: MIT Press, 2015. 416 p.

BORGMANN, Albert. Focal things and practices. In: SCHARFF, R. C.; DUSEK, Val. **Philosophy of technology**: the technological condition: an ontology. Oxford: Blackwell Publishing, 2006.

BORKO, H. Information Science: What is it? **American Documentation**, v.19, n.1, p.3-5, Jan. 1968.

BOURDIEU, Pierre. O campo científico. In: ORTIZ, Renato (Org.). **A sociologia de Pierre Bourdieu**. São Paulo: Olhos D'Água, 2003. p. 112-143.

BOURDIEU, Pierre. **Os usos sociais da ciência**: por uma sociologia clínica do campo científico. São Paulo: Unesp, 2004.

BRAMAN, Sandra. Defining information policy. **Journal of Information Policy**, n. 1, p. 1-5, 2011.

BRASIL. Lei nº 7.232 de 29 de outubro de 1984. Dispõe sobre a Política Nacional de Informática e dá outras providências. **Diário Oficial [da] República Federativa do Brasil**. Poder Executivo, Brasília, DF, 29 out. 1984, Seção 1.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. **Estratégica nacional de ciência, tecnologia e inovação 2012-2015**. Brasília, 2011.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. **Plano de ação em ciência, tecnologia e inovação**: principais resultados alcançados 2007-2010. Brasília, 2010.

BUCHAN, Iain; WINN, John; BISHOP, Chris. Uma abordagem unificada da modelagem de serviços de saúde com uso intensivo de dados. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011. p. 113-119.

BUCKLAND, Michael K.; Information as thing. **Journal of the American Society for information science**, v. 42, n.5, p. 351-360, 1991.

BUSH, Vannevar. As we may think. **Atlantic Monthly**, v. 176, n. 1, p. 101-108, July, 1945.

CARIBÉ, Rita de Cássia do Vale; MUELLER, Suzana Pinheiro Machado. Comunicação cientifica para o público leigo: história breve. **Informação & Informação**, v. 15, n. esp., p. 13-30, 2010.

CASTELLS, Manuel. **A galáxia da internet**: reflexões sobre a internet, os negócios e a sociedade. Rio de Janeiro: J. Zahar, 2003.

CÉSAR JÚNIOR, Roberto Marcondes. Apresentação à edição brasileira. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011. p. 7- 8.

CHARMAZ, Kathy. **A construção da teoria fundamentada**: guia prático para análise qualitativa. Porto Alegre: Artmed, 2009.

COLLIS, Jill; HUSSEY, Roger. **Pesquisa em administração**: um guia prático para alunos de graduação e pós-graduação. 2. ed. Porto Alegre: Bookman, 2005. (Métodos de pesquisa).

CORDEIRO, Daniel *et al*. Da ciência à e-ciência: paradigmas da descoberta do conhecimento. **Revista USP**, n. 97, p. 71-80, mar./maio 2013.

CORRÊA, Fabiano Couto. **Gestión de datos de investigación**. Barcelona: UOC, 2016. (Colección EPI scholar).

CORRÊA, Maíra Baumgarten. **O Brasil na era do conhecimento**: política de ciência, tecnologia e desenvolvimento sustentável. 2003. 309f. Tese (Doutorado em Sociologia)- Programa de Pós-Graduação em Sociologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

COSTA, Antônio Roberto F. Política nacional de informação científica e tecnológica: necessidade versus realidade. **Informação & Sociedade**, v. 1, n. 1, p. 30-37, jan./dez. 1991.

COSTA, Maíra Murrieta. **O social bookmarking como instrumento de apoio à elaboração de guias de literatura na Internet**. 2001. 251 f. Dissertação (Programa de Pós-Graduação em Ciência da Informação) – Universidade de Brasília, Brasília, 2011.

COSTA, Maíra Murrieta *et al.* Considerações iniciais sobre a e-science e a sua relação com a biblioteconomia e a ciência da informação. In: ENTRONTRO INTERNACIONAL DADOS, TECNOLOGIA E INFORMAÇÃO, 2013, São Paulo. **Anais...** Marília: UNESP, 2013.

Disponível em: <http://gpnti.marilia.unesp.br:8085/index.php/DTI/DTI/paper/viewFile/276/101>. Acesso em: 12 set. 2013.

COSTA, Maira Murrieta; CUNHA, Murilo Bastos da. A literatura internacional sobre e-Science nas bases de dados LISA e LISTA. **Revista Encontros Bibli**, v. 20, n. 44, p. 127-144, set./dez. 2015. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2015v20n44p127/30493>. Acesso em: 20 dez. 2016.

COSTA, Maíra Murrieta; CUNHA, Murilo Bastos da. O bibliotecário no tratamento de dados oriundos da e-science: considerações iniciais. **Perspectivas em Ciência da Informação**, v. 19, n. 3, p. 189-206, jul./set. 2014.

CRESPO, Isabel Merlo; CAREGNATO, Sônia Elisa. Padrões de comportamento de busca e uso de informação por pesquisadores de biologia molecular e biotecnologia. **Ciência da Informação**, v. 35, n. 3, p. 30-38, set./dez. 2006.

CUNHA, Murilo Bastos da. A biblioteca universitária na encruzilhada. **DataGramaZero – Revista de Ciência da Informação**, v. 11, n. 6, dez. 2010. Disponível em: <http://dgz.org.br/dez10/Art_07.htm>. Acesso em: 25 jul. 2012.

CUNHA, Murilo Bastos da. IBICT: 51 anos. **Ciência da Informação**, v. 34, n. 1, p. 7-8, jan./abr. 2005.

CUNHA, Murilo Bastos da. **Para saber mais**: fontes de informação em ciência e tecnologia. Brasília: Briquet de Lemos, 2001.

CUNHA, Murilo Bastos da. Prefacio. In: AMARAL, Sueli Angélica de. **Marketing da informação na internet**: ações de promoção. Campo Grande: UNIDERP, 2004.

CUNHA, Murilo Bastos da; COSTA, Maíra Murrieta. A preservação digital da e-science: aspirina ou vitamina? 2014. In: SEMINÁRIO INTERNACIONAL DE PRESERVAÇÃO DIGITAL, 1., Brasília, 2014; ENCONTRO NACIONAL DA REDE CARINIANA, 1.; 3., Brasília, 2014. **Palestra...** Brasília, 2014. Disponível em: <http://cariniana.ibict.br/images/sinpred/1/cunhaemaira.pdf>. Acesso em: 3 jun. 2014.

DATAONE. **Best practices**. 2016. Disponível em: <https://www.dataone.org/best-practices>. Acesso em: 3 jun. 2016.

DAVENPORT, Thomas H. **Ecologia da informação**. São Paulo: Futura, 2001.

DAVENPORT, Thomas H. **Big data no trabalho**: derrubando mitos e descobrindo oportunidades. Rio de Janeiro: Elsevier, 2014.

DAVIDOVICH, Luiz. **De olho no futuro**: a 4ª Conferência Nacional de Ciência, Tecnologia e Inovação. Revista da USP, São Paulo, n. 89, mar. 2011.

DOZIER, Jeff; GAIL, William B. A ciência emergente das aplicações ambientais. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011. p. 41-46.

ERCILIA, Maria; GRAEFF, Antonio. **A Internet**. 2. ed. São Paulo: Publifolha, 2008. (Folha explica).

FAPESP. E-Science é tema de seminário organizado pela FAPESP. **Agencia Notícias**, 9 maio de 2013. Disponível em: <http://agencia.fapesp.br/escience_e_tema_de_seminario_ organizado_pela_fapesp_e_pela_microsoft_research/17246/>. Acesso em: 25 nov. 2015.

FAPESP. FAPESP anuncia chamada para programa em eScience. **Agencia Notícias**, 2014. Disponível em: <www.fapesp.br/8361.phtml>. Acesso em: 25 nov. 2015.

FIRESTONE, Charles M. Foreword. In: BOLLIER, David. **The promise and peril of big data**. Washington, DC: Aspen Institute, 2010.

FLICK, Uwe. **Introdução à pesquisa qualitativa**. 3.ed. Porto Alegre: Artmed, 2009. 405p. (Métodos de Pesquisa)

FOX, Peter, HENDLER, James. E-science semântica: o significado codificado na próxima geração de ciência digitalmente aprimorada. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011.

FREITAS, Graça. Vagas em e-Science: Laboratório Nacional de Ciência e Tecnologia do Bioetanol. **Blog Antenado: nada definido, de tudo um pouco**. 29 jul. 2011. Disponível em <http://www.temosvagasdeemprego.com.br/2011/07/fwd-sbc-l-vagas-em-e-science.html>. Acesso em: 20 jul. 2012.

FUNDAÇÃO DE AMPARO A PESQUISA DO ESTADO DE SÃO PAULO. **Programa FAPESP de pesquisa em eScience**. [2015?]. Folder eletrônico do programa. Disponível em <http://www.fapesp.br/publicacoes/2015/folder_escience.pdf>. Acesso em: 20 dez. 2015.

FUNDAÇÃO DE AMPARO A PESQUISA DO ESTADO DE SÃO PAULO. **Programa FAPESP pesquisa em e-Science**. 2014. Slides de apresentação do evento.

FUNG, Margaret. National information policy: some basic considerations. In: _____. **Library cooperation and development seminar**. Taipei, Taiwan: [s.n.], 1986.

GARCIA, Maria Lúcia Andrade. A informação científica e tecnológica no Brasil. **Ciência da Informação**, v. 1, n. 1/2, p. 41-81, 1980.

GARVEY, W. D. **Communication**: the essence of science; facilitating information among librarians, scientists, engineers and students. Oxford: Pergamon, 1979.

GASQUE, Kelley Cristine Gonçalves Dias. **O pensamento reflexivo na busca e no uso da informação na comunicação científica**. 2008. 242 f. Tese (Programa de Pós-Graduação em Ciência da Informação) -- Universidade de Brasília, Brasília 2008.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2006.

GLASER, Barney G.; STRAUSS, Anselm L. **Awareness of dying**. Chicago: Aldine, 1965. 307 p.

_____. **The discovery of grounded theory**: strategies for qualitative research. New York: Aldine de Gruyter, 1967. 271p.

GÓES, Luís Fabrício Wanderley *et al*. **Computação em grade**: conceitos, tecnologias, aplicações e tendências. Minas Gerais: Escola Regional de Informática de Minas Gerais, 2005. Disponível em: <http://www.ppgee.pucminas.br/lsdc/artigos/goes_erimg05.pdf>. Acesso em: 10 maio 2016.

GOMES, Maria Yêda F. S. de Filgueiras. O estados e o processo de implantação de uma política nacional de informação científica e tecnológica no Brasil. **Ciência da Informação**, v. 17, n. 2, p. 105-117, jul./dez. 1988.

GONZÁLES DE GÓMEZ, Maria Nélida. Novos cenários políticos para a informação. **Ciência da Informação**, v. 31, n. 1, p. 27-40, jan./abr. 2002.

GONZÁLES DE GÓMEZ, Maria Nélida. O papel do conhecimento e da informação nas formações políticas ocidentais. **Ciência da Informação**, v. 16, n. 2, p. 157-167, jul./dez. 1987.

GRANVILLE, Vincent. **Vertical vs. horizontal data scientists**. Data Science Central: the online research for big data practitioners. 2013. Disponível em <http://www.datasciencecentral.com/profiles/blogs/vertical-vs-horizontal-data-scientists>. Acesso em: 30 ago. 2013.

GRAY, Jim. eScience: a transformed scientific method. In: CONSELHO NACIONAL DE PESQUISA DOS ESTADOS UNIDOS (NRC-CSTB), Califórnia, 2007. **Palestra...** 2007. Disponível em: <http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt>. Acesso em: 30 ago. 2012.

GREEN, Daron. Infraestrutura-científica: introdução. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011. p. 129-130.

HERRERA, Amílcar O. Los determinantes sociales de la política científica em America Latina: política cientifica explícita y política implícita. **Red de Revistas Científicas de América Latina, el Caribe, España y Portugal**, v. 2, n. 5, p, 117-131, dic. 1995.

HEY, Tony. **The UK e-science program**: next generation grid applications. [200-?]. Disponível em: <http://users.ecs.soton.ac.uk/ajgh/LyonGridTalk.pdf>. Acesso em: 27 dez. 2016. (Slides de apresentação da palestra do e -Science Core Program).

HEY, Tony. Why engage in e-science? **Library & Information Update**, 2004.

HEY, Tony; HEY, Jessie. E-science and its implications for the library community. **Library Hi-Tech**, 2006.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011.

HEY, Tony; TREFETHEN, Anne E. The UK e-science core programme and the grid. **Future Generation Computer Systems**, n. 18, p. 1017–1031, 2002.

HEY, Tony; TREFETHEN, Anne. E-science and its implications. **Philosophical Transactions of the Royal Society (A)**, v. 361, p. 1809-1825, June 2003.

HILBERT, Martin; LOPEZ, Priscila. How to measure the worlds technological capacity to communicate, store and compute information part I: results and scope. **International Journal of Communication**, v. 6, p. 956-979, April 2012.

HUNT, James R.; BALDOCCHI, Dennis D.; VAN INGEN, Catharine. Redefinição da ciência ecológica com o uso intensivo em dados. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). **O quarto paradigma**: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011. p. 47-51.

INDE. **Sig Brasil**: o portal brasileiro de dados espaciais. Apresentação. 2017. Disponível em: <http://www.inde.gov.br/a-inde/apresentacao.html>. Acesso em: 8 mar. 2017.

INGWERSEN, Peter.; JÄRVELIN, K. **The turn**: integration of information seeking and retrieval in context. New York: Springer-Verlag, 2005.

INSTITUO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. **Manifesto de acesso aberto a dados da pesquisa brasileira para ciência cidadã**. Outubro de 2016. Disponível em < http://www.ibict.br/Sala-de-Imprensa/noticias/2016/ibict-lanca- manifesto-de-acesso-aberto-a-dados-da-pesquisa-brasileira-para-ciencia-cidada >

JANKOWSKI, Nicholas W. **Exploring e-science**: an introduction. Journal of Computer-Mediated Communication, v. 12, p. 549-562, 2007.

JARDIM, José Maria. Políticas públicas de informação: a (não) construção da política nacional de arquivos públicos e privados no Brasil (1994-2006). In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2008, São Paulo. **Anais...** São Paulo: ANCIB, 2008.

JARDIM, José Maria. **Transparência e opacidade do Estados no Brasil**: usos e desusos da informação governamental. Niterói, RJ: EdUFF, 1999.

KUHN, Thomas S. **A estrutura das revoluções científicas**. São Paulo: Perspectiva, 2009. (Debates; 115).

LEMOS, Antônio Agenor Briquet de. Planejamento e coordenação da informação científica e tecnológica no Brasil. **Ciência da Informação**, v. 15, n. 2, p. 107-116, jul./dez. 1986.

LIMA, Paulo Gomes. **Política científica & tecnológica no Brasil no Governo Fernando Henrique Cardoso (1995-1998)**. Dourados: UFGD, 2011.

LIPPINCOTT, Joan K. Library and information technology support of e-science in the western context. In: MARCUM, Deanna B.; GEORGE, Gerald. **The data deluge**: can libraries cope with e-science? Santa Barbara, California: ABC CLIO, 2010.

LOPES, José Leite. **Ciência e liberdade**: escritos sobre ciência e educação no Brasil. Rio de Janeiro: UFRJ, 1998.

LUCE, Richard E. Grand challenges and new roles for the twenty-first-century research library in an era of e-science. In: MARCUM, Deanna B.; GEORGE, Gerald (Ed.). **The data deluge**: can libraries cope with e-science? Santa Barbara, California: Libraries Unlimited, 2010. cap. 1.

LYMAN, Peter; VARIAN, Hal R. **How much information 2003?** Berkeley, California: University of California at Berkeley, 2003. (Relatório produzido pelos estudantes da Escola de Gestão da Informação e Sistemas da Universidade da Califórnia em Berkeley). Disponível em: <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm>. Acesso em :18 jul. 2012.

MACHLUP, F.; MANSFIELD, U. **The study of information**: interdisciplinary messages. New York: Wiley, 1983.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 6. ed. São Paulo: Atlas, 2007.

MARCUM, Deanna B.; GEORGE, Gerald (Ed.). **The data deluge**: can libraries cope with e-science? Santa Barbara, California: Libraries Unlimited, 2010.

MARQUES, Rodrigo Moreno; PINHEIRO, Marta Macedo Kerr. Política de informação nacional e assimetria de informação no setor de telecomunicações brasileira. **Perspectivas em Ciência da Informação**, v. 16, n. 1, p. 65-91, jan./mar. 2011.

MATIAS-PEREIRA, José. **Manual de metodologia da pesquisa científica**. São Paulo: Atlas, 2007.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big data**: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Elsevier, 2013.

McDERMOTT, Irene E. Ransomware: tales from the cryptolocker. **Online Searcher**, v. 39, n. 3, p. 35-37, May/June 2015.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos, 1999.

MEADOWS, A. J. Os periódicos científicos e a transição do meio impresso para o eletrônico. **Revista de Biblioteconomia de Brasília**, v. 25, n. 1, p. 5-14, 2001.

MEDEIROS, Jackson da Silva; CAREGNATO, Sônia Elisa. Compartilhamento de dados de e-science: explorando um novo conceito para a comunicação científica. **Liinc em Revista**, v. 28, n. 2, p. 311-322, set. 2012.

MEIS, Leopoldo de; LETA, Jacqueline. **O perfil da ciência brasileira**. Rio de Janeiro: UFRJ, 1996.

MERTON, Robert. **Ensaios de sociologia da ciência**. São Paulo: Editora 34, 2013.

MINAYO, Maria Cecília de Souza (Org.). **O desafio do conhecimento**: pesquisa qualitativa em saúde. 10. ed. São Paulo: Hucitec, 2007.

MINAYO, Maria Cecília de Souza (Org.). **Pesquisa social**: teoria, método e criatividade. 6. ed. Rio de Janeiro: Vozes, 1996.

MOREL, Regina Lúcia de Moraes. **Ciência e estado**: a política científica no Brasil. São Paulo: T. A. Queiroz, 1979.

MORIN, Edgar. **Ciência com consciência**. 14. ed. Rio de Janeiro: Bertrand Russel, 2010.

MOURA, Maria Aparecida. Interoperabilidade semântica e ontologia semiótica: a construção e o compartilhamento de conceitos científicos em ambientes colaborativos online. **Informação & informação**, v. 16, n. esp., p. 165-179, jan,/jun. 2011.
MUELLER, Suzana Pinheiro Machado. A publicação da ciência: áreas científicas e seus canais preferenciais. **Datagramazero**, v. 6, n. 1, 2005.

MUELLER, Suzana Pinheiro Machado. Literatura cientifica, comunicação cientifica e ciência da informação. In: TOUTAIN, Lídia Maria Batista Brandão (Org.). **Para entender a ciência da informação**. Salvador: EDUFBA, 2007. p. 125-144.

MUELLER, Suzana Pinheiro Machado. O crescimento da ciência, o comportamento científico e a comunicação científica: algumas reflexões. **Revista da Escola de Biblioteconomia da UFMG**, v. 24, n. 1, p. 63-84, jan./jun. 1995.

MUNIZ, Nancy Aparecida Campos. **O CNPq e a sua trajetória de planejamento e gestão em C&T**: histórias para não dormir, contadas pelos seus técnicos (1975-1995). 2008. Tese (Doutorado em História) - Programa de Pós-Graduação em História, Universidade de Brasília, Brasília, 2008.

PELAEZ, Vitor; SZMRECSÁNYI, Tamás (Org.). **Economia da inovação tecnológica**. São Paulo: Hucitec, 2006.

PRINCIPE, Pedro *et al*. Estudo sobre os dados científicos gerados no âmbito da investigação produzida na Universidade do Minho. **Cadernos BAD**, n. 2, p. 3-17, jul./dez. 2014.

PSB-40. Câmara dos Deputados realiza seminário sobre extensão tecnológica. **Notícia**. 2011. Disponível em: <http://www.psb40.org.br/noticias/camara-dos-deputados-realiza-seminario-sobre-extensao-tecnologica/>. Acesso em: 18 out. 2016.

REDE CARINIANA. **The Dataverse project: orientações básicas para o uso**. Brasília, Rede Cariniana, 20017. (Slides apresentados durante a 4ª Reunião Técnica da Rede Brasileira de Repositórios Dataverse ocorrida em Brasília, no auditório do IBICT em de agosto de 2017).

REZENDE, Sérgio Machado. Apresentação. In: **Livro Azul da 4º Conferência de Ciência, Tecnologia e Inovação para o Desenvolvimento Sustentável**. Brasília: MCTI/CGEE, 2010a. Disponível em: <http://www.cgee.org.br/publicacoes/livroazul.php>. Acesso em: 03 junho 2014.

REZENDE, Sérgio Machado. **Momentos da ciência e tecnologia no Brasil**: uma caminhada de 40 anos pela C&T. Rio de Janeiro: Vieira & Kent, 2010.

RHEINGOLD, Howard. **Comunidade virtual**. Lisboa: Gradiva, 1996.

ROBREDO, Jaime; VILAN FILHO, Jaime Leyro. Metrias da informação: História e tendências. In: ROBREDO, Jaime; BRÄSCHER, Marisa (Org.). **Passeios no bosque da informação**: estudos sobre representação e organização da informação e do conhecimento – EROIC. Brasília: IBICT, 2010. cap. 10. p. 184-258. Disponível em: <http://www.ibict.br/publicacoes/eroic.pdf>. (Edição comemorativa dos 10 anos do Grupo de Pesquisa EROIC).

ROSENBEERG, Victor; CUNHA, Murilo Bastos da. **Uso de informação técnica e científica no Brasil**. Brasilia: IBICT, 1983. (mimeografado).

ROSENBERG, Victor. National information policies. **Annual Review of Information Science and Technology**, v. 17, p. 3-32, 1982b.

ROSENBERG, Victor. Política de informação nos países em desenvolvimento: o caso do Brasil visto por um americano. **Ciência da Informação**, v. 11, n. 2, p. 37-43, 1982a.

ROWLANDS, Ian. Some compass bearings for information policy orienteering. **Aslib Proceedings**, v. 50, n. 8, p. 230-237, Sept. 1998.

SALERNO, Mario Sergio; KUBOTA, Luís Claudio. Estado e inovação. In: DE NEGRI, João Alberto; KUBOTA, Luís Claudio. **Políticas de incentivo à inovação tecnológica no Brasil**. Brasília: IPEA, 2008. p. 13-64.

SALES, Luana Farias. **Integração semântica de publicações científicas e dados de pesquisa**: proposta de modelo de publicação ampliada para a área de ciências nucleares. 2014. 268 f. Tese (Programa de Pós-Graduação em Ciência da Informação) – Universidade Federal do Rio de Janeiro, 2014.

SALES, Luana Farias; SAYÃO, Luís Fernando. Ciberinfraestrutura para integração, acesso, compartilhamento e reuso de dados de pesquisa da área nuclear. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2013, Maceió. **Anais...** Porto Alegre: SBCC, 2013. p. 1-5.

SAMPIERI, Roberto Hernández; COLLADO, Carlos Fernández; LUCIO, Pilar Baptista. **Metodologia de pesquisa**. 3. ed. São Paulo: McGraw Hill, 2006.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, v.1, n.1, p. 41-62, 1996.

SAYÃO, Luís Fernando; SALES, Luana Farias. Curadoria geral: um novo patamar para a preservação de dados digitais de pesquisa. **Informação & Sociedade**, v. 22, n. 3, p. 179-191, set./dez. 2012.

SAYÃO, Luís Fernando; SALES, Luana. Farias. Dados abertos de pesquisa: ampliando o conceito de acesso livre. RECIIS. **Electronic Journal of Communication Information and Innovation in Health**, v. 8, p. 76-92, 2014.

SAYÃO, Luís Fernando; SALES, Luana Farias. Ciberinfraestrutura de informação para a pesquisa: proposta de integração entre repositório institucional, repositório de dados e CRIS. **Informação & Sociedade**, v. 25, p. 163-184, 2015.

SCHWARTZMAN, Simon. **Um espaço para a ciência**: a formação da comunidade científica no Brasil. Brasília: MCTI, 2001. Disponível em: <http://www.schwartzman.org.br/>. Acesso em: 6 ago. 2013.

SHEARER, Kathleen. **Comprehensive brief on research data management policies**. Canada Government, 2015. Disponível em : <http://docplayer.net/17594465-Comprehensive-brief-on-research-data-management-policies.html> Acesso em 28 julho 2016.

SILVA, Marta Benjamim da; BERNARDINO, Maria Cleide Rodrigues; NOGUEIR, Carine Rodrigues. **Políticas públicas para a leitura no Brasil**: implicações sobre a leitura infantil. Ponto de Acesso, v. 6, n. 3, p. 20-46, abr. 2012.

SILVA, Terezinha Elizabeth da. Política de informação na pós-modernidade: reflexões sobre o caso do Brasil. **Informação & Sociedade**, v. 1, n. 1, p. 8-13, jan./dez. 1991.

SIMBERLOFF, Daniel et al. **Long-lived Digital data collections**: enabling research and education in the 21st century. National Science Board, National Science Foundation, 2005. Disponível em < https://www.nsf.gov/pubs/2005/nsb0540/nsb0540_1.pdf >

SOCIEDADE BRASILEIRA DE PROGRESSO DA CIÊNCIA. Helena Nader: "Não vamos desistir do MCTI". **Notícias**, jun. 2016. Disponível em: <www.sbpcnet.org.br>. Acesso em: 25 ago. 2016.

SOEHNER, Catherine; STEEVES, Catherine; WARD, Jennifer. **E-science and data support services**: a study of ARL member institutions. Washington, DC: Association of Research Libraries, 2010. Disponível em: <http://www.arl.org/bm~doc/escience_report2010.pdf>. Acesso em: 18 jul. 2012.

SOLLA-PRICE, Derek J. **O desenvolvimento da ciência**: análise histórica, filosófica, sociológica e econômica. Rio de Janeiro: Livros Técnicos e Científicos, 1976.

STRASSER, Carly *et al*. **Primer on data management**: what you always wanted to know. UC Office of the President: California Digital Library, 2012. (**CDL Staff Publications**). Disponível em < http://escholarship.org/uc/item/7tf5q7n3>

STRASSER, Carly. **Research data management**: a primer publication of the national information standards organization. Baltimore: NISO, 2015. (NISO Premier Series). Disponível em < http://wiki.lib.sun.ac.za/images/2/24/PrimerRDM-2015-0727.pdf >

STRAUSS, Anselm; CORBIN, Juliet. **Pesquisa qualitativa**: técnicas e procedimentos para o desenvolvimento da teoria fundamentada. 2. ed. Porto Alegre: Artmed, 2008. 288 p.

SZIGETI, Kathy; WHEELER, Kathy. Science and technology resources on the Internet: essential readings in e-Science. **Issues in Science and Technology Librarianship**, v. 64, 2011. Disponível em: < http://dx.doi.org/10.5062/F400001J >. Acesso em : 30 abr. 2014.

TALIA, Domenico. Workflow systems for science: concepts and tools. **International Scholarly Research Notices Software Engineering**, 2013. Disponível em: <http://dx.doi.org/10.1155/2013/404525>. Acesso em: 03 fev. 2017.

TAPSCOTT, Don ; WILLIAMS, Anthony D. **Wikinomics**: como a colaboração em massa pode mudar o seu negócio. Rio de Janeiro: Nova Fronteira, 2007.

TARAPANOFF, Kira. A política científica e tecnológica no Brasil: o papel do IBICT. **Ciência da Informação**, v. 21, n. 2, p. 87-166, maio/ago. 1992.

TENOPIR, Carol; BIRCH, Ben; ALLARD, Suzie. **Academic libraries and research data services:** current practices and plans for the future. Association of College and Research Libraries (ACRL) White Paper. June 2012. Disponível em http://www.ala.org /acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

UNITED STATES. White House. Executive Office of the President. **Big data**: seizing opportunities, preserving values. Washington, 2014. Disponível em: <http://www.whitehouse.gov/sites/ default/files/docs/big_data_privacy_report_may_1_2014.pdf>. Acesso em: 20 maio 2014.

UNIVERSIDADE ESTADUAL DE SÃO PAULO – UNESP. E-Science é tema de seminário a FAPESP. **Portal de Notícias da UNESP**, 10 maio 2013. Disponível em: <http://www.unesp.br/portal#!/noticia/10868/e-science-e-tema-de-seminario-na-fapesp/>. Acesso em: 05 jul. 2013.

VAZ, Glauber José. **E-Science na Embrapa**. Campinas: Embrapa Informática Agropecuária, 2011. (Documentos, 117).

VERGARA, Sylvia Constant. **Projetos e relatórios de pesquisa em administração**. 5. ed. São Paulo: Atlas, 2004.

VIDEIRA, Antonio Augusto Passos. **25 anos de MCT**: raízes históricas da criação de um ministério. Rio de Janeiro: Centro de Gestão e Estudos Estratégicos, 2010.

WERSIG, G.; NEVELING, U. The phenomena of interest to information science. **The Information Scientist**, v. 9, n. 4, 1975.

ZIMAN, John. **An introduction to science studies**: the philosophical and social aspects of science and technology. Cambridge: Cambridge University, 1984.

ZIMAN, John. **Real science**: what it is and what it means. Cambridge: Cambridge University Press, 2000.

**APÊNDICE 1 – QUESTIONÁRIO DISPONIBILIZADO PARA OS PROFESSORES DA SCHOOL OF INFORMATION – UNIVERSITY OF MICHIGAN**

# Scientific Data Management Policy

*Obrigatório

1. In this university , you see yourself as: *

- ◯ Professor
- ◯ PhD. Student
- ◯ Master Degree Student
- ◯ Information Technology Professional
- ◯ Librarian
- ◯ Outro: _____

2. What type of raw data you produce? *
National Science Board Taxonomy

- ◯ Observation data
- ◯ Simulation data
- ◯ Generated data in the laboratory
- ◯ Automatically collected data by specialized sensors
- ◯ Outro: _____

3. As a researcher, what is the main source of data for your research project? *
National Science Board Taxonomy

- ◯ Observation data
- ◯ Simulation data
- ◯ Generated data in the laboratory
- ◯ Automatically collected data by specialized sensors
- ◯ Outro: _____

4. As a researcher, you would like to have access to raw data of other research *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

5. You have had access to raw data from other researchers. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

6. You would trust in the authenticity of raw data from other research. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

7. You would trust in the transparency of raw data from other research. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

8. As a researcher, you would share the raw data of your research with other researchers. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

9. If you wish, please make a comment on question 08.

Sua resposta

10. You have any kind of rule to share the raw data from your research. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

11. Identify the reasons why you would have no interest in sharing the raw data from your research. *

Mark as many answers as necessary.

- [ ] Do not know who will use them
- [ ] Not sure how the data will be used
- [ ] No interest in sharing the details of your research
- [ ] Not sure how this works
- [ ] Outro: _____

12. Your institution provides infraestructure for the scientific data management. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

13. Your institution has a designated department to offer the service of data curation. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

14. If possible, please comment about question above (n. 13).

Who performs this activity? To whom the department is administratively subordinated? In your opinion, has this department skilled human resources for this purpose?

Sua resposta

15. Your institution has a scientific data management policy. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

16. NSF has a scientific data management policy. *

NSF - National Science Foundation

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

17. You have had access to any document that provides guidance on the storage of raw scientific data. *

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Agree | ○ | ○ | ○ | ○ | ○ | Disagree |

18. If possible, please comment about questions 11, 12 and 13 noting the aspects of access to information, preserving information and guidelines for information reuse.

Sua resposta

19. In your opinion, what are the difficulties of implementing the online scientific data management policy? *

Sua resposta

20. As a researcher, what does the term data curation mean to you?

Sua resposta

21. If you receive funding for your research, the funding agency would require you to have a plan for the management and preservation of scientific data. What would you do? *

Sua resposta

22. If you have interest in participating in the next steps of this research, please let me know your email, skype or telephone number.

Sua resposta

ENVIAR

Nunca envie senhas pelo Formulários Google.

**APENDICE 2 – FORMULÁRIO DE COLETA DE DADOS – AGÊNCIAS DE FOMENTO E/OU FUNDAÇÕES DE AMPARAO À PESQUISA NO BRASIL**

1. Informe seus contatos e vínculo empregatício

Nome

Empresa

Endereço de *email*

N° telefone

*2. Durante a sua trajetória em agências de fomento para pesquisa, observou se haviam pesquisadores preocupados com a gestão e a preservação dos dados produzidos por suas respectivas pesquisas?

*3. Enquanto servidor de uma agencia de fomento, o (a) senhor (a) conhece e/ou conheceu projetos desenvolvidos pelas universidades e/ou instituições de pesquisa brasileiras que tem/tiveram a necessidade de uma política que norteasse a gestão de dados científicos? Se possível, de exemplos.

*4. Na sua opinião, as agências de fomento estão atentas a necessidade de tratamento, armazenamento e a preservação digital de dados científicos brutos que estão sendo produzidos pelas instituições brasileiras? As agencias precisam fomentar essa discussão?

*5. Na sua opinião, na CAPES (ou outra agencia de fomento) existe alguma diretriz de coleta, tratamento técnico, armazenamento e preservação de dados científicos gerados pela pesquisa financiada pela agencia?

*6. A sua Agencia de Fomento (CAPES, CNPq, FAPs) possui algum sistema que recupere os *dados brutos* (*raw data*) das pesquisas por ela financiada/apoiada?

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

*7. A sua Agencia de Fomento possui um sistema de informação gerencial que reúna informações sobre as pesquisas correntes financiadas pela instituição.

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

*8. Em caso da sua Agencia de Fomento possuir um Sistema de Informação Gerencial, esse sistema registra o tipo de dado que o pesquisador está produzindo.

*São exemplos de tipos de dados: dados de observação, dados de simulação, dados de laboratório. São exemplo da área de conhecimento do dado : biodiversidade brasileira, dados espaciais, dados de engenharia nuclear dentre outros.*

*9. Na sua opinião, a gestão dos dados científicos, gerados por pesquisas financiadas por agências de fomento, é um *dever* da respectiva agencia.

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

10. Por favor, se possível, comente sobre a questão anterior.

*11. Na sua opinião o Brasil deveria ter um repositório central de dados de pesquisa (*raw data*)? Qual o grau de relevância dessa iniciativa? Caso concorde com a iniciativa, qual a instituição teria capacidade técnica para conduzir essa ação?

*Essa pergunta refere-se ao desenvolvimento de um repositório nacional que integre os dados brutos de pesquisa (raw data) gerados por pesquisas financiadas por agências de fomento. Como exemplo de iniciativa semelhante, cita-se a Biblioteca Digital de Teses e Dissertações (BDTD) responsável por armazenar documentos bibliográficos (teses e dissertações), que por sua vez, são diferentes de dados científicos primários.*

*12. Está no planejamento da sua Agencia de Fomento desenvolver softwares de acesso aos dados brutos de pesquisa?

*13. Na sua opinião, a sua Agência de Fomento precisa de uma política para a gestão de dados científicos?

*A pergunta-se refere-se a uma política que norteie a gestão dos dados brutos (raw data)coletados pelo pesquisador. Não se aplica aqui nessa questão as políticas inerentes ao tratamento de documentos bibliográficos resultantes de pesquisas, como, por exemplo, tese, dissertação, relatório de pesquisa, artigo dentre outros.*

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

14. Se desejar, utilize o espaço abaixo para comentários gerais sobre a necessidade de uma política institucional para a gestão de dados científicos.

*15. Na sua opinião, o Brasil precisa de uma política nacional de gestão de dados científicos.

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

16. Se desejar, utilize o espaço abaixo para comentários gerais sobre a necessidade de uma Política Nacional para a gestão de dados científicos.

*17. Na sua opinião, deve fazer parte da atual agenda de política de ciência e tecnologia:

| | Irrelevante | Pouco Importante | Indiferente | Importante | Muito Importante |
|---|---|---|---|---|---|
| Desenvolver e disponibilizar um repositório de dados científicos | ○ | ○ | ○ | ○ | ○ |
| Desenvolver diretrizes para a coleta, tratamento técnico, armazenamento, preservação dos dados científicos gerados por pesquisas financiadas pelo governo | ○ | ○ | ○ | ○ | ○ |
| Desenvolver diretrizes para a re-utilização dos dados, para além do contexto inicial em que foram criados, com o objetivo de poupar recursos públicos de financiamento | ○ | ○ | ○ | ○ | ○ |
| Discutir com a comunidade acadêmica questões relacionadas a propriedade do dado (quem coletou é um técnico de coleta ou é autor do dado? Ou ainda, o dado é de propriedade do governo brasileiro?) | ○ | ○ | ○ | ○ | ○ |
| Desenvolver uma tabela de temporalidade para o prazo de carência dos dados (quando o dado pode ser divulgado) por área de conhecimento | ○ | ○ | ○ | ○ | ○ |
| Desenvolver regras para o compartilhamento de dados em nível nacional | ○ | ○ | ○ | ○ | ○ |
| Desenvolver regras para o compartilhamento de dados em nível internacional | ○ | ○ | ○ | ○ | ○ |
| Desenvolver mecanismos de reconhecimento ao pesquisador que coleta dados (a exemplo do pesquisador que publica artigos | ○ | ○ | ○ | ○ | ○ |

*18. Na sua opinião, quais são as dificuldades para a implementação de uma política nacional de gestão de dados científicos.

*19. Na sua opinião, o pesquisador brasileiro dedica algum tempo da sua pesquisa para administrar os seus dados?

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

*20. Sobre a pergunta anterior, o(a) senhor(a) acha que seria necessário o pesquisador dispusesse de um tempo para gerenciar seus dados de pesquisa? Em caso negativo, que o(a) senhor(a) julga que deveria conduzir essa atividade?

*21. Dentre os motivos abaixo relacionados, selecione quais justificam o dado científico não estar disponibilizado para a consulta on-line ao público em geral.

*Responda até três opções.*

- alta de recurso financeiro
- alta de infraestrutura tecnológica
- alta de diretrizes institucionais sobre o que pode ser disponibilizado
- alta de tempo para disponibilizar os dados
- alta de equipe para tal finalidade
- alta de incentivos administrativos para compartilhar os dados
- .usência de conhecimento sobre o local para disponibilizar os dados para acesso público
- )s dados são de uso restrito devido a questões de propriedade intelectual (desenvolvimento ıecnológico) ou segurança nacional.
- ) fato de não ser uma exigência das agências de fomento brasileiras.
- ) fato de não ser pré-requisito para obter financiamento internacional.
- ) pré-julgamento de que os dados não serão de interesse para outros pesquisadores.

22. Se desejar, utilize o espaço abaixo para comentários gerais sobre a pergunta anterior

*23. Dentre os motivos abaixo relacionados, quais incentivariam o depósito de dados em um repositório de acesso público.

*Responda até três opções.*

- [ ] er um requisito da agência de fomento.
- [ ] ) fato de haver um incremento no valor de financiamento para preparar os dados para serem compartilhados.
- [ ] ) fato de haver um incremento no valor financiado para viabilizar o gerenciamento dos dados depois que este é depositado em um repositório.
- [ ] .uxílio institucional para o gerenciamento de dados
- [ ] .econhecimento acadêmico para o pesquisador que coletou o dado.
- [ ] ) repositório oferecer o serviço de armazenamento de longo prazo.
- [ ] ) repositório oferecer a funcionalidade/capacidade de restringir o acesso aos dados (embargo dos dados/ carência dos dados) de acordo com as regras pré-estabelecidas (temporalidade).

24. Se desejar, utilize o espaço abaixo para comentários gerais sobre a pergunta anterior.

*25. Considerando o processo de publicação científica no Brasil, bem como o processo de publicação científica internacional (a janela de tempo entre a submissão do artigo, o aceite e a publicação do mesmo), na sua opinião, qual seria o prazo razoável para o embargo dos dados (período de carência para acesso a totalidade dos dados)

- ○ Menos de 1 ano
- ○ Entre 1 e 2 anos
- ○ Entre 3 e 5 anos
- ○ Entre 5 e 8 anos
- ○ Entre 8 e 10 anos
- ○ Não tenho conhecimento para responder
- ○ Outro (especifique)

26. Se desejar, utilize o espaço abaixo para comentários gerais sobre a pergunta anterior.

*27. No que diz respeito aos dados de pesquisa que culminam com o desenvolvimento tecnológico de um produto e, considerando o prazo de concessão de uma patente pelo INPI (em média 8 anos), qual o prazo de embargo dos dados (carência) da pesquisa que o(a) senhor(a) julga adequado? O senhor tem algum comentário específico sobre alguma área do conhecimento?

*28. Na sua opinião, qual o perfil do cientista de dados? Quais são as características desse profissional?

*29. Na sua opinião quem é o profissional capacitado a tratar o dado científico, ou seja, realizar a curadoria de dados.

30. Na sua opinião, as universidades no Brasil têm contribuído para a formação desse profissional capaz de realizar o gerenciamento de dados científicos, bem como a curadoria de dados?

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

31. Se desejar, utilize o espaço abaixo para comentários gerais sobre a pergunta anterior.

*32. O que o termo "curadoria de dados" significa para você?

*33. O que o termo "gestão de dados científicos" significa para você?

34. Em função da sua experiência, quem o(a) senhor(a) indicaria para participar dessa pesquisa?

# APÊNDICE 3 – FORMULÁRIO DE COLETA DE DADOS – DOUTORES E/OU DOUTORANDOS ENVOLVIDOS COM A GESTÃO DE DADOS CIENTÍFICOS NO BRASIL

Convite - Questionário - Tese de Doutorado sobre política para gestão de dados científicos

*1. No exercício de suas atividades profissionais (acadêmicas), você produz algum tipo de dado de pesquisa (*dados científicos*)? Em caso positivo, identifique abaixo quais os tipos de dados. Caso, na sua opinião, seu tipo de dado não se encaixe nessa taxonomia, por favor, deixe me saber. Utilize o espaço abaixo para comentários.

*ATENÇÃO: Se você for profissional de TI, Gestão ou bibliotecário, você provavelmente trabalha com curadoria e armazenamento dos dados. Responda com esse viés. Obrigada!*

**As opções de respostas estão fundamentadas na National Science Board Taxonomy e, em caso de dúvidas, podem ser consultadas Simberloff *et al*. (2005) ou em Borgman (2015).**

SIMBERLOFF, Daniel *et al*. **Long-lived digital data collections**: enabling research and education in the 21st Century. Arizona: National Science Board, 2005. / BORGMAN, Christine L. **Big data, little data, no data**: scholarship in the networked world. Massachusetts: Cambridge, 2015.

- [ ] ıados de Observação
- [ ] ıados de Simulação
- [ ] ıados Produzidos em Laboratórios
- [ ] ıados Coletados automaticamente por sensores especializados (*exemplo, dados sobre a terra*)
- [ ] ıados abertos do governo

Por favor, deixe me saber quais são os tipos de dados que você produz e as peculiaridades de tratamento do mesmo.

*2. Na sua opinião, os dados produzidos pela sua pesquisa são de alguma forma preservados?

*A pergunta se refere aos dados brutos coletados pelo pesquisador. Não se aplica aqui nessa questão, a preservação digital de documentos bibliográficos resultantes da pesquisa, como, por exemplo, tese, dissertação, relatório de pesquisa, artigo dentre outros.*

Discordo — Não Sei — Concordo

3. Se desejar, comente sobre a questão anterior.

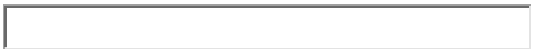
*Você pode utilizar esse espaço para comentar como tem procurado preservar seus dados, as dificuldades encontradas, bem como relatar experiências de sucesso que teve a oportunidade de conhecer.*

*4. Existe algum sistema de busca que recupere os dados brutos (***raw data***) da sua pesquisa?

*A pergunta se refere aos dados brutos coletados pelo pesquisador. Não se aplica aqui nessa questão o uso de sistemas de recuperação de documentos bibliográficos.*


5. Se desejar, comente sobre a questão anterior.

*Esse sistema é um sistema de busca institucional? Se aplica as demais pesquisas da instituição? Ou, é uma iniciativa isolada?*

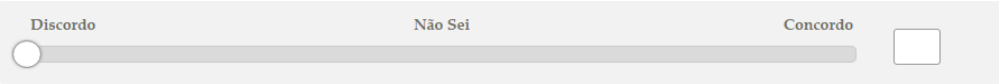


*6. Você trabalha com algum workflow científico. Em caso positivo, qual? Qual o benefício da utilização do software para a pesquisa?

*São exemplos de workflow científico: Askalon, Chiron, DIS3GNO,DVega, Galaxy, GWES, Karajan, Kepler, Pegasus, Swift, Taverna, Triana, Weka2WS.*

*7. Na sua opinião, os dados brutos produzidos pela sua pesquisa possuem classificação quanto ao ciclo de vida?

*A pergunta se refere aos dados brutos coletados pelo pesquisador. Não se aplica aqui nessa questão a classificação de documentos bibliográficos resultantes da pesquisa, como, por exemplo, tese, dissertação, relatório de pesquisa, artigo dentre outros. São exemplos de ciclos de vida o proposto pelo Projeto DATAONE, ou mesmo o do Dicgital Curation Center.*




8. Se desejar, comente sobre a questão anterior.

*Pontos para reflexões: É um plano de classificação institucional? É uma iniciativa isolada do projeto de pesquisa? Foi exigido pela agência de fomento? Em caso de projetos com apoio internacional, foi uma exigência para receber esse apoio?*

9. Enquanto pesquisador, qual a principal fonte de dados para o seu projeto de pesquisa?

*Caso, na sua opinião, seu tipo de dado não se encaixe nessa taxonomia, por favor, deixe me saber. Utilize o espaço abaixo para comentários.*

- Dados de Observação
- Dados de Simulação
- Dados Produzidos em Laboratório
- Dados coletados automaticamente por sensores especializados
- Dados abertos do governo

Por favor, deixe me saber quais são os tipos de dados que você produz e as peculiaridades de tratamento do mesmo.

*10. Enquanto pesquisador, você gostaria de ter acesso aos dados brutos de outras pesquisas.

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

*11. Você já obteve acesso aos dados brutos (***raw data***) de outras pesquisas?

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo
- Nesse caso, você obteve acesso a dados de pesquisa de um mesmo grupo de pesquisadores? Foi por afinidade de tema de pesquisa? Como se deu esse compartilhamento?

*12. Você confia na autenticidade dos dados brutos de outra pesquisa

*Entende-se autenticidade como: acesso a dados verdadeiros, ou reproduções totalmente fiéis ao conteúdo original.*

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

*13. Você confia na transparência dos dados brutos de outra pesquisa disponibilizados na Rede?

*O conceito de transparência está relacionado ao de autenticidade, mas extrapola esse quando também se refere a obrigação imposta a todos os administradores públicos no que diz respeito ao zelo e à necessidade de publicidade do dado, bem como, da informação pública, de forma que a sociedade tenha a capacidade para exercer seu poder de fiscalização. Refere-se também ao Decreto nº 5.482, de 30 de junho de 2005 que tem como finalidade divulgar dados e informações dos órgãos e entidades da administração pública federal na Internet.*

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

14. Se desejar, comente sobre as questões acima referentes a acesso aos dados brutos de outras pesquisas.

*15. Você compartilha os dados brutos de sua pesquisa com outros pesquisadores.

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

Por favor, comente - O compartilhamento se dá com pesquisadores que possuem afinidade com o seu tema de pesquisa? Compartilha apenas entre os membros de seu grupo de pesquisa? O compartilhamento é feito com qualquer outro pesquisador?

*16. Você possui "condições/regras" para compartilhar os dados brutos de sua pesquisa.

*Por exemplo, conhecer o pesquisador; exigir a citação dos dados brutos originais; compreender como o dado será reutilizado dentre outros.*

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

Se possuir regras e/ou condições, comente sobre elas.

*17. Sua instituição fornece infraestrutura para a gestão de dados científicos.

*A pergunta se refere aos dados brutos coletados pelo pesquisador. Não se aplica aqui nessa questão infraestrutura para a gestão de documentos bibliográficos (software de bibliotecas, biblioteca de teses e dissertações). - Por exemplo, possui banco de dados com conceito de processamento em grid? Utilizam o MapReduce?*

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

18. Na sua opinião, quais as ferramentas computacionais utilizadas na gestão de dados científicos na sua instituição?

*A pergunta se refere aos dados brutos coletados pelo pesquisador. Não se aplica aqui nessa questão software para tratamento de documentos bibliográficos e/ou arquivísticos.*

*19. Sua instituição possui um departamento dedicado a oferecer um serviço de curadoria de dados.

*A pergunta se refere ao serviço de curadoria para dados brutos coletados pelo pesquisador. Não se aplica aqui nessa questão a curadoria de documentos bibliográficos resultantes da pesquisa, como, por exemplo, tese, dissertação, relatório de pesquisa, artigo dentre outros. São exemplos de metodologias para a curadoria de dados o modelo do Digital Curation Center, ou ainda o do Projeto DataONE.*

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

20.Se possível, faça comentários sobre a questão anterior.

*Pontos para Refletir - Quem executa a atividade de curadoria de dados? É um setor específico? Qual? A quem esse setor está subordinado administrativamente? Esse setor possui recursos humanos capacitados para a curadoria de dados?*

*21. Sua instituição possui uma política para a gestão de dados científicos.

*A pergunta se refere a uma política que norteie a gestão dos dados brutos coletados pelo pesquisador (a espécie de uma planta na Amazônia etc.). Não se aplica aqui nessa questão políticas de segurança da informação, ou ainda políticas inerentes ao tratamento de documentos bibliográficos.*

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

*22. Você já teve acesso a documentos que fornecem diretrizes para o armazenamento e gestão de dados científicos (***raw data***).

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

23. Se desejar, comente sobre a questão anterior.

*Que documento foi esse? Está disponível para consulta pública? Se tiver, por favor, referencie o mesmo.*

*24. Por favor, utilize o espaço abaixo para comentar sobre os aspectos de acesso aos dados científicos brutos, a preservação digital desses dados, bem como diretrizes para reutilização dos dados.

*Qual o seu posicionamento perante essa nova realidade, principalmente ao se considerar que revistas internacionais de ponta têm exigido acesso aos dados de pesquisa brutos para se publicar o artigo.*

*25. Na sua opinião, quais são as dificuldades para a implementação uma política de gestão de dados científicos?

*26. O Brasil precisa de uma política nacional de gestão de dados científicos.

- 1 Discordo
- 2
- 3
- 4
- 5 Concordo

*27. Comente sobre a necessidade dessa Política Nacional para a Gestão de Dados Científicos. Como ela se daria? Quais seriam os principais interlocutores na sua opinião?

*28. Caso a agência de fomento exigisse um plano de gestão de dados no projeto de sua pesquisa, você teria condições de elaborar esse documento?

*Se você precisasse de ajuda. Nesse caso, a que tipo de profissional você iria recorrer?*

*29. Como pesquisador, o que o termo "***curadoria de dados***" significa para você?

*30. Como pesquisador, o que o termo "***gestão de dados científicos***" significa para você?

31.Você tem interesse em participar das próximas etapas dessa pesquisa?

Em caso positivo, fica a seu critério colocar seu ***email***, ***telefone,*** **Skype**. Enfim, a forma que for mais conveniente para você.

***Meus sinceros agradecimentos por participar dessa pesquisa. Aproveito para informar que os dados serão tratados no anonimato.***

# ANEXO 1 – INFRA- ESTRUTURA NACIONAL DE DADOS ESPACIAIS - INDE

**Presidência da República**
**Casa Civil**
**Subchefia para Assuntos Jurídicos**

**DECRETO Nº 6.666, DE 27 DE NOVEMBRO DE 2008.**

Institui, no âmbito do Poder Executivo federal, a Infra-Estrutura Nacional de Dados Espaciais - INDE, e dá outras providências.

**O PRESIDENTE DA REPÚBLICA**, no uso da atribuição que lhe confere o art. 84, inciso VI, alínea "a", da Constituição, e tendo em vista o disposto no Decreto nº 89.817, de 20 de junho de 1984, e no Decreto de 1º de agosto de 2008, que dispõe sobre a Comissão Nacional de Cartografia - CONCAR,

**DECRETA:**

Art. 1º Fica instituída, no âmbito do Poder Executivo federal, a Infra-Estrutura Nacional de Dados Espaciais - INDE, com o objetivo de:

I - promover o adequado ordenamento na geração, no armazenamento, no acesso, no compartilhamento, na disseminação e no uso dos dados geoespaciais de origem federal, estadual, distrital e municipal, em proveito do desenvolvimento do País?

II - promover a utilização, na produção dos dados geoespaciais pelos órgãos públicos das esferas federal, estadual, distrital e municipal, dos padrões e normas homologados pela Comissão Nacional de Cartografia - CONCAR? e

III - evitar a duplicidade de ações e o desperdício de recursos na obtenção de dados geoespaciais pelos órgãos da administração pública, por meio da divulgação dos metadados relativos a esses dados disponíveis nas entidades e nos órgãos públicos das esferas federal, estadual, distrital e municipal.

§ 1º Para o atingimento dos objetivos dispostos neste artigo, será implantado o Diretório Brasileiro de Dados Geoespaciais - DBDG, que deverá ter no Portal Brasileiro de Dados Geoespaciais, denominado "Sistema de Informações Geográficas do Brasil - SIG Brasil", o portal principal para o acesso aos dados, seus metadados e serviços relacionados.

Art. 2º Para os fins deste Decreto, entende-se por:

I - dado ou informação geoespacial: aquele que se distingue essencialmente pela componente espacial, que associa a cada entidade ou fenômeno uma localização na Terra, traduzida por sistema geodésico de referência, em dado instante ou período de tempo, podendo ser derivado, entre outras fontes, das tecnologias de levantamento, inclusive as associadas a sistemas globais de posicionamento apoiados por satélites, bem como de mapeamento ou de sensoriamento remoto?

II - metadados de informações geoespaciais: conjunto de informações descritivas sobre os dados, incluindo as características do seu levantamento, produção, qualidade e estrutura de armazenamento, essenciais para promover a sua documentação, integração e disponibilização, bem como possibilitar a sua busca e exploração?

III - Infra-Estrutura Nacional de Dados Espaciais - INDE: conjunto integrado de tecnologias? políticas? mecanismos e procedimentos de coordenação e monitoramento? padrões e acordos, necessário para facilitar e ordenar a geração, o armazenamento, o acesso, o compartilhamento, a disseminação e o uso dos dados geoespaciais de origem federal, estadual, distrital e municipal?

IV - Diretório Brasileiro de Dados Geoespaciais - DBDG: sistema de servidores de dados, distribuídos na rede mundial de computadores, capaz de reunir eletronicamente produtores, gestores e usuários de dados geoespaciais, com vistas ao armazenamento, compartilhamento e acesso a esses dados e aos serviços relacionados? e

V - Portal Brasileiro de Dados Geoespaciais, denominado "Sistema de Informações Geográficas do Brasil - SIG Brasil": portal que disponibilizará os recursos do DBDG para publicação ou consulta sobre a existência de dados geoespaciais, bem como para o acesso aos serviços relacionados.

§ 1º Os dados estatísticos podem, a critério do órgão produtor, ser considerados como dados geoespaciais, desde que estejam de acordo com a definição do inciso I do caput.

§ 2º Serão considerados dados geoespaciais oficiais aqueles homologados pelos órgãos competentes da administração pública federal, e que estejam em conformidade com o inciso I do caput.

Art. 3º O compartilhamento e disseminação dos dados geoespaciais e seus metadados é obrigatório para todos os órgãos e entidades do Poder Executivo federal e voluntário para os órgãos e entidades dos Poderes Executivos estadual, distrital e municipal.

§ 1º Constituem exceção a esta obrigatoriedade as informações cujo sigilo seja imprescindível à segurança da sociedade e do Estado, nos termos do art. 5º, inciso XXXIII, da Constituição e da Lei nº 11.111, de 5 de maio de 2005.

§ 2º Os dados geoespaciais disponibilizados no DBDG pelos órgãos e entidades federais, estaduais, distritais e municipais devem ser acessados, por meio do SIG Brasil, de forma livre e sem ônus para o usuário devidamente identificado, observado o disposto no § 1º.

Art. 4º Os órgãos e entidades do Poder Executivo federal deverão:

I - na produção, direta ou indireta, ou na aquisição dos dados geoespaciais, obedecer aos padrões estabelecidos para a INDE e às normas relativas à Cartografia Nacional? e

II - consultar a CONCAR antes de iniciar a execução de novos projetos para a produção de dados geoespaciais, visando a eliminar a duplicidade de esforços e recursos.

Art. 5º Compete ao Instituto Brasileiro de Geografia e Estatística - IBGE, como entidade responsável pelo apoio técnico e administrativo à CONCAR:

I - construir, disponibilizar e operar o SIG Brasil, em conformidade com o plano de ação para implantação da INDE, de que trata o inciso VIII do art. 6º?

II - exercer a função de gestor do DBDG, por meio do gerenciamento e manutenção do SIG Brasil, buscando incorporar-lhe novas funcionalidades?

III - divulgar os procedimentos para acesso eletrônico aos repositórios de dados e seus metadados distribuídos e para utilização dos serviços correspondentes em cumprimento às diretrizes definidas pela CONCAR para o DBDG?

IV - observar eventuais restrições impostas à publicação e acesso aos dados geoespaciais definidas pelos órgãos produtores?

V - preservar, conforme estabelecido na Lei nº 5.534, de 14 novembro de 1968, o sigilo dos dados estatísticos considerados dados geoespaciais de acordo com o § 1º do art. 2º?e

VI - apresentar as propostas dos recursos necessários para a implantação e manutenção da INDE.

Parágrafo único. O IBGE enviará à CONCAR, anualmente, relatório das atividades realizadas com base neste artigo.

Art. 6º Compete à CONCAR:

I - estabelecer os procedimentos para a avaliação dos novos projetos de que trata o inciso II do art. 4º?

II - homologar os padrões para a INDE e as normas para a Cartografia Nacional, nos termos do Decreto-Lei nº 243, de 28 de fevereiro de 1967, e do Decreto nº 89.817, de 20 de junho de 1984?

III - definir as diretrizes para o DBDG, com o objetivo de subsidiar a ação do IBGE, nos termos do inciso III do art. 5º?

IV - garantir que o DBDG seja implantado e mantido em conformidade com os Padrões de Interoperabilidade de Governo Eletrônico, mantidos pela Secretaria de Logística e Tecnologia da Informação, do Ministério do Planejamento, Orçamento e Gestão?

V - promover o desenvolvimento de soluções em código aberto e de livre distribuição para atender às demandas do ambiente de servidores distribuídos em rede, utilizando o conhecimento existente em segmentos especializados da sociedade, como universidades, centros de pesquisas do País, empresas estatais ou privadas e organizações profissionais?

VI - coordenar a implantação do DBDG de acordo com o plano de ação para implantação da INDE, de que trata o inciso VIII deste artigo?

VII - acompanhar, na forma do parágrafo único do art. 5º, as atividades desempenhadas pelo IBGE previstas no referido artigo? e

VIII - submeter ao Ministério do Planejamento, Orçamento e Gestão plano de ação para implantação da INDE, para atender ao estabelecido neste Decreto, até cento e oitenta dias após a sua publicação, contendo, entre outros, os

seguintes aspectos:

a) prazo para implantação das estruturas física e virtual do DBDG e do SIG Brasil?

b) prazo para a CONCAR homologar normas para os padrões dos metadados dos dados geoespaciais?

c) prazo para os órgãos e entidades do Poder Executivo federal disponibilizarem para a CONCAR e armazenarem, no servidor do sistema de sua responsabilidade, os metadados dos dados geoespaciais de seu acervo?

d) prazo para início da divulgação dos metadados dos dados geoespaciais e da disponibilização dos serviços relacionados, pelo SIG Brasil?

e) regras para disponibilização na INDE dos metadados de novos projetos ou aquisições de dados geoespaciais? e

f) recursos financeiros necessários para a implantação da INDE, ouvido o IBGE, nos termos do inciso VI do art. 5º, incluindo as necessidades do DBDG e do SIG Brasil, bem como os recursos financeiros necessários ao desenvolvimento de padrões, para divulgação da INDE, capacitação de recursos humanos e promoção de parcerias com entidades e órgãos públicos federais, estaduais, distritais e municipais.

Art. 7º Caberá à Secretaria de Planejamento e Investimentos Estratégicos, do Ministério do Planejamento, Orçamento e Gestão, promover, junto aos órgãos das administrações federal, distrital, estaduais e municipais, por intermédio da CONCAR, as ações voltadas à celebração de acordos e cooperações, visando ao compartilhamento dos seus acervos de dados geoespaciais.

Art. 8º Este Decreto entra em vigor na data de sua publicação.

Brasília, 27 de novembro de 2008?187º da Independência e 120º da República.

LUIZ INÁCIO LULA DA SILVA
*Paulo Bernardo Silva*

# ANEXO 2 – POLÍTICA DE DADOS DE COLEÇÕES E ACERVOS CIENTÍFICOS BIOLÓGICOS DO MUSEU PARAENSE EMÍLIO GOELDI – MPEG

## POLÍTICA DE DADOS DE COLEÇÕES E ACERVOS CIENTÍFICOS BIOLÓGICOS DO MUSEU PARAENSE EMÍLIO GOELDI – MPEG

Esta política diz respeito aos dados e metadados das coleções biológicas do Museu Paraense Emílio Goeldi, sob a responsabilidade da Coordenação de Pesquisa e Pós-graduação, que mantém, gerencia e desenvolve as diversas coleções científicas biológicas da Instituição, sendo assessorada pelo Conselho de Curadoria. Esta política baseia-se nos seguintes princípios:

- Respeito à legislação brasileira pertinente;
- Cooperação e sinergismo são elementos importantes na promoção do conhecimento científico;
- O conhecimento científico é incrementado por meio de ampla disseminação dos dados sobre a biodiversidade;
- A valorização dos dados como recurso institucional aumenta com o seu uso amplo e adequado, e diminui com o mau uso, má interpretação ou com desnecessárias restrições ao seu acesso;
- Os dados sobre biodiversidade, obtidos com o emprego de recursos públicos, devem ser de uso público e aberto, restringindo-se o acesso aos dados ainda não divulgados através de matéria impressa, comunicação oral ou meio eletrônico e outros que sejam oriundos de conhecimento tradicional;
- A gestão da informação sobre biodiversidade é um bem intelectual da instituição que detém a coleção.

**Objetivos desta política de dados**

- Assegurar que a comunidade científica, governo e sociedade em geral tenham acesso ao conjunto de dados científicos das coleções em tempo hábil e que os devidos créditos sejam atribuídos à instituição detentora.
- Orientar todas as instâncias envolvidas com as coleções biológicas do MPEG quanto à abrangência, propriedade, gestão, acesso, utilização, integração, e restrições de uso dos dados e metadados dessas coleções;
- Assegurar que a legislação brasileira pertinente seja obedecida, evitando-se o uso indevido dos dados pelos que a eles têm acesso.

**Definições gerais**

- Esta política abrange os dados e metadados biológicos, curatoriais, ambientais, espaciais e bibliográficos, tanto analógicos quanto digitais, vinculados aos espécimes, lotes, peças e observações que integram as coleções botânicas e zoológicas do MPEG, considerados neste documento como ***Dados das Coleções do MPEG***;
- Esta política aplica-se a todas as coleções biológicas do MPEG;

- Os usuários ao utilizarem os dados das coleções do MPEG, assumem sua concordância com os termos desta política e com normas adicionais específicas eventualmente estabelecidas no âmbito de cada coleção;
- Os dados das coleções do MPEG, analógicos e digitais, podem ser categorizados em três tipos: (a) *dados internos* (dados gerenciais de uma determinada coleção e não necessariamente destinados a uso do público externo à instituição); (b) *dados de acesso restrito* (dados com determinadas restrições de uso e que necessitam de autorização prévia individual para serem acessados pelo público externo); (c) *dados públicos* (dados sem qualquer restrição para serem acessados pelo público externo);
- Caberá a Coordenação de Pesquisa e Pós-Graduação, assessorada pelo Conselho de Curadoria, arbitrar eventuais conflitos ou resolver casos omissos ou questões relacionadas a esta política de dados.

**Propriedade dos dados**

- Os dados e informações mantidos sob a guarda e responsabilidade do MPEG são patrimônio da União e considerados de domínio público, respeitando-se as três categorias estabelecidas no item Definições Gerais deste documento.
- A informação contida nos bancos de dados das coleções do MPEG é regida pela legislação brasileira de propriedade intelectual, sendo que a propriedade e os direitos autorais sobre essa informação pertencem ao MPEG;
- Os bancos de dados eletrônicos das coleções representam extensões lógicas das coleções biológicas do MPEG e da sua documentação física, constituindo parte integrante dessas coleções, mantendo o MPEG sua propriedade e todos os direitos dela decorrente.

**Gestão dos dados**

- O armazenamento e a preservação dos dados digitais devem ser feitos em bancos de dados eletrônicos, a serem mantidos em computadores no âmbito de cada coleção biológica, sob a responsabilidade dos respectivos curadores. Devem também ser mantidos no(s) servidor(es) das coleções biológicas do MPEG, sob a responsabilidade do Serviço de Processamento de Dados (SPD);
- Os bancos de dados digitais de cada coleção estarão ligados a um servidor dedicado, mas não necessariamente interligados entre si. O servidor fará a integração dos bancos de acordo com as políticas de acesso, respeitando a definição de dados sensíveis estabelecida por cada curador e proporcionando a interface institucional necessária para a disponibilização ao público.
- Compete aos Curadores, em conjunto com os assessores de bio-geo-informática do Programa de Pesquisa em Biodiversidade e técnicos do SPD, onde convir, estabelecer e seguir normas e padrões que garantam a segurança, acessibilidade, qualidade, longevidade, integridade e interoperabilidade dos dados da coleção sob sua responsabilidade;
- Os curadores, com o apoio dos assessores em bio-geo-informática, onde convir, devem promover esforços para garantir a confiabilidade, qualidade e atualidade dos dados da coleção sob sua responsabilidade.

**Acesso e utilização dos dados**

- Os dados das coleções do MPEG são de utilização restrita a fins científicos, educacionais, gerenciais, de divulgação científica e de gestão pública. Nenhum dado das coleções (envolvendo, por exemplo, informação textual, digital, fotografia, imagem, reprodução ou publicação em qualquer formato) poderá ser utilizado com intenção comercial sem autorização expressa do representante legal da Instituição, sendo condição normal para tal autorização que os direitos autorais sejam atribuídos ao MPEG;
- O MPEG poderá, no âmbito de cada coleção, estabelecer medidas de controle, monitoramento e documentação de todo o acesso e uso de suas coleções biológicas e respectivos bancos de dados, sendo que o estabelecimento de tais medidas é de responsabilidade do respectivo curador, em conjunto com o NBGI, onde convir;
- O uso de qualquer dado de coleções do MPEG deve ser devidamente creditado à coleção do MPEG provedora do dado mediante a citação do seu nome e/ou acrônimo. Esse crédito deve ser feito em qualquer publicação, anúncio, correspondência ou demonstração pública que faça alusão ou mencione tal dado. Entretanto, o MPEG não permite a usuários externos a reprodução, publicação, distribuição ou re-impressão do total – ou parte substancial do total – das informações, registros, imagens, sons e observações de um ou mais dos bancos de dados de suas coleções biológicas, salvo exceções com a devida autorização do representante legal do MPEG;
- Os usuários das coleções do MPEG devem enviar aos respectivos curadores cópia de todas as publicações que façam uso de dados ou espécimes dessas coleções;
- Termos e condições adicionais para o uso dos dados das coleções do MPEG poderão ser necessárias, a critério de cada coleção provedora dos dados. No caso de acesso que implique em alteração de conteúdo dos bancos de dados, o usuário deverá demonstrar sua concordância com esses termos e condições mediante a assinatura de uma declaração formal (Termo de Compromisso de Uso dos Dados);
- Os usuários devem respeitar qualquer período de carência ou restrição de acesso que um conjunto de dados possa conter;
- O MPEG não poderá ser responsabilizado em nenhuma hipótese por qualquer dano, conseqüência ou prejuízo que a utilização dos dados de suas coleções biológicas tornados públicos venha, eventualmente, causar a pessoas físicas e/ou jurídicas.

**Integração de banco de dados**

- Os dados digitais das coleções devem ser armazenados em plataforma computacional, tanto de "hardware" quanto de "software", compatível com as orientações da política computacional do Serviço de Processamento de Dados (SPD);
- O MPEG promoverá esforços para colocar os bancos de dados de suas coleções biológicas para consulta aberta e "on line", primeiramente por meio de acesso direto via página eletrônica institucional, e, posteriormente por meio de integração a redes multi-institucionais provedoras de dados de biodiversidade. A instituição disponibilizará apenas os dados considerados públicos, de acordo com o item Definições Gerais do presente documento.

**Salvaguardas e restrições de uso**

- O acesso aos equipamentos ("hardware"), sistemas operacionais, programas e códigos subjacentes que suportam os bancos de dados eletrônicos das coleções é restrito a pessoal do quadro institucional. A possibilidade de acessar e manipular o conteúdo dos bancos de dados eletrônicos de uma coleção é limitado a usuários autorizados pelo curador da coleção em questão e mediante a assinatura de um Termo de Compromisso de Uso dos Dados (em anexo);
- Apesar dos esforços para prover dados acurados, o MPEG não fornece nenhuma garantia, expressa ou implícita, acerca da confiabilidade, integralidade e atualidade da informação contida nos bancos de dados eletrônicos das suas coleções, ou mesmo da sua aplicabilidade a qualquer propósito em especial. Os dados das coleções não devem, portanto, ser considerados como dados primários, cabendo ao usuário a responsabilidade pela conferência dos mesmos, antes de utilizá-los para qualquer finalidade pretendida.
- As informações dos bancos de dados eletrônicos do MPEG são fornecidas com o intuito de complementar, ao invés de substituir, o uso das coleções propriamente ditas;
- O MPEG reserva-se o direito de suspender, restringir ou bloquear o acesso a dados sensíveis dos bancos de dados eletrônicos de suas coleções a qualquer tempo. Esses dados sensíveis podem abranger informações referentes a determinados registros ou campos em função de situações tais como: (i) registros referentes a espécies com status de vulnerabilidade; (ii) registros com sérios problemas de qualificação dos dados; (iii) registros essenciais para pesquisas em andamento por pesquisadores do, ou vinculados ao, MPEG; (iv) registros que tenham alguma restrição específica (compromissos assumidos quando do recebimento do material como parte de uma doação por terceiros; propriedade intelectual; usos e outras).
- O curador da coleção provedora dos dados é o responsável pelo estabelecimento de critérios para a definição dos dados sensíveis dessa coleção, bem como pela tomada de decisão para suspender, restringir ou bloquear o acesso a tais dados.
- Ao critério do curador da coleção provedora dos dados, dados sensíveis poderão ser liberados para finalidades científicas e de gestão pública, mediante solicitação formal e escrita, na qual o usuário se comprometa a não utilizar a informação obtida de forma a prejudicar a conservação ambiental e/ou a alterar irreversivelmente o equilíbrio ecológico de uma região.

**Revisão desta política de dados**

- Os termos desta Política de Dados devem ser revisados e, se necessário, atualizados anualmente, ou extraordinariamente a qualquer momento, inclusive integrando outros acervos não biológicos da instituição.
- Caberá à Coordenação de Pesquisa e Pós-Graduação do MPEG coordenar essa revisão, após o que o documento deverá passar pelo Conselho de Curadoria, e a aprovação caberá ao Diretor do MPEG.

**Autores**

Ricardo de S. Secco

Alexandre B. Bonaldo

Alexandre Aleixo

Ana Lúcia C. Prudente

Ely Simone Gurgel

João Ubiratan Santos

Maria Inês Ramos

Maria de Nazaré Bastos

Orlando T. Silveira

Regina T. Lobato

Suely A. Marques

Wolmar B. Wosiacki

**Revisão**

Ima Célia Guimarães Vieira

Nilson Gabas Júnior

**Assessoria**

Benedita Barros

*Aprovado pela Diretoria em 18.05.2007*

| | |
|---|---|
| Ima Célia Guimarães Vieira | Nilson Gabas Jr. |
| Diretora | Coordenador de Pesquisa e Pós-Graduação |

# ANEXO 3 - POLÍTICA DE ACESSO A DADOS E INFORMAÇÕES CIENTÍFICAS DO INSTITUTO DE PESQUISAS JARDIM BOTÂNICO DO RIO DE JANEIRO

MINISTÉRIO DO MEIO AMBIENTE
INSTITUTO DE PESQUISAS JARDIM BOTÂNICO DO RIO DE JANEIRO

## PORTARIA JBRJ Nº 077/2012, DE 19 DE JULHO DE 2012.

O **PRESIDENTE DO INSTITUTO DE PESQUISAS JARDIM BOTÂNICO DO RIO DE JANEIRO**, no uso das atribuições que lhe conferem a Lei nº 10.316, de 06 de dezembro de 2001, publicada no Diário Oficial da União de 07 de dezembro de 2001, o Decreto nº 6.645, de 18 de novembro de 2008, publicado no Diário Oficial da União de 19 de novembro de 2008, retificado no DOU de 20 de novembro de 2008 e no DOU de 27 de fevereiro de 2009 e o disposto no Regimento Interno aprovado pela Portaria Ministerial nº 401, de 11 de novembro de 2009, publicada no Diário Oficial da União, de 13 de novembro de 2009,

**RESOLVE:**

**Art. 1º** Instituir a Política de Acesso a Dados e Informações Científicas do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro – JBRJ.

**Art. 2º** Esta Política diz respeito aos dados das coleções do JBRJ, sob a responsabilidade da Diretoria de Pesquisa Científica – DIPEQ e se baseia nos seguintes princípios:

I - Respeito à legislação brasileira pertinente;

II - Cooperação e sinergismo como elementos importantes na promoção do conhecimento científico;

III - O incremento do conhecimento científico como resultado da ampla disseminação dos dados sobre a biodiversidade.

**Art. 3º** São objetivos desta Política:

I - Assegurar que a comunidade científica, governo e sociedade em geral tenham acesso ao conjunto de dados científicos das coleções em tempo hábil e que os devidos créditos sejam atribuídos;

II - Orientar todas as instâncias envolvidas com as coleções biológicas do JBRJ quanto à abrangência, propriedade, gestão, acesso e utilização, integração, e restrições de uso dos dados dessas coleções.

**Art. 4º** Esta Política abrange os dados de coleções, tanto analógicos quanto digitais, vinculados aos espécimes e amostras que integram as coleções sob a responsabilidade da DIPEQ/JBRJ, considerados neste documento como Dados das Coleções do JBRJ.

**Art. 5º** Os Dados das Coleções do JBRJ são enquadrados nas seguintes categorias:

I - Dados sem Restrição: são dados cujo acesso público e sua publicação em formato analógico ou digital não possuem qualquer restrição de acesso;

II - Dados em Carência: são aqueles cuja restrição ao acesso e publicação é temporária e necessária para garantir o tratamento, análise e utilização em publicação original por parte dos seus autores;

III - Dados Sensíveis: são aqueles referentes à localização e uso de espécies de interesse econômico, comercial e ameaçadas de extinção; e referentes a dados sobre espécies obtidos a partir do conhecimento tradicional associado, e de sub-amostras representativas do patrimônio genético acessado, conforme estabelecido pela legislação e pelo Conselho de Gestão do Patrimônio Genético - CGEN.

**Art. 6º** Cabe ao Diretor de Pesquisa Científica, ouvidos o Assessor de Informações Científicas e o Assessor de Coleções Científicas, definir e revisar os prazos de carência dos dados e informações em carência, e as espécies cujos dados são enquadrados na categoria de Dados Sensíveis.

**§1°** Apesar dos esforços para prover dados acurados, o JBRJ não fornece nenhuma garantia, expressa ou implícita, acerca da confiabilidade, integralidade e atualidade da informação contida nos dados das suas coleções, ou mesmo da sua aplicabilidade a qualquer propósito em especial.

**§2°** O uso de qualquer dado de coleções e de informações a partir de pesquisas geradas no JBRJ deve ser devidamente creditado ao JBRJ e, quando for o caso, ao(s) autor(es) provedor(es) das informações mediante a citação do(s) nome(s) e/ou acrônimo.

**§3°** Os usuários, sejam internos ou externos, ao fazerem uso dos dados e informações das coleções da DIPEQ/JBRJ, assumem sua concordância com os termos, diretrizes, normas e procedimentos adicionais e específicos desta Política.

**§4°** Cabe à DIPEQ/JBRJ, em conjunto com o Assessor de Informações Científicas, arbitrar eventuais conflitos ou resolver casos omissos ou questões relacionadas a esta política de dados.

**Art. 7º** Os termos desta Política de Acesso a Dados e Informações devem ser revisados e, se necessários, atualizados anualmente, ou extraordinariamente a qualquer momento.

**Parágrafo único.** Cabe à Assessoria de Informações Científicas, coordenar a revisão, cuja aprovação compete ao Diretor de Pesquisa Científica.

**Art. 8º** Esta Política se vincula à Política de Coleções da DIPEQ, estabelecida pela Assessoria de Coleções, incluindo as diretrizes, normas e procedimentos definidos na Política de Coleções.

**Art. 9º** Revogam-se as disposições em contrário.

**Art. 10.** Esta Portaria entra em vigor na data de sua assinatura.

**LISZT B. VIEIRA** Presidente do
Instituto de Pesquisas Jardim
Botânico do Rio de Janeiro

# ANEXO 4 – POLÍTICA DE DADOS E INFORMAÇÕES SOBRE BIODIVERSIDADE DO INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE

**INSTRUÇÃO NORMATIVA Nº 2, DE 25 DE NOVEMBRO DE 2015.**

Institui a Política de Dados e Informações sobre Biodiversidade do Instituto Chico Mendes de Conservação da Biodiversidade e dispõe sobre sua disponibilização, acesso e uso (Processo nº 02070.001239/2015- 93).

**O PRESIDENTE DO INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE** - Instituto Chico Mendes, no uso das competências que lhe confere o Art. 21, Anexo I, do Decreto n°7.515 de 8 de julho de 2011, publicado no Diário Oficial da União de 11 de julho de 2011, e pela Portaria n° 899, de 14 de maio de 2015, do Ministro de Estado Chefe da Casa Civil da Presidência da República, publicada no Diário Oficial da União de 15 de maio de 2015, resolve:

**Art. 1°** Instituir a Política do Instituto Chico Mendes de Conservação da Biodiversidade para Dados e Informações sobre Biodiversidade, visando regulamentar a disponibilização, o acesso e o uso dos dados e informações custodiados pelo Instituto em suas bases e sistemas de informação.

**Parágrafo único.** As unidades do Instituto Chico Mendes responsáveis pela gestão de sistemas de informação ou bases de dados sobre biodiversidade poderão elaborar definições e regras específicas para disponibilização, acesso e uso dos dados e informações, desde que em conformidade com o regramento disposto na presente Instrução Normativa.

**Art. 2°** Para fins desta Instrução Normativa, considera-se:

I - autor: pessoa(s) ou instituição(ões) a quem se atribui a responsabilidade sobre a geração de um determinado dado ou informação, conforme definido na norma ou forma de funcionamento de cada sistema de informação ou base de dados;

II - biodiversidade: variabilidade de organismos vivos de todas as origens, compreendendo, dentre outros, os ecossistemas terrestres, marinhos e outros ecossistemas aquáticos e os complexos ecológicos de que fazem parte; compreendendo ainda a diversidade dentro de espécies, entre espécies e de ecossistemas;

III - dado: sequencia de símbolos quantificados ou quantificáveis referentes a um objeto ou evento, podendo consistir em textos, números, datas, imagens, arquivos vetoriais, entre outros;

IV - informação: afirmação realizada a partir da organização, análise ou interpretação de um conjunto dados;

V - carência: período no qual o acesso por terceiros ou a publicação de dados ou informações sobre biodiversidade custodiados pelo Instituto estão temporariamente restritos, para garantir o tratamento, análise e utilização em publicação por seus autores.

VI - dados ou informações sensíveis: são aqueles para os quais a disponibilização pode comprometer a proteção de espécies ou ecossistemas.

**Art. 3º** Os autores de dados ou informações sobre biodiversidade, ao inseri-los nos sistemas de informações geridos pelo Instituto Chico Mendes, autorizam a sua custódia pelo Instituto nos termos desta Instrução Normativa.

**Art. 4°** Os dados e informações custodiados serão enquadrados nas seguintes categorias:

I - "sem carência"

II - "em carência"

**§1°** Os autores de dados ou informações poderão definir um período de carência de até cinco anos para sua disponibilização.

**§2°** não existirá período de carência para dados e informações resultantes de pesquisas ou trabalhos técnicos contratados pelo Instituto Chico Mendes.

**Art. 5°** Os dados e informações inseridos nas bases de dados ou nos sistemas de informação previamente à publicação desta Instrução Normativa e para os quais não havia no sistema de origem a possibilidade de definição de período de carência pelo autor, seguirão o seguinte regramento:

I - para os dados e informações inseridos até 2011, passa a vigorar o período de carência de um ano a partir da data de publicação da presente Instrução;

II - para os dados e informações inseridos a partir de 2012, passa a vigorar o período de carência de cinco anos a partir da data de inserção dos dados nas bases ou sistemas.

**Parágrafo único** - Os períodos de carência poderão ser reduzidos mediante autorização dos autores dos dados e informações.

**Art. 6°** Os dados e informações em período de carência poderão ser usados pelo Instituto, independente da autorização dos seus autores, nas seguintes hipóteses:

I - para o planejamento de ações voltadas à gestão das unidades de conservação federais e à conservação da biodiversidade, desde que não implique na publicação dos dados ou informações;

II - para publicações técnicas ou científicas envolvendo análises e sínteses de informação sobre animais e plantas em níveis taxonômicos igual ou superior à Classe.

**Art. 7°** O Instituto Chico Mendes poderá restringir temporariamente a divulgação de dados ou informações considerados sensíveis, mesmo fora do período de carência.

**Parágrafo único.** O período e as formas de restrição de dados e informações sensíveis serão formalizados em ato administrativo específico.

**Art. 8°** O Instituto Chico Mendes é responsável por organizar e disponibilizar os dados e informações inseridos em suas bases e sistemas, cabendo ao cidadão que acessá-los aferir a sua confiabilidade, integralidade e atualidade.

**Art. 9°** O Instituto Chico Mendes tornará disponível a identificação dos autores dos dados e informações custodiados, assim como dos sistemas de informação que são fonte original do conteúdo sobre biodiversidade, para seu devido referenciamento nas publicações que fizerem uso deste material.

**§1°** Os autores de dados ou informações que não desejarem ser citados deverão assim indicar ao Instituto.

**§2°** Os autores das publicações que utilizarem os dados ou informações de que trata o caput são responsáveis pela citação da sua autoria e fonte.

**Art. 10º** As unidades gestoras das bases e sistemas de informação sobre biodiversidade do Instituto terão o prazo de doze (12) meses para realizarem os ajustes necessários à sua adequação a esta Instrução Normativa.

**Art. 11º** Esta Instrução Normativa entra em vigor na data de sua publicação.

CLÁUDIO CARRERA MARETTI

DOU nº 226 de 26 de novembro de 2015 – Seção 1 – p. 83.

# ANEXO 5 – POLÍTICA DE DADOS DO PROGRAMA DE PESQUISA EM BIODIVERSIDADE - PPBIO

PORTARIA Nº 693, DE 20 DE AGOSTO DE 2009

Institui, no âmbito do Programa de Pesquisa em Biodiversidade - PPBio, a Política de Dados.

O MINISTRO DE ESTADO DA CIÊNCIA E TENOLOGIA,
no uso de suas atribuições legais, em especial as que lhe confere o art. 87, parágrafo único, inciso II, da Constituição Federal, e tendo em vista o disposto no Decreto 4.339 de 22 de agosto de 2002 e o cumprimento dos dispositivos da Convenção sobre Diversidade Biológica, aprovada pelo Decreto Legislativo nº 2, de 3 de fevereiro de 1994, e promulgada pelo Decreto nº 2.519, de 16 de março de 1998; e

Considerando que o Programa de Pesquisa em Biodiversidade - PPBio tem entre seus objetivos gerais o fomento à geração e disseminação de informações e conhecimento sobre a biodiversidade brasileira para diferentes segmentos da sociedade;

Considerando que o Programa de Pesquisa em Biodiversidade - PPBio fomenta a criação de sistemas de informação, de bases de dados e gerenciamento de repositórios da informação sobre a biodiversidade brasileira;

Considerando que os dados gerados no âmbito do Programa de Pesquisa em Biodiversidade - PPBio podem ter interesse comercial e o seu uso pode gerar consequências econômicas e ambientais;

Considerando a necessidade de um arcabouço de princípios, regras e orientações para todos os participantes do Programa de Pesquisa em Biodiversidade e usuários das bases de dados geradas no âmbito do Programa, no que diz respeito à abrangência, à coleta, ao armazenamento, à propriedade, à autoria, ao compartilhamento, à citação, ao acesso e uso dos dados e das bases de dados;

Considerando a necessidade de evitar conflitos e de obter compromissos sobre as questões de propriedade intelectual;

Considerando a necessidade de observar o que dispõe a legislação vigente de propriedade intelectual e inovação, especialmente no que concerne à proteção aos direitos autorais, à propriedade industrial e à informação confidencial; resolve:

Art. 1º Instituir, no âmbito do Programa de Pesquisa em Biodiversidade - PPBio, a Política de Dados, com o objetivo de promover o gerenciamento das informações para os dados coletados sobre a biodiversidade brasileira e gerados no âmbito do Programa, seus acessos, usos e disseminação, na forma do anexo a esta portaria.

Art. 2o Esta portaria entra em vigor na data de sua publicação.

SERGIO MACHADO REZENDE

ANEXO

DA POLÍTICA DE DADOS DO PROGRAMA DE PESQUISA EM BIODIVERSIDADE - PPBio

DAS DEFINIÇÕES GERAIS

Metadados: conjunto de informações que acompanham e descrevem as características dos dados biológicos, ambientais, socioambientais e espaciais e as condições de sua coleta, por exemplo: local de coleta, data de coleta, nome do coletor, latitude e longitude, imagens digitais ou fotos, entre outras.

Dados: informações biológicas, ambientais, socioambientais ou espaciais adquiridas com recursos financeiros ou logísticos do PPBio ou por ações amparadas por este. Podem ser caracterizadas como dados digitais ou conjuntos de dados armazenados e gerenciados por computadores; dados analógicos, oriundos de atividades do PPBio, ainda que não digitalizados, como anotações de campo, planilhas, cadernetas de coleta; e quaisquer relatórios ou mapas produzidos, em formato digital ou analógico, resultado da compilação, análise, reunião ou organização, utilizando como fonte conjuntos de dados do PPBio.

Dados preliminares: são aqueles capazes de fornecer informações básicas descritivas do material biológico coletado (exemplos: morfologia, coloração, tamanho), ou ainda, as informações associadas a esse, sem garantir a identificação taxonômica precisa do mesmo.

Dados consolidados: são aqueles capazes de fornecer informações refinadas e completas e, tanto quanto possível, definitivas sobre o material coletado, incluindo a identificação taxonômica.

Dados ostensivos: são dados preliminares ou consolidados que após respeitado o período de embargo podem ser utilizados sem restrição cujo acesso pode ser franqueado ao público em geral.

Dados sensíveis: são dados preliminares ou consolidados que, se liberados ao acesso público, possam resultar em efeito adverso ao local e/ou às comunidades de origem da mesma e por isso, passível de restrição. Podem ser considerados dados sensíveis (a) a localização de espécies que estejam na lista de espécies ameaçadas de extinção; (b) dados de espécie que possa ser roubada ou traficada por sua raridade ou valor econômico (considerando sua potencialidade: como fornecedora de produtos que venham a ser utilizados na indústria farmacêutica ou química; como agente de controle biológico; entre outras); (c) a localização de habitats e sítios arqueológicos, culturais ou históricos cujo acesso possa ameaçar sua integridade; (d) informações utilizadas em decisões de Política de Estado que possam vir a interferir no alcance das metas e objetivos da mesma.

Casos particulares que não estejam listados nessa Política deverão ser encaminhados à Coordenação Executiva para avaliação pelo Comitê Científico e aprovação do Conselho Diretor.

Núcleos de Biogeoinformática: unidades de gerenciamento de sistemas informatizados, aplicativos, bases de dados e metadados, instituídos e mantidos pelos núcleos executores e núcleos regionais.

Comitê Gestor de Informação: colegiado responsável pela deliberação sobre questões técnicas, administrativas, infra-estruturais e operativas que venham a ocorrer durante a operacionalização e gerenciamento de dados e informações do PPBio.

Participantes do Programa: os Núcleos Executores, os Núcleos Regionais, coordenadores de projetos e coordenadores de redes temáticas e todos os pesquisadores, colaboradores, estudantes, técnicos e bolsistas vinculados a esses núcleos que assinarem o termo de compromisso com esta Política.

Núcleos Executores: instituições assim designadas por Termo de Compromisso e Gestão ou Convênio firmado no âmbito do PP-Bio.

Núcleos Regionais: instituições que trabalham em parceria e de forma coordenada com os núcleos executores.

Indicador de confiabilidade do dado: parâmetro que qualifica o dado quanto à precisão e acurácia da informação e que demonstra a confiabilidade e qualidade dos dados inseridos na base de dados do PPBio.

Divulgação ampla: disponibilização de metadados e dados ostensivos a todos os interessados a partir do portal do PPBio.

Divulgação restrita: disponibilização de dados sensíveis permitida mediante autorização ou senha de acesso.

Período de embargo: período no qual, dados sob restrições de uso e acesso, não são disponibilizados pelo portal, mas são passíveis de visualização pelo Comitê Gestor de Informação.

DAS OBRIGAÇÕES, ATRIBUIÇÕES E COMPETÊNCIAS

2. Sobre as obrigações, atribuições e competências.

2.1- Dos Núcleos de Biogeoinformática.

Seguir normas, padrões e procedimentos estabelecidos para sua atuação pelo Comitê Gestor de Informação;

Implementar mecanismos que garantam a segurança, acessibilidade, qualidade, longevidade, integridade e interoperabilidade dos dados e metadados do PPBio;

Cadastrar os usuários e participantes do programa na instituição na qual estão situados, mantendo seus termos de compromisso;

Monitorar a atualização dos indicadores de qualidade associados aos dados que integram a base de dados;

Respeitar o sigilo de informações consideradas sensíveis;

Criar e manter um portal de acesso aos dados e metadados do PPBio na Internet;

Desenvolver, avaliar e adotar ferramentas computacionais e os aplicativos necessários ao registro, gestão, documentação, análise, integração, busca, acesso, armazenamento, segurança e publicação dos dados do PPBio;

Manter registro das publicações que utilizaram dados do PPBio.

2.2 Do Comitê Gestor de Informação.

Estabelecer normas, padrões e procedimentos para atuação dos núcleos de biogeoinformatica;

Revisar e atualizar anualmente, ou quando necessário discutir, juntamente com o Comitê Científico, os termos desta Política de dados;

Aprovar os parâmetros definidos pelos NBGIs em relação às ferramentas computacionais e os aplicativos necessários ao registro, gestão, documentação, análise, integração, busca, acesso, armazenamento, segurança e publicação dos dados e metadados do PPBio;

Estabelecer normas e padrões que garantam a segurança, acessibilidade, qualidade, longevidade, integridade e interoperabilidade dos dados e metadados do PPBio.

Definir e detalhar processos e padrões de armazenagem, segurança, recuperação, análise e publicação dos dados e metadados do PPBio;

Definir normas e procedimentos para o processo de replicação das bases de dados e metadados em servidores de outras instituições;

Promover o acesso e utilização eficiente dos dados por parte dos participantes, observando a legislação vigente;

Decidir, juntamente com o responsável pela inserção dos dados, sobre seu acesso por participantes em período de embargo;

Credenciar os participantes responsáveis pela inserção de dados no banco do PPBio conforme indicação do coordenador do projeto;

Decidir, juntamente com a instituição responsável pela obtenção dos dados, ouvido o Comitê Científico, sobre dados sensíveis que devam ser divulgados de forma restrita, assim como sobre os pedidos de acesso de terceiros;

Controlar o acesso às bases de dados mediante o fornecimento de senhas ou de emissão de autorizações de acesso.

2.3. Do Conselho Diretor.

Aprovar as revisões da Política de Dados propostas pelo Comitê Gestor de Informação e pelo Comitê Científico;

Sugerir mudanças nos aspectos da Política de Dados;

Arbitrar eventuais conflitos, resolver casos omissos, excepcionais ou questões relacionadas a esta Política de Dados.

2.4. Do Comitê Científico.

Orientar o Comitê Gestor de Informação e assessorar o Conselho Diretor;

Definir as categorias de dados sensíveis utilizados nesta Política e deliberar sobre a inclusão ou exclusão de dados da categoria de dados sensíveis para fins de divulgação restrita;

Sugerir ações ao Comitê Gestor de Informação e aos Núcleos de Biogeoinformática;

Revisar e atualizar anualmente ou quando necessário, juntamente com o Comitê Gestor de Informação, os termos desta política de dados;

Sugerir ao Conselho Diretor ações e estratégias de comunicação para disseminar conhecimentos de biodiversidade à sociedade.

2.5. Dos Participantes.

Aceitar o teor dessa política de dados por meio de assinatura de termo de compromisso;

Repassar ao PPBio todos os dados, metadados ou conjunto de dados gerados com recursos do programa, respeitados os prazos e condições estipulados nesta Política;

Classificar os dados gerados em relação às categorias de dados sensíveis definidas pelo comitê científico;

Responsabilizar-se pela qualidade e repasse de todas as informações ao banco de dados do PPBio;

A inclusão de dados obtidos por outros programas nas bases do PPBio deverá respeitar preceitos legais e políticas institucionais, devendo ser acompanhada de autorização escrita da instituição de origem.

Este documento pode ser verificado no endereço eletrônico http://www.in.g o v. b r / a u t e n t i c i d a d e . h t m l , Documento assinado digitalmente conforme MP no2.200-2 de 24/08/2001, que institui a pelo código 00012009082100008 Infraestrutura de Chaves Públicas Brasileira - ICP-Brasil.

Nº 160, sexta-feira, 21 de agosto de 2009 1 9ISSN 1677-7042 Nº 160, sexta-feira, 21 de agosto de 2009 1 9ISSN 1677-7042

DA GESTÃO E AUTORIA DOS DADOS

3. Sobre a gestão e autoria dos dados.

Os dados/metadados ou conjunto de dados/metadados gerados com recursos do PPBio são de interesse público para o desenvolvimento científico -tecnológico e sua gestão é de responsabilidade do Ministério da Ciência e Tecnologia.

O PPBio deverá resguardar a autoria dos dados nas bases de dados e nas publicações resultantes.

A co-autoria ou outras formas de citação da participação na geração, análise e publicação dos dados deverão ser definidas pelas partes envolvidas, refletindo a participação intelectual, de acordo com o código de ética da ciência.

DAS BASES DE DADOS - USOS E ACESSOS

4. Sobre as condições de uso e acesso das bases de dados.

4.1. As bases de dados e metadados do PPBio serão protegidas por mecanismos adequados de prevenção e proteção à acessos não autorizados.

4.2 Os dados coletados, gerados e disponibilizados no âmbito do PPBio são de utilização prioritária para fins educacionais, culturais, científicos, de divulgação e de gestão pública. O acesso e uso com intenção comercial ou de forma que possa resultar na geração de produtos ou processos passíveis de exploração econômica, deverá ocorrer mediante a celebração de contrato entre as partes interessadas, observada a legislação pertinente e as disposições desta Política.

4.3. O acesso aos dados sensíveis e àqueles em período de embargo far-se-á de forma restrita por meio de autorização do Comitê Gestor de Informação mediante consulta ao responsável pela inserção de dados no banco do PPBio e demais partes interessadas.

4.4. O acesso aos dados via portal na Internet deverá ser feito mediante declaração de aceitação das condições de uso e acesso por meio da assinatura de um Termo de Compromisso disponível no portal.

4.5. Recomenda-se a todo usuário, no caso de encontrar um dado que julgue incorreto, informar por meio de formulário próprio disponível no portal, ao pesquisador (ou grupo de pesquisa) responsável pela inserção do dado e ao Comitê Gestor de Informação, para que estes possam avaliar e, quando couber, providenciar a correção sugerida.

4.6. Os pesquisadores, as instituições participantes do PPBio, assim como o Ministério da Ciência e Tecnologia, não poderão ser responsabilizados em nenhuma hipótese por qualquer dano, conseqüência ou prejuízo que a utilização dos dados tornados públicos venha eventualmente causar, seja a pessoas físicas, seja a pessoas jurídicas.

4.7. Todos os produtos resultantes da utilização de dados e metadados do PPBio deverão ser acompanhados dos devidos créditos ao PPBio.

DA PROPRIEDADE INTELECTUAL

5. Sobre a proteção e propriedade intelectual.

5.1. Os produtos e processos decorrentes de informações e pesquisas concebidas ou executadas no âmbito do PPBio poderão ser protegidos e/ou patenteados segundo a legislação vigente, desde que seja observado o disposto no subitem 4.2.

DOS PRAZOS

6. Sobre os prazos.

6.1. Os metadados devem ser disponibilizados ao Comitê Gestor de Informação no prazo máximo de 30 dias após a coleta dos dados, tornando-se passíveis de consulta pública via portal na Internet em um prazo máximo de 7 dias após o repasse ao Comitê Gestor de Informação.

6.2. Os dados preliminares devem ser disponibilizados pelos autores ao Comitê Gestor de Informação no prazo máximo de 12 meses após a coleta, podendo ser nesse período visualizado pelo CGI. Findo esse período, os dados serão tratados como ostensivos, exceto se houver solicitação de prorrogação ao Conselho Diretor. Dados consolidados devem ser disponibilizados pelos autores ao Comitê Gestor de Informação no prazo máximo de 24 meses após a data da coleta. Em casos excepcionais, o prazo poderá ser estendido, desde que autorizado pelo Conselho Diretor do PPBio.

# ANEXO 6 – POLITICA DE DADOS DO PROGRAMA DE PESQUISAS ECOLÓGICAS DE LONGA DURAÇÃO - PELD

## POLITICA DE DADOS DO PROGRAMA DE PESQUISAS ECOLÓGICAS DE LONGA DURAÇÃO - PELD

RN-009/2016

Institui a Política de Dados do Programa de Pesquisa Ecológica de Longa Duração (PELD), com o objetivo de regulamentar as formas de disponibilização, acesso e uso dos dados gerados pelos pesquisadores da rede PELD.

O Presidente do CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPq, no uso das atribuições que lhe são conferidas pelo Estatuto aprovado pelo Decreto nº 7.899, de 04/02/2013?

considerando que os dados coletados no âmbito da rede PELD são de interesse público, tendo valor inestimável para a gestão ambiental sustentável, e devem, portanto, estar disponíveis para a sociedade?

considerando que a disponibilização de dados em um repositório de acesso público deve ser regulamentada no sentido de se proteger a autoria e assegurar a perenidade dos dados dados, além de promover ampla colaboração científica?

considerando que o PELD é membro da International Long Term Ecological Research (ILTER), que congrega 40 países membros e suas redes de Pesquisa Ecológica de Longa Duração?

considerando que o Brasil é membro do Global Biodiversity Information Facility (GBIF), que é uma rede global de países e organizações criada para facilitar a mobilização, acesso, descoberta e uso da informação sobre a ocorrência de espécies ao redor do planeta?

considerando ainda a Lei de Acesso à Informação (Lei nº 12.527, de 18/11/2011), destinada a assegurar o direito fundamental de acesso à informação, bem como a divulgação de informações de interesse público, entre outros fins?e

e em conformidade com decisão da Diretoria Executiva em sua 32ª (trigéssima segunda) reunião de 16/12/2015,

**R E S O L V E:**

**1.** Instituir a Política de Dados do Programa de Pesquisa Ecológica de Longa Duração (PELD), com o objetivo de regulamentar as formas de disponibilização, acesso e uso dos dados gerados pelos pesquisadores da rede PELD.

**2**. Para efeitos desta RN, colocam-se as seguintes definições:

**2.1**. Dados ¿ Informações biológicas, ambientais, sócio-ambientais ou espaciais adquiridas com recursos financeiros ou logísticos do PELD ou ações amparadas por este, reunidos em um arquivo digital, do tipo tabela ou planilha, para alimentação do repositório de dados.

**2.2.** Existem três tipos de dados: **preliminares, públicos e sensíveis**:

2.2.1. **Dados preliminares** são aqueles que já se encontram inseridos no repositório, sendo relacionados a trabalhos

ainda não publicados e que, por este motivo, são considerados preliminares, e serão de acesso restrito durante o período de embargo.

2.2.2**. Dados públicos** são aqueles sem restrições de acesso pelo usuário.

2.2.3**. Dados sensíveis** são aqueles que, se liberados ao acesso público, podem resultar em efeitos adversos à biota, ecossistemas ou populações humanas locais, e por isso são passíveis de restrições de acesso pelo usuário. Podem ser considerados sensíveis os seguintes dados:

2.2.3.1. Localização geográfica de espécies ameaçadas de extinção?

2.2.3.2. Dados sobre espécies de elevado valor ou potencial econômico que possam ser objeto de tráfico ou caça?

2.2.3.3 Localização de sítios arqueológicos.

**2.3. Metadados -** Conjunto de informações autoexplicativo sobre o pacote de dados, definido no repositório, que compila as principais características do pacote de dados em padrões internacionalmente aceitos, permitindo conhecer a origem destes e as condições de sua amostragem. Os metadados permitem localizar um determinado pacote de dados através da ferramenta de busca do repositório, disponibilizando o contato com o gestor e autor dos dados.

**2.4. Pacote de dados -** Conjunto de documentação estruturada que contém os metadados e uma ou mais tabelas de dados associados.

**2.5. Autor dos dados -** Indivíduo responsável pela produção dos dados, bem como sua disponibilização ao gestor de dados do sítio.

**2.6. Usuário dos dados -** Pessoa a quem é dado o acesso aos conjuntos de dados, observadas eventuais restrições de acesso, e mediante cadastro e aceitação dos termos e condições de uso.

**2.7. Gestor de dados do sítio -** Pessoa indicada pelo coordenador do sítio como responsável pela articulação dos pesquisadores de sítio com o repositório de dados, apoio à alimentação de dados, e verificação e orientação aos autores quanto à adequação dos conjuntos de dados às características exigidas pelo repositório.

**2.8. Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr)** - Sistema –online- destinado a integrar informações sobre a biodiversidade e os ecossistemas brasileiros, através da articulação de diversas bases de dados nacionais e estrangeiras, a fim de subsidiar a pesquisa e apoiar os tomadores de decisão na criação e implementação de políticas públicas.

**2.9. Repositório de dados ecológicos -** Sistema desenvolvido com finalidades de arquivamento, organização e acesso à informação referente aos dados ecológicos, disponibilizado pelo SiBBr, mediante cadastro e aceitação dos termos e condições de uso do repositório.

**3. Alimentação de Dados**

**3.1.** Todos os pacotes de dados relativos à pesquisa financiada com recursos da Rede PELD deverão ser incluídos no Repositório de Dados PELD tão logo sejam disponibilizados ao gestor de dados do sítio, respeitando-se os seguintes prazos máximos:

3.1.1. **Metadados:** devem ser disponibilizados todos até a metade da vigência original do projeto ou sempre que solicitado pelo CNPq, em função de ações de Acompanhamento & Avaliação de projetos.

3.1.2. **Dados**: devem ser integralmente disponibilizados até o prazo final para prestação de contas técnico-financeira do projeto, ou seja, até 60 dias após a vigência final do projeto.

**4. Autorizações de acesso**

**4.1.** Metadados serão de acesso público tão logo sejam disponibilizados?

**4.2.** Dados públicos poderão ser acessados por qualquer usuário, mediante cadastro e aceite dos termos e condições de uso.

**4.3.** Dados preliminares serão de acesso restrito ao(s) autor e gestor de dados, durante o período de embargo de até dois anos após o término da vigência original do projeto, prorrogável por mais um ano.

4.3.1. Após o período de embargo, os dados preliminares serão tornados públicos automaticamente.

**4.4.** Dados sensíveis são, a priori, de acesso restrito ao(s) autor(es) e gestor de dados do sítio. O acesso poderá ser franqueado a tomadores de decisão, mediante cadastro específico no SiBBr.

4.4.1. Para ter acesso a um pacote de dados sensíveis, o tomador de decisão deverá informar a justificativa, que será encaminhada ao(s) autor(es) dos dados, junto com uma notificação de que aquele pacote de dados, a priori de acesso restrito, foi acessado por determinado tomador de decisão.

**4.5.** As solicitações de restrição de acesso devem ser feitas pelo autor dos dados ao gestor de dados do sítio, mediante autorização emitida pelo coordenador de sítio.

**4.6**. As restrições de acesso aos dados serão especificadas sobre cada registro ou, eventualmente, uma tabela de dados inteira, mas não sobre todo o conjunto de dados de um determinado sítio PELD. Cabe ao coordenador do sítio decidir sobre as restrições de acesso a cada registro/tabela.

**4.7.** O autor ou gestor de dados depositados no repositório de dados ecológicos são os responsáveis pela atualização e correção dos dados, sempre que necessário. Somente o próprio autor ou gestor poderá editar seus dados.

**5. Termos e condições de uso do repositório PELD - dados públicos**

**5.1.** O uso dos dados é restrito a fins educacionais, acadêmicos, de pesquisa, recreacionais e outras finalidades não-lucrativas. O uso para quaisquer finalidades lucrativas requer autorização explícita do autor dos dados.

5.1.1 A utilização dos dados para finalidades lucrativas sujeita o usuário às sanções legais cabíveis, conforme o disposto na Lei nº 9.279/1996.

**5.2.** A qualidade e/ou veracidade dos dados não pode ser garantida pelo repositório, sendo o seu uso de responsabilidade do usuário.

**5.3.** Os dados são cedidos somente ao usuário, mediante cadastro. A redistribuição de dados a terceiros não é

permitida sem autorização explícita do autor dos dados.

**5.4.** Os usuários são encorajados a convidar o autor dos dados a participar intelectualmente dos trabalhos desenvolvidos a partir do compartilhamento dos dados.

**5.5.** Quando houver dúvida sobre inclusão de autores, o critério de maior abrangência deve ser preferido.

**5.6** O reconhecimento da autoria dos dados é obrigatório em todas as utilizações de conjuntos de dados. As citações devem conter obrigatoriamente as seguintes informações: autor(es)? ano de publicação, título do pacote de dados, sítio PELD (nome e sigla, coordenador do sítio), Identificador SiBBr, repositório PELD no SiBBR, data de acesso.

**5.7.** As fontes de financiamento também devem ser citadas em todas as utilizações de conjuntos de dados, conforme referenciadas nos metadados associados.

**5.8.** O repositório enviará notificações automáticas ao autor dos dados sempre que um pacote de dados for gravado por um usuário.

**5.9** O usuário compromete-se a notificar o autor dos dados sempre que um trabalho for publicado utilizando, parcial ou integralmente, dados acessados a partir do repositório.

**5.10.** Ao acessar um pacote de dados, o usuário concorda com os termos acima e assume todas as responsabilidades legais pela utilização indevida dos dados.

**6. Disposições Finais**

**6.1.** Esta Política de Dados deverá ser revisada regularmente, cabendo à Coordenação do Programa de Pesquisa em Gestão de Ecossistemas (COGEC/CNPq) a competência para revisá-la, com o auxílio de especialistas?e ao Comitê Gestor do PELD a competência para aprovar as versões revisadas.

**6.2.** Casos omissos ou excepcionais serão analisados pelo CNPq.

**6.3.** Esta Resolução entra em vigor na data de sua publicação.

Brasília, 13 de abril de 2016.

**HERNAN CHAIMOVICH**

Publicada no DOU de 15/04/2016, Seção 1, pág. 10