DADOS CIENTÍFICOS >
perspectivas e desafios

Guilherme Ataíde Dias
Bernardina Maria Juvenal Freire de Oliveira
Organizadores

Editora UFPB

DADOS CIENTÍFICOS:
perspectivas e desafios

Guilherme Ataíde Dias
Bernardina Maria Juvenal Freire de Oliveira
(Organizadores)

DADOS CIENTÍFICOS:
perspectivas e desafios

Editora UFPB
João Pessoa
2019

Impresso no Brasil. Printed in Brazil.

Projeto Gráfico — Editora da UFPB

Editoração Eletrônica e Design da Capa — Emmanuel Luna

Imagem — Imagem editada a partir da foto de Heather Ford no site Unsplash, maio de 2019.

Catalogação na fonte:
Biblioteca Central da Universidade Federal da Paraíba

D121 Dados científicos : perspectivas e desafios / Guilherme Ataíde Dias, Bernardina Maria Juvenal Freire de Oliveira (organizadores). - João Pessoa : Editora UFPB, 2019.
216 p. : il.

ISBN: 978-85-237-1411-6

1. Ciência da informação. 2. Dados científicos. 3. Inteligência artificial. 4. Informação e conhecimento – Ensino superior. I. Dias, Guilherme Ataíde. II. Oliveira, Bernardina Maria Juvenal Freire de. III. Título.

UFPB/BC CDU:007

EDITORA DA UFPB

Cidade Universitária, Campus I – s/n
João Pessoa – PB
CEP 58.051-970
www.editora.ufpb.br
editora@ufpb.br
Fone: (83) 3216.7147

Editora filiada à:

Associação Brasileira das Editoras Universitárias

Livro aprovado para publicação através do Edital Nº 4/2017-2018, financiado pelo Programa de Apoio a Produção Científica - Pró-Publicação de Livros da Pró-Reitoria de Pós-Graduação da Universidade Federal da Paraíba.

SUMÁRIO

PREFÁCIO

A partir da última década do século XX um processo de mudanças globais denominado de convergência digital traz à cena social grandes mudanças na forma como os saberes e as informações científicas e tecnológicas são estocadas, processadas e disseminadas. O atual estágio de desenvolvimento das tecnologias digitais tem levado à busca de estudos teóricos e debates filosóficos sobre as consequências éticas e práticas da coexistência entre humanos e máquinas inteligentes, fazendo surgir um campo de estudos instigante sobre a inteligência humana e a inteligência dos agentes artificiais. A literatura da área de Inteligência Artificial (IA) aponta que para a compreensão deste conceito é necessário relacionar os sistemas que pensam como humanos; os sistemas que agem como humanos; os sistemas que pensam logicamente e os sistemas que agem logicamente. As possibilidades do uso da IA e de sistemas inteligentes no cotidiano da sociedade traz o aumento exponencial das conexões em rede de bancos de dados digitais, fazendo emergir formas de comunicação cada vez mais abertas e multimodais.

Desde então a tendência de utilização de uma única infraestrutura de tecnologia de telecomunicações para prover serviços que em momentos anteriores requeriam equipamentos, canais de comunicação, protocolos e padrões independentes como o rádio, a televisão, as redes de computadores e telefonia traz em seu bojo profundas mudanças na forma como as pessoas e as organizações criam, armazenam, usam e disseminam a informação e as transformam em conhecimento. Emerge, deste contexto, uma economia da informação que está caminhando através de ciclos, *bits* e algoritmos

para uma nova tessitura de ideias virtuais, infindáveis e cooperativas que Pierre Lèvy vai denominar de ideografia dinâmica, isto é, um novo mundo de signos e cognição que trazem problemas filosóficos variados para se pensar a linguagem e o pensamento. Textos, imagens, vozes, atores, cenários sociais e artificiais mediados pelas tecnologias digitais apontam para um estreitamento nas relações entre humanos e máquinas. As ecologias cognitivas plurais propiciadas por novas estruturas tecnológicas de transmissão, armazenamento e compartilhamento de informações vão crescendo gradativamente, gerando esferas comunicacionais ativas e espaços de aprendizagens sociais fluidos e participativos nos fazendo repensar o mundo, as mutações culturais e relações sociais. Os suportes de escritas dinâmicas dos sistemas inteligentes trazem desafios e metamorfoses para a representação física e multidimensional do universo abstrato da informação nos fluxos das existências das memorias culturais. Esse espaço cibernético composto por seres humanos, máquinas e produtos de *software*, traz novos processos de aprendizagem de mundo uma vez que parte considerável da memória da humanidade já está sendo construída em formatos digitais. O ciberespaço elimina as distâncias, desterritorializa e se virtualiza, tornado-se um espaço mutável que comprime-se e dilata-se abrindo um novo espaço para gestão do conhecimento e para a gestão de dados em ambientes distribuídos em rede com vistas à preservação desta memória.

E é neste contexto de metamorfoses culturais que o livro "Dados científicos: perspectivas e desafios" organizado pelos pesquisadores Guilherme Ataíde Dias e Bernardina Maria Juvenal Freire de Oliveira afirma sua marca e relevância. Traz um conjunto de textos relacionados com a questão dos dados cientifico no ambiente acadêmico, abrindo ricos caminhos interpretativos sobre possíveis vias de produção e disseminação do conhecimento nas Universidades e Centros de Pesquisa.

Forjado não na simples interpretação de fatos mas na luta cotidiana e árdua da pesquisa científica, os autores da coletânea nos desafiam a compreender essa rede complexa que envolve ciclos de vida dos dados; representação de dados e suas especificidade; privacidade de dados, reuso e curadoria digital de dados, *e-Science* e outros temas afins numa linguagem científica densa, mas ao mesmo tempo elegante e simples que permite ao leitor um passeio agradável por temas que embora circunscritos ao jargão acadêmico da área da Ciência da Informação pode ser compreendido por outras campos interdisciplinares e afins.

A leitura do conjunto de textos leva à muitas abstrações sobre o futuro da humanidade e ao complexo mundo físico que habitamos com todos os seus ecossistemas memoriais. Nos leva a desconfiar do que não somos capazes de enxergar este emaranhado de nós que as redes digitais nos fazer imergir. Nos leva ao desafio de tentar perceber se em realidade estamos diante de algo semelhante ao mito da caverna de Platão ou presos no sistema Matrix.

Edna Gusmão de Góes Brennand

APRESENTAÇÃO

A produção deste livro foi resultado da nossa interlocução com diversos pesquisadores brasileiros que investigam as mais diversas questões associadas aos dados científicos, o documento inclui textos que estão na fronteira do conhecimento acerca dos respectivos temas abordados. É possível que este seja um dos livros pioneiros no que diz respeito a discussão da questão dos dados científicos no âmbito da Ciência da Informação brasileira.

Nos quatro capítulos iniciais da obra são abordadas questões de vertente mais teóricas, questões estas que são relevantes e absolutamente necessárias para a compreensão das dinâmicas associadas com os dados científicos. Os capítulos seguintes abordam questões relacionadas com a curadoria dos dados científicos, repositórios eletrônicos de dados e o Ciclo de Vida dos Dados (CVD).

Esta obra é recomendada não só para os profissionais da área da Ciência da Informação, mas para pesquisadores de qualquer área do conhecimento, visto que a produção de dados científicos está virtualmente presente nas atividades de qualquer pesquisador, a sua leitura vai possibilitar uma melhor compreensão das questões associadas com a produção, tratamento, armazenamento e disseminação dos dados científicos.

Agradecemos aos autores que contribuíram com os capítulos da obra, bem como a Pró-Reitoria de Pós-graduação da Universidade Federal da Paraíba e ao Programa de Pós-graduação em Ciência da Informação pelo apoio recebido na elaboração desta obra.

João Pessoa, 14 de janeiro de 2019

Guilherme Ataíde Dias
Bernardina Maria Juvenal Freire de Oliveira

1

CAMPO INFORMACIONAL RESULTANTE DA INTERAÇÃO DE CICLOS DE VIDA DOS DADOS

Ricardo César Gonçalves Sant'Ana

> O real bem definido elimina a ansiedade; o virtual a faz nascer e adensar. Amamos viver em redes, em cruzamentos separados por caminhos de distâncias calculáveis; hoje somos seres errantes num mundo edificável sem referências nem distâncias. Michel Serres (2003 p.143)

A construção de uma percepção da realidade é baseada no conjunto de intuições que nossa sensibilidade proporciona como insumos na instanciação de conceitos e de fenômenos, principalmente em sua dimensão material, de onde emerge sua profunda dependência da relação com os fluxos informacionais. Essa visão kantiana, antes de abrir espaço para discussões mais profundas, permite construir um cenário a partir de elementos envolvidos em nossa relação com os diversos fluxos informacionais que nos cercam.

> Nos dias atuais, as técnicas e as ciências reúnem sua imensa massa de dados como se essas informações não viessem de nossos corpos - como afirmava o antigo empirismo, quando o conhecimento emergia das sensações -, mas de um tipo de corpo

> global formado pela soma futura das espécies e dos reinos. Por meio das técnicas que aparelham nossos corpos, ampliamos nossas percepções em toda sua biocapacidade empírica possível. O velho empirismo encontra-se transformado: ele ultrapassa os cinco ou sete sentidos dos corpos individuais e amplia seu alcance às espécies vivas. (SERRES, 2003, p.129)

Entre as informações desses fluxos, as tratadas (resultantes de camadas de abstração sobrepostas ou de camadas de interpretação) apresentam um custo de obtenção mais baixo do que as baseadas em dados primários, que necessitam de todo um processo de filtragem, de tratamento e de interpretação. A mesma correspondência pode ser estabelecida em relação à dependência tecnológica desses fluxos, que se sustenta na necessidade de recursos potencializadores de nossa capacidade de lidar com grande e variado volume de conteúdos com baixa carga semântica intrínseca, o que exige um poder de processamento que atenda a requisitos como versatilidade e agilidade.

> Não concebemos mais o organismo por meio de máquinas simples, roldanas e alavancas, como faziam Descartes ou La Mettrie, nem por intermédio de motores termodinâmicos ou elétricos, como no final do Século XIX, mas sim como máquinas informáticas. A representação da vida segue nossas capacidades técnicas. (SERRES, 2003, p.64)

As tecnologias da informação não só vêm atendendo a esses requisitos como também têm se transformado em fator-chave nos ciclos de vida dos dados, indo além de suportar a coleta e armazenamento, e participar efetivamente do tratamento, do filtro, da seleção e, até mesmo, geração de novos conteúdos, são, como menciona Castells (1999), "tecnologias para agir sobre a informação, não apenas informação para agir sobre a tecnologia", retroalimentando todo o processo e destacando-se nos discos de acreção de cada acesso a dados.

Dessa perspectiva, vislumbra-se o motivo da crescente presença dos ambientes digitais em nossas relações com a informação e a crescente penetrabilidade dessas tecnologias em todo o tecido social (CASTELLS, 1999, p.108), configurando uma conjuntura totalmente distinta em que o fator-chave de sucesso é caracterizado pelo alto potencial de acesso a recursos informacionais resultantes da migração de estrutura baseada em alta disponibilidade, antes, de energia, e agora, de informação (FREEMAN, 2001).

Apesar do custo relativamente baixo, em termos de volume de conteúdo disponibilizado, esta participação das tecnologias implica em emprego de recursos adicionais aos naturalmente disponíveis e requeridos em um acesso direto, gerando custo adicional de obtenção no processo de acesso a dados que pode ser percebido em termos de dispositivos, capacidade de conexão e conhecimentos específicos. Esse custo adicional é assumido por aqueles que detêm a posse ou o controle sobre meios para construir pontes entre os usuários e os dados primários e, como tal, precisam ser justificados. Os dados e seus derivados podem, ainda, ser mercantilizados como retorno direto sobre o custo de instanciamento desses recursos.

Os argumentos para justificar tais investimentos podem, ainda, ser amparados por seu potencial de retorno, tendo em vista o valor de se ter acesso à construção dessas camadas de interpretação, que se converte em instrumento de convencimento ou de influência no senso comum. Esse capital simbólico é baseado mais em sua definição do que em sua posse, já que a vantagem reside na condução de como ele será construído, direcionamento esse que pode levar à construção e ao reforço de esquemas de poder e de fluxo de outros capitais, como o político, o econômico ou o social.

O custo de obtenção leva à construção de estruturas de manutenção de sistemas de interpretação dos dados que, por sua essência centralizadora,

requer o apoio de bens de produção que, por sua vez, dependem, entre outros fatores, da participação do capital econômico, o que leva a uma retroalimentação da assimetria informacional entre os diferentes elementos envolvidos nas relações sociais, constituindo meios de construção de forma privilegiada de poder. Serres explicita a questão do protagonismo na construção das interpretações apresentando questões como:

> Como pode o real construir-se sob a forma de signos? Em que condições alguns signos que se tornam matemáticos acessam o real e o representam? Que estatuto conceder àqueles que não lhe têm acesso a não ser por esse conjunto de características? (SERRES, 2003, p.70)

Justifica-se, assim, a necessidade de ampliar o conhecimento e controle sobre os processos envolvidos no acesso a dados de tal forma que se possa sustentar um modelo de disseminação de explicitação sobre os diversos processos baseados no uso dos dados para construção de camadas de interpretação (camadas de abstração) e os atores e intenções envolvidos nesses processos, aumentando a proporção entre o consumo consciente e insciente de informações tratadas.

A relevância social da explicitação dos fluxos informacionais é ainda maior no caso dos dados por representar processos essencialmente sociais baseados em uma forte dependência tecnológica, propiciando uma redução da ilusão naturalista que tende a minimizar o impacto da violência simbólica[1] protagonizada pelos detentores dos recursos:

---

1 "A violência simbólica é essa coerção que se institui por intermédio da adesão que o dominado não pode deixar de conceder ao dominante (portanto, à dominação), quando dispõe apenas, para pensá-lo e para pensar a si mesmo, ou melhor, para pensar sua relação com ele, de instrumentos de conhecimento partilhados entre si e que fazem surgir essa relação como natural, pelo fato de serem, na verdade, a forma incorporada da estrutura da relação de dominação; ou então, em outros termos, quando os esquemas por ele empregados no intuito de se perceber e de se apreciar,

> O poder de apropriação simbólica do mundo, garantido pela visão perspectiva ao situar o diverso sensível na unidade ordenada de uma síntese, se apóia, como que sobre um pedestal invisível, no privilégio social, o qual constitui a condição de emergência dos universos escolásticos, bem como da aquisição e do exercício das disposições correlatas. (BOURDIEU, 2001, p.34).

Assim, parte-se do estudo das fases dos ciclos de vida dos dados e dos resultados da interação desses ciclos em indivíduos ou instituições, buscando elementos que possam compor uma estrutura de referência que possa servir de apoio para estudos e até mesmo constituição de novos ciclos de vida dos dados voltados para interpretação deste cenário, ou seja, ciclos de segunda ordem que possam amparar um referencial teórico sobre essas interações.

Acesso a dados e o Ciclo de Vida dos Dados

Ao se deparar com processo tão complexo como o acesso a dados, torna-se premente a necessidade de buscar elementos que permitam construir uma estrutura de referências que propicie não só estabelecer pontos de vista sobre o todo como também identificar dimensões de posicionamento como as relacionadas ao tempo e aos objetivos e características do contexto em que o processo se insere.

Assim, ao pensar na dimensão tempo, é possível estabelecer um processo de classificação e, principalmente, de ordenação de atividades, esforços e competências envolvidos no processo, o que propicia, no caso do acesso a dados, considerar a percepção de fases que agrupam tais fatores.

---

ou para perceber e apreciar os dominantes (elevado/baixo, masculino/feminino, branco/negro etc.), constituem o produto da incorporação das classificações assim naturalizadas, cujo produto é seu ser social." (BOURDIEU, 2001, p.206)

Coleta

Para analisar essas fases, é possível partir de um primeiro momento em que a partir de uma necessidade surge a possibilidade de se estabelecer o escopo da informação necessária, sobre a qual se baseia, então, atividades como o planejamento e execução da obtenção do conteúdo esperado. Emerge, assim, uma delimitação que estabelece o que se pode entender como a fase da coleta dos dados. Nessa fase são alocados esforços e competências específicas, que incluem o processo de definição dos requisitos a atender, a delimitação do conteúdo a ser obtido, os procedimentos para localização, seleção, filtro e análise da qualidade do que for coletado e, ainda, outras tarefas envolvidas nesta fase.

Uma vez obtido o conteúdo desejado, seu uso pode ser imediato ou, descrevendo de outra forma, sua existência mantida somente pelo tempo da necessidade imediata e em meios voláteis, sendo que, ao fim da utilização, o mesmo pode ser imediatamente descartado.

Armazenamento

Após da obtenção e eventual uso do conteúdo, pode ser necessário que este seja mantido de tal forma que possa vir a ser acessado novamente, sem a necessidade de se recorrer a sua fonte original. Essa necessidade pode ser ainda mais clara, quando se tratar de processos de obtenção que resultam de contextos únicos e impossíveis de se replicar, seja por suas variáveis espaciais seja pela própria condição de seu contexto temporal.

Passa-se, então, a um segundo momento, com novas atividades que envolvem características, objetivos e competências distintas das anteriores, e que exigem proximidade maior com os

aspectos tecnológicos. Essa fase passa a ter como foco a manutenção dos dados coletados de tal forma que possam ser acessados em um momento futuro e, portanto, podemos entender esta fase como a fase de armazenamento dos dados.

Recuperação

Garantir que os dados estão em um determinado suporte e que poderão ser acessados no futuro não inclui preocupações relacionadas ao 'como' este acesso poderá ocorrer, e esta é a característica de um outro momento do acesso a dados que é definida, novamente, por novas atividades que envolvem objetivos e competências distintas das anteriores. Essa fase é baseada nas questões relacionadas a como tornar estes dados acessíveis, incluindo o que poderá ser acessado, quem poderá fazê-lo e como este acesso será disponibilizado, preocupações, portanto, sobre a recuperação daquele dado armazenado o que permite que possamos identificá-la como fase de recuperação.

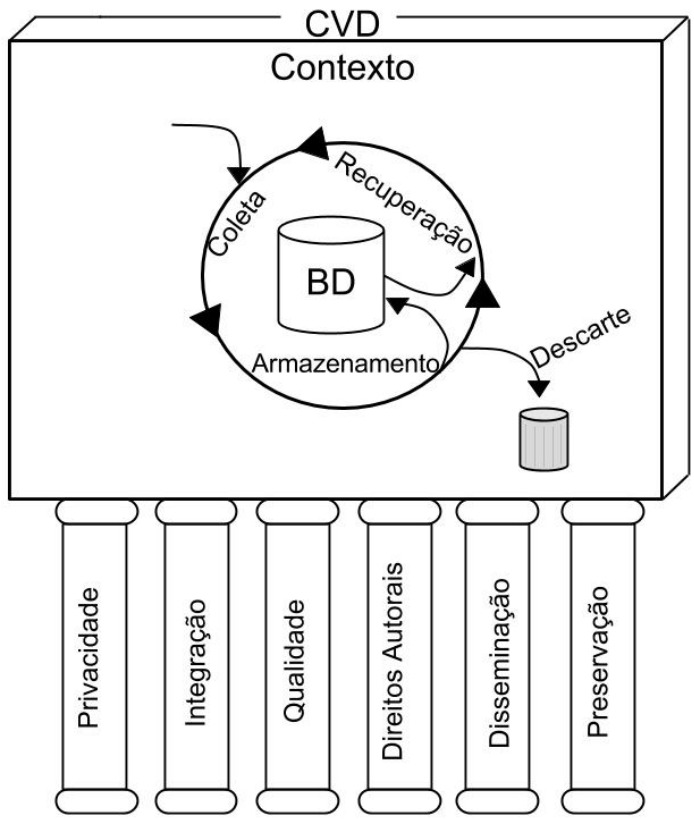
Descarte

Mesmo sendo cada vez mais baixo o custo de armazenamento de dados, persiste a necessidade de que os dados coletados em determinado momento sejam eliminados. Esse processo pode ter muitas motivações, desde aquelas relacionadas ao espaço de armazenamento, até as relacionadas ao direito de que envolvidos/referenciados pelos dados decidam que os mesmos não estejam mais disponíveis. Nessa fase, como nas anteriores, são requisitadas atividades e competências específicas e podemos, então, entender este momento como fase de descarte.

Característica cíclica

Essas fases não podem ocorrer de forma independente e são profundamente entrelaçadas, a ponto de exigir, por exemplo, que, durante o planejamento de cada uma delas, deverão ser levados em conta elementos contextuais das demais. Também podem ocorrer de forma concomitante, sendo que depende de cada contexto se cada uma delas se configura como projeto ou processo.

Figura 1: Ciclo de Vida dos Dados



Fonte: Sant'Ana (2016).

As fases apresentam, ainda, uma forte vinculação entre elas, assim, o dado coletado pode então ser armazenado e disponibilizado para recuperação, sendo que, qualquer uma destas fases pode seguir para a fase de descarte. Essa característica permite que o processo, como um todo, possa ser percebido como um ciclo (figura 1) que representa todos os momentos envolvidos no acesso a dados, configurando o que se pode denominar de ciclo de vida dos dados. No entanto, existem fatores que permeiam todas as fases de forma transversal e que, por estarem presentes em todas elas, em menor ou maior grau, precisam ser considerados em todas elas, principalmente no planejamento, funcionando como elemento de ligação entre as fases. Esses fatores são: privacidade, integração, qualidade, direitos autorais, disseminação e preservação.

Nessa mesma linha de raciocínio, e considerando que para cada contexto distinto de conjuntos de dados que necessitamos, acessamos e utilizamos compõe-se de um ciclo de vida distinto, temos a possibilidade de uma estrutura maior, resultante desses ciclos e que ainda deve considerar nossa participação, consciente ou não, em processos de coleta, o que corrobora na sustentação da percepção de um espaço próprio da resultante de todos esses fluxos informacionais.

Campo Informacional

O surgimento do espaço virtual retirou do humano a capacidade de estender, como fazia em situações anteriores, a projeção de suas percepções do espaço que o cerca, tirando sua condição de *anthropos*. Desse novo espaço, elaborado como um constructo de segunda ordem do próprio espaço humano, emergem novas dimensões, muitas das vezes, distintas até mesmo das dimensões canônicas do espaço físico.

A percepção destas novas dimensões pode ser observada em situações como a de mensurar a distância entre dois nós em uma rede social. Nesse ambiente, distintamente do físico, a distância geográfica desses nós tem pouco impacto direto e nem mesmo as possibilidades apresentadas por técnicas de mensuração de dois pontos em redes podem ser aplicadas eficazmente, já que, ainda que dois nós estejam a uma distância mínima de n links, uma simples busca em mecanismos de recuperação faz com que essa distância se reduza a uma unidade, tornando aqueles que até então estavam "distantes" em vizinhos.

Se considerarmos que a composição deste campo resultante será definida por dimensões que definem o vetor resultante em função de elementos que realmente interferem neste vetor, é possível inferir que o impacto das coordenadas físicas do elemento alvo para composição do campo interfere de forma muito sutil, ou até mesmo que as coordenadas até então utilizadas podem ser substituídas por uma dimensão que sintetize os efeitos da posição geográfica do elemento alvo, tal como o nível de conectividade. Assim, as dimensões formadas pelas coordenadas do elemento alvo poderiam ser substituídas pela dimensão "conectividade" corroborando com a percepção de que a alta disponibilidade de conectividade ofusca a questão geográfica.

Esse mesmo raciocínio pode ser estendido a outras dimensões que podem ser candidatas à composição do campo informacional resultante dos ciclos de vida dos dados a qual um elemento alvo está relacionado. Portanto, além da conectividade é possível considerar candidatos a dimensões na composição do campo informacional fatores como: competência do elemento alvo; características culturais; forma de controle sobre recursos necessários como por exemplo de infra, entre outros.

No caso da dimensão competência do elemento alvo, pode-se considerar a capacidade que o mesmo tem de acessar, ou mesmo

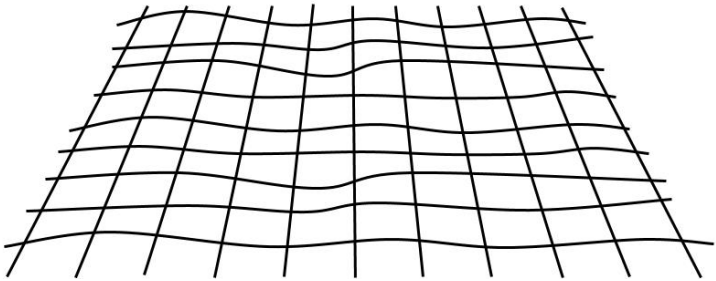
interagir, com os mecanismos de captação e de disponibilização de dados, fazendo com que o elemento possa ter diferenciais, principalmente em termos de potencial de utilização dos dados que estão ao seu alcance. Esse uso diferenciado pode levar, ainda, a uma interação maior com os dispositivos tecnológicos, o que pode facilitar os processos de coletas de dados sobre ele ou sobre suas ações. Resulta que a distância entre dois elementos alvo não pode mais ser definida somente por suas localizações geográficas e nem mesmo pela distância definida pela análise da arquitetura das redes que estão envolvidos.

As reflexões que emergem ao se analisar as relações da vida em sociedade levam a distinção de duas dimensões de análise para que elas ocorram: espaços físicos e simbólicos. Enquanto limitados por parcas alternativas comunicacionais, os espaços físicos detinham grande importância na configuração das relações. No entanto, a base, principalmente nas relações mais complexas em que ocorre maior volume de informações transacionadas, sempre foram os espaços simbólicos.

No espaço simbólico, construído a partir das relações sociais, "pessoas ocupam posições diferentes, e esses desníveis levam à noção de campo" (MARTINO, 2009, p.147). Esta definição de campo se baseia em um "espaço estruturado onde agentes em disputa buscam a hegemonia simbólica das práticas, ações e representações" (MARTINHO, 2009, p.147)

Para facilitar sua interpretação e estudo, propõe-se uma vinculação do conceito de campo ao conceito definido pela teoria de campo pela Física, agregando, assim, elementos estruturantes de segunda ordem a estes campos que permitam propostas de formas de se construir perspectivas sincrônicas do campo em estudo.

Figura 2: Espaço em que se propicia a percepção da relação de distintos elementos em relação a uma variável subjacente.



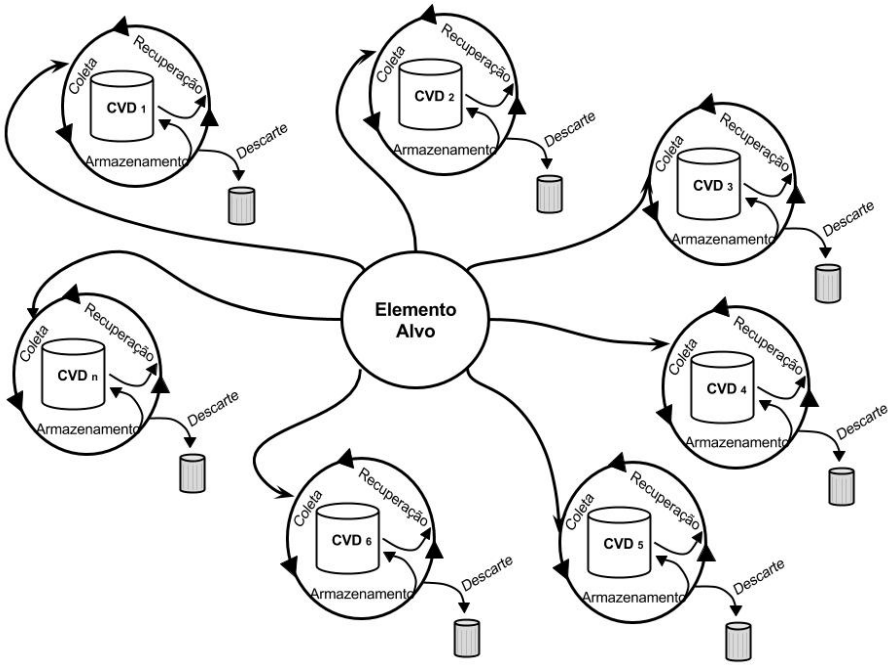
Fonte: Autor

Acrescenta-se, ainda, a possibilidade de, por meio do registro de dados sobre a composição de um campo em um determinado momento e uma análise deste mesmo campo em um momento distinto, compor um olhar histórico, diacrônico, ou seja, uma análise horizontal a partir de seus dados.

Ao considerar o conjunto de CVDs aos quais um elemento alvo está, de alguma forma relacionado, seja por meio de exposição à fase de coleta de CVDs de governos, empresas ou indivíduos, seja por meio da possibilidade de termos acesso a dados que possam alimentar a fase de coleta dos CVDs deste elemento, pode-se inferir uma resultante que pode ser entendida como o nível de interação de dados deste elemento alvo com o meio externo.

Partindo-se da composição destes Ciclos de Vida dos Dados - CVDs, pode-se então conceber que um determinado elemento alvo é submetido a um conjunto de CVDs em suas fases de coleta, abrindo a possibilidade de que este elemento alvo forneça dados a estes CVDs (figura 3). O elemento alvo, que também poderia ser denominado como agente, transcende características pessoais ou institucionais, englobando as várias categorias daquele que estiver sendo analisado enquanto alvo de coleta ou interagindo com a fase de recuperação de CVDs ao seu alcance.

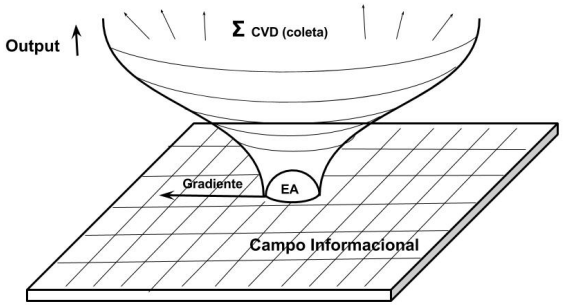
Figura 3: CVDs e seus processos de coleta de dados



Fonte: Autor

Esta interação é denominada, neste estudo, como parte do campo informacional a qual o elemento alvo está submetido, e que responde pelos dados coletados sobre, com ou sem seu consentimento ou mesmo ciência. Assim, tem-se uma dimensão *output* do campo que identifica o volume e variedade de dados e que pode resultar, inclusive, na percepção de densidade desse campo. A composição de CVDs em um amálgama resultante representa a densidade, conforme ilustrado na figura 4, de onde é possível esperar, inclusive, a identificação de elementos próprios de campo, como por exemplo, seu gradiente.

Figura 4: Fases de coleta de CVDs obtendo dados de/sobre o elemento alvo

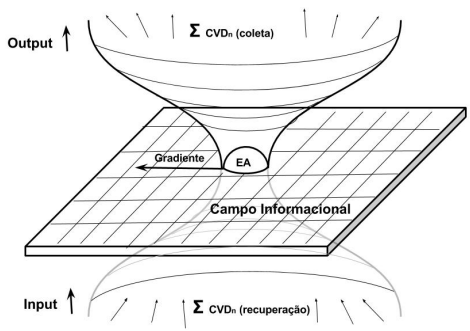


Fonte: Autor

Da mesma forma podemos considerar o conjunto dos processos de coleta dos CVDs que os elementos alvo detêm para compor uma segunda dimensão deste campo, conforme ilustra a figura 5, e que pode ser entendida como a dimensão *input*.

Cabe destaque a distinção que devemos manter em relação a possibilidade de coleta direta a partir de sensores ou dispositivos que obtenham conteúdos do mundo real por interação ou capacidade de sensibilização e, por outro lado, temos, ainda, a coleta feita de outros CVDs, o que configura reuso de dados.

Figura 5: Composição das fases de coleta dos CVDs que o elemento alvo detêm



Fonte: Autor

Para que se possa ter uma mensuração, ainda que inicial, pode-se considerar a utilização do volume de dados transacionados em todos os CVDs identificados durante um determinado período, gerando assim um resultante que seria calculado em bits/tempo. Essa medida poderia ser obtida em termos de quantidade de dados que fluíram, ou que, em potencial, poderiam ter fluido nos CVDs identificados, dependendo do objetivo da mensuração ser voltada para dados efetivamente transacionados ou potencialmente transacionáveis.

Ao analisar essas dimensões percebe-se a concepção de novos espaços que não estão baseados nas antigas bases físicas, impactando na zona de conforto e retirando a possibilidade de correlação direta entre os espaços geográficos e os novos ambientes virtuais, que passam a requisitar uma nova percepção de construção de espaços multidimensionais e com limitações e características ainda por serem mapeadas.

Desta forma temos, entre outros resultados, a possibilidade de percepção direta da assimetria deste elemento alvo no campo informacional, por meio do nível de desequilíbrio entre as duas dimensões *input/output* configurando um desequilíbrio entre ele e os demais elementos da sociedade que faz parte. Esse desequilíbrio pode ser positivo como no caso dos que possuem maior densidade na dimensão *input* do que na dimensão *output,* o que lhe proporciona vantagem competitiva. O mesmo ocorre no sentido contrário para os casos em que a dimensão *output* apresenta maior densidade, proporcionando desvantagem nas relações.

Campo Informacional e os dados científicos

Na prática da pesquisa, em especial na Ciência da Informação, é possível identificar uma oportunidade para reforçar a questão da identidade da área uma vez que, responsável pela camada de informação, deve atender

agora às especificidades que o acesso a dados acrescenta e pode responder a questões relacionadas às dimensões *input* e *output,* por exemplo, e, ainda, refletir sobre possíveis requisitos, barreiras, atores, relações e assimetrias resultantes do processo.

O cenário dos campos informacionais também pode contribuir para a percepção da relação entre os diversos atores envolvidos no universo das pesquisas científicas. Cada vez mais dependentes do uso intensivo de dados, as assimetrias no acesso a dados podem explicitar as desigualdades encontradas nos diferentes ambientes em função do controle sobre os processos em cada uma das fases dos CVDs, em especial armazenamento e recuperação. Mesmo a coleta sendo realizada nos países periféricos, se o controle sobre o armazenamento não for muito claramente compartilhado, corre-se o risco de limitações derivadas de interesses centrais transformarem-se em barreiras, não necessariamente explícitas, quando da construção das possibilidades de acesso na fase de recuperação.

Definições de padrões e a própria responsabilidade pela manutenção dos recursos de armazenamento não podem estar submetidos a governos ou mesmo entidades - modelo *top-down* - que, por melhores e maiores que sejam os compromissos de gestão compartilhada, ficarão submetidas a um potencial controle futuro inerente a interesses de seus detentores. Uma forte justificativa para este tipo de solução pode ser amparada na disponibilidade de recursos financeiros (como nos casos de governos) e de proximidade de competências específicas (como nos casos de instituições de pesquisa), o que não descaracteriza o potencial de controle futuro.

As soluções para estas questões devem sempre ser acompanhadas de abertura para modelos que permitam a interoperabilidade de características que, mesmo específicas para determinadas necessidades, possam ser integradas e replicadas de forma descentralizada e balizada

por processos *bottom-up*. A própria adoção disseminada pela comunidade científica pode ser facilitada em modelos descentralizados.

Considerações

A composição do espaço vetorial com o qual o campo vai ser percebido pode ultrapassar a característica tridimensional, o que leva a dificuldade de correlação deste espaço com o espaço físico sobre o qual construímos nossa percepção de realidade. Essa característica faz com que, desse espaço virtual, aflore a necessidade de novos olhares. Esse novo espaço pode ser composto por um conjunto de vetores base que tem sua constituição definida por fatores que interferem no resultante do campo em um elemento alvo.

Entre as questões que emergem estão: como mensurar o impacto de uma determinada dimensão no resultante deste espaço vetorial sobre o elemento alvo? Existe interdependência entre os diversos fatores/dimensões? É possível estabelecer um valor de gradiente dentro deste campo? Entende-se que todas essas questões podem ser estudadas, em um primeiro momento, por meio de mapeamento e monitoramento de elementos alvo durante determinado período e com controle sobre as dimensões envolvidas no estudo, utilizando-se o volume de dados transacionados conforme proposto neste texto.

Sistemas políticos e interesses de grandes corporações, mesmo que indiretamente, geram dificuldades para o uso efetivo de dados para definição clara de relações, tais como entre o uso de combustíveis fósseis e mudanças climáticas, controle de armas e índices de violência, promessas eleitorais e resultados de gestões públicas, consumo de tabaco e câncer, entre tantas outras. Em casos como o da relação tabaco/doenças, grandes conquistas foram obtidas no estabelecimento da correlação, não sem

um longo e difícil embate entre os diversos envolvidos no processo de utilização dos dados e seus detentores.

É preciso estar atento para a oportunidade que se abre, já que o impacto das tecnologias digitais no acesso a dados não só acelera e retroalimenta a concentração de poder e o desequilíbrio entre os diversos atores que o compõem como, por outro lado, propiciam a abertura de canais de acesso a dados sobre o processo em si, criando condições para que pesquisas possam buscar sustentação que vão além do escopo de mero uso para as reflexões e inferências. Emerge deste cenário a necessidade de se considerar estudos futuros que façam aportes de outras áreas, como a identificação de estruturas nos campos informacionais, como por exemplo, eventuais discos de acreção informacionais, o que permitirá o balizamento de pesquisas mais abrangentes sobre os elementos envolvidos nos fluxos informacionais.

Vivemos imersos em um universo complexo e, quanto mais estruturas desenvolvemos para controlá-lo e alterá-lo, mais necessitamos de elementos que nos permitam identificar ordem e relação entre seus elementos.

Referências

BOURDIEU, P. *Meditações pascalinas*. Rio de Janeiro: Bertrand Brasil, 2001. 324 p.

CASTELLS, M. *A sociedade em rede*: a era da informação: economia, sociedade e cultura. São Paulo: Paz e Terra, 1999. V. 1.

FREEMAN, C. A hard landing for the 'New Economy'? Information technology and the United States national system of innovation. *Structural Change and Economic* **Dynamics**. v.12, n.2, Julho,2001. p.115-139.

MARTINO, L. M. S. *Teoria da comunicação*: idéias, conceitos e métodos. Petrópolis, RJ: Vozes, 2009. 286 p.

SANT'ANA, R.C.G. Ciclo de vida dos dados: uma perspectiva a partir da Ciência da Informação. *Informação & Informação*. v.21, n.2, 2016. p. 116-142.

SERRES, M. *Hominescências:* o começo de outra humanidade? Rio de Janeiro: Bertrand Brasil, 2003.

2

A CIÊNCIA INVISÍVEL: os dados da cauda longa da pesquisa científica

Luís Fernando Sayão
Luana Farias Sales

Introdução

Há uma parcela considerável do trabalho científico que não está visível nem para a sociedade, em termos de benefícios e de qualidade de vida, nem para os pares no contexto da dinâmica de uma comunidade científica. "Estudos recentes indicam que mais de 50% das descobertas científicas não aparecem na literatura publicada, ao invés disso, residem nas gavetas e nos computadores pessoais dos pesquisadores" (FERGUSON *et al*, 2014, p. 1443).

Esse fenômeno tem os contornos mais nítidos no segmento da ciência conhecido como "cauda longa da ciência", em que um grande número de pequenas equipes de pesquisadores e laboratórios independentes gera, em seu dia a dia de pesquisas, uma ampla variedade de coleções de dados. Apesar do dimensionamento individual, essas pequenas coleções de dados estão sendo reconhecidas como ativos informacionais de alto valor, que, coletivamente, têm o potencial

de ser mais relevantes do que a soma de suas partes (WYBORN; LEHNERT, 2016). Os dados da cauda longa representam a maior parcela de dados produzida pela ciência e constituem um território de constante criatividade e inovação que precisam ser revelados, integrados e compartilhados.

São muitas as razões para que a opacidade se instale nesse segmento da ciência e crie obstáculos para que os dados sejam razoavelmente integrados a outros ativos, compartilhados e reusados. Começando com a falta de infraestruturas tecnológicas e gerenciais e de políticas institucionais que assegurem a estabilidade, a persistência e a interoperabilidade dos dados; o controle de qualidade e a padronização, porque a natureza heterogênea e fragmentada dessas coleções exige estratégias diversificadas para sua gestão; políticas voltadas para a publicação de dados, a ausência de esquemas de reconhecimento da autoria e políticas de recompensa pela organização e pela disseminação dos dados e a falta de interesse dos pesquisadores em divulgar dados além dos limites profissionais mais próximos, dados sobre hipóteses não confirmadas e resultados negativos e dados considerados auxiliares de estudos publicados em artigos.

A descontinuidade provocada por essa invisibilidade de um segmento importante da pesquisa científica fragiliza a ciência, como um empreendimento social e humanístico, e tem muitos desdobramentos. Os mais contundentes são: deixar registros incompletos e tendenciosos dos processos de geração de conhecimento, criando uma lacuna nas memórias acadêmicas das instituições; duplicar esforços que alonguem desnecessariamente o ciclo de comunicação científica e, sobretudo, impossibilitar que se efetivem os princípios básicos da reprodutibilidade dos experimentos científicos e da autocorreção da ciência (FERGUSON *et al*, 2014; THE ROYAL SOCIETY, 2012).

Nesse contexto, observa-se, com clareza, que há uma diferença acentuada nos processos de gestão de dados. Enquanto um grande cuidado é devotado frequentemente à formação de coleções, à preservação e ao reuso de dados provenientes dos grandes projetos, comparativamente, pouca atenção é dedicada aos dados que são gerados pela maioria dos cientistas que, cotidianamente, desenvolvem projetos em menor escala. Novas estruturas sociais e desenvolvimentos técnicos podem aumentar significantemente a disponibilidade e o valor dos dados dos "pesquisadores individuais" e de seus projetos de pesquisa. O desafio que se coloca para a política científica e para as instituições de pesquisa é desenvolver infraestruturas e práticas como repositórios digitais disciplinares, que tornem esses dados úteis para a sociedade (HEIDORN, 2008).

O presente ensaio tem como objetivo analisar brevemente as causas e os desdobramentos da opacidade desses ativos informacionais e as necessidades infraestruturais para que eles se revelem e possam ser compartilhados e reusados em domínios disciplinares distintos dos quais eles foram originalmente gerados ou coletados. Nessa direção, na primeira seção, serão apresentadas as diferenças estruturais entre *Big Science* e *Little Science*; em seguida, serão analisadas as causas e as consequências da opacidade dos dados da cauda longa e, por fim, por fim serão estabelecidos alguns requisitos tecnológicos, gerenciais e organizacionais para a gestão e a curadoria desses dados.

Big Science e *Little Science*: formas diferentes de estruturar dados

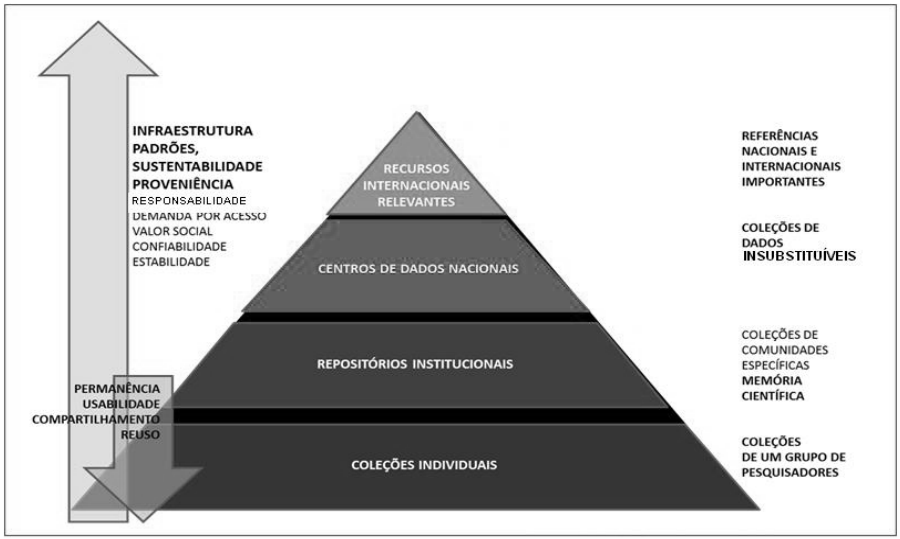
O quarto paradigma científico, ou eScience, é descrito na literatura como a reconfiguração dos paradigmas anteriores – ciência experimental, ciência teórica e ciência baseada em simulação - em torno de um mundo

rico em dados. São muitas as expectativas em torno dessa nova forma de fazer ciência dado a possibilidade que oferece de *insights* revolucionários que vão dos fenômenos relacionados às mudanças climáticas e às descobertas de novas drogas até as metodologias que possibilitem examinar novos ângulos da história e da cultura (BORGMAN, 2012).

Esse novo modelo de ciência tem os contornos mais nítidos nas "big sciences", como Astronomia e Física, e nos campos de estudos híbridos que vão surgindo, como Astroinformática, Bioinformática e Humanidades digitais. Entretanto, os pequenos laboratórios que formam as estruturas de pesquisa das universidades também se tornam mais e mais intensivos na geração e no uso de dados, e novos métodos e instrumentação possibilitam que pesquisadores individuais e pequenas equipes coletem volumes sem precedentes de observações (BORGMAN, 2012). Porém, as diferenças de gestão e de compartilhamento dos dados produzidos por esses dois universos de pesquisa científica são muito diferentes.

Nos ambientes científicos, é consenso que o compartilhamento de dados pode ser complexo e custoso e precisa ser reconfigurado por estimativas realísticas de demandas e de usabilidade que justifiquem a permanência e a curadoria das coleções de dados de pesquisa. Nessa direção, o relatório da The Royal Society (2012) propõe um modelo com quatro camadas, o qual explicita os padrões atuais de gestão de dados de pesquisa e sua relação com as demandas sobre eles. As condições de demanda conduzem a diferentes níveis de profundidade de curadoria. "As camadas de atividades refletem a escala, o custo e o alcance internacional dos dados gerenciados e, em algum grau, a percepção de sua importância" (p.60). Cada camada exige diferentes aportes financeiros e infraestruturais. A figura 1 é uma adaptação dos autores da figura originalmente apresentada no Relatório e enfatiza as características que interessam aos propósitos deste estudo.

Figura 1: Pirâmide de gestão de dados



Fonte: Autores, baseado em The Royal Society (2012)

Na figura 1, a amplitude do valor dos dados aumenta no sentido do vértice superior da pirâmide. Sua escala de valor se inicia em um nível individual e passa para o comunitário até alcançar um valor para toda a sociedade. Isso implica mais responsabilidades e demandas por acesso, além de atenção aos padrões, sustentabilidade, estabilidade e proveniência (THE ROYAL SOCIETY, 2012).

Na primeira camada, estão as coleções massivas de dados geradas pelos grandes programas internacionais de pesquisa, como os que estão subjacentes ao Grande Colisor de Hádrons[1] e ao Wordwilde Protein Data Bank[2], que se caracterizam como atividades que se desenrolam no

1 <https://home.cern/topics/large-hadron-collider>
2 <http://www.wwpdb.org/>

contexto da *big science*; na segunda camada, estão incluídos os centros de dados e recursos gerenciados pelos órgãos nacionais; na terceira, estão representadas as coleções geradas por programas de pesquisa conduzidos por instituições individuais, como universidades e institutos de pesquisa, cuja gestão – pela própria diversidade dos dados - é ampla e variada; na quarta, os pesquisadores individuais ou grupos de pesquisadores que coletam e armazenam seus próprios dados.

Nessa última camada, considerados coletivamente, está o maior volume de dados produzidos pela ciência – a maior parte composta de *datasets* com menos de 1 GB, cuja metade está armazenada em seus laboratórios de origem (HORSTMANN, 2015). O fluxo de compartilhamento de dados só se realiza tipicamente entre os colaboradores mais próximos ou por meio dos *websites* dos projetos ou das instituições. A gestão é apoiada por ferramentas *off-the-shelves*, como planilhas eletrônicas, porém uma grande parcela desses dados permanece sem nenhuma ação concreta de gestão e curadoria e desaparece – tragada pelo tempo e pela obsolescência tecnológica - nos dispositivos portáteis de armazenamento e nos discos dos computadores pessoais dos pesquisadores. Aí se identifica, também, a grande parcela de dados obscuros e sem qualquer gestão que os torne visíveis e reutilizáveis.

As bases de dados da parte superior gerenciam os dados gerados tipicamente pelas atividades da chamada *big science*, cujas características marcantes são o uso de grandes dispositivos e instrumentos de pesquisa, o alto custo envolvido e a longa duração dos seus programas, que envolvem grandes equipes de pesquisadores distribuídos em escala planetária, com uma grande expertise do domínio.

As atividades da *big science* têm uma correlação estrita com o fenômeno do *big data* – o *big data* científico - que compreende as disciplinas que geram volumes massivos de dados observacionais e computacionais e

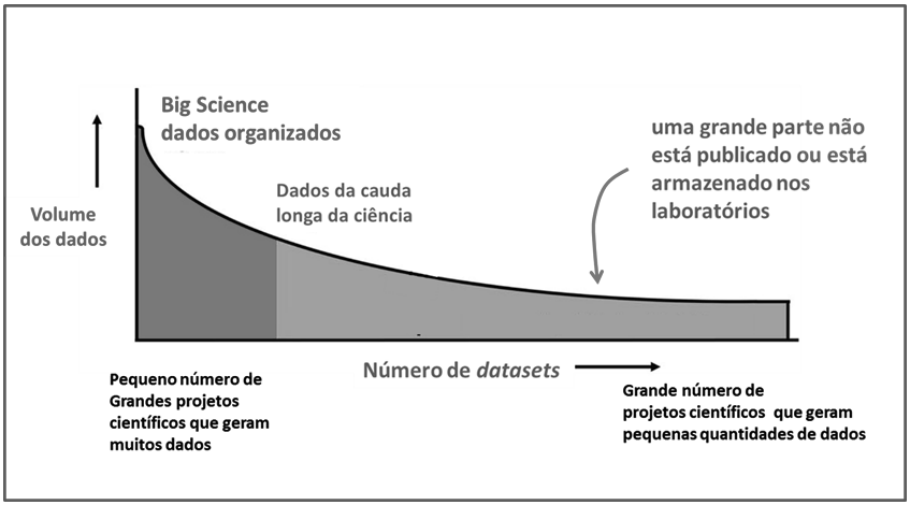
usam em larga escala instrumentos compartilhados, como redes globais de sensores, satélites e computação de alto desempenho. Os dados gerados ou coletados são tipicamente padronizados e relativamente bem gerenciados e curados por estruturas sofisticadas de dados (WYBORN; LEHNERT, 2016) comunitárias, nacionais e internacionais, como centros de dados.

Na base da pirâmide, em escala bem menor, identificam-se os dados primariamente adquiridos por pesquisadores individuais ou por pequenos grupos de pesquisa que trabalham cotidianamente em projetos que estão circunscritos aos laboratórios das universidades e dos institutos de pesquisa. Ao contrário dos aparatos da *big science*, os instrumentos são dimensionalmente menores e, muitas vezes, construídos pelos próprios pesquisadores ou fornecidos por uma grande diversidade de fabricantes; têm baixo orçamento, curta duração e se desenrolam localmente sem pouca conexão com os grandes projetos da *big science*.

As análises da distribuição da geração de dados pelos segmentos científicos da grande e da pequena ciência são comumente baseadas no conceito estatístico de "cauda longa" (*long tail*), muito utilizado nas áreas de marketing, especialmente, de produtos digitais. A cauda longa é uma forma específica de distribuição estatística em que uma pequena parte da população tem muitas ocorrências (a cabeça), enquanto grande parte tem pouca (a cauda).

No mundo da pesquisa, a cauda longa é caracterizada por pequenas quantidades de dados heterogêneos com padrões e fluxos específicos e não estão regulados mais amplamente. Isso implica, primordialmente, a necessidade de gestão, de curadoria e de controles personalizados que devem ser realizados, idealmente, em ambientes orientados por disciplina (e-IRG Task Force, 2016, HEIDORN, 2008). A figura 2 exibe a distribuição estatística aplicada à geração de dados de pesquisa.

Figura 2: Dados da cauda longa da Ciência



Fonte: Adaptado de Ferguson (2014)

Com o advento do *big data*, o foco de atenção das agências de fomento e dos formuladores de políticas científicas se voltou, prioritariamente, para os segmentos da pesquisa que estão fundamentados na geração e no uso intensivo de dados. Como desdobramento dessa supervalorização, uma grande quantidade de pequenas coleções de dados sai do campo de visão desses *stakeholders* (e-IRG Task Force, 2016), em contraposição à crescente importância desses recursos informacionais como insumo essencial para diversos segmentos da pesquisa contemporânea.

De forma geral, a demanda por dados de pesquisa - na qualidade do produto de pesquisa de primeira grandeza, e não, apenas, como um subproduto dos fluxos das atividades científicas - pela ciência contemporânea e, mais especificamente, pelas metodologias de geração de conhecimento da eScience - define a concepção de um espaço de dados, compartilhados e integrados, que necessita da diversidade informacional que caracteriza a pesquisa da cauda longa, principalmente no que tange

às pesquisas interdisciplinares. A *diversidade de dados*, como argumentam Boyd e Crawford (2012), pode ser considerada como o princípio organizacional mais geral da pesquisa científica. Num extremo do espectro dessa diversidade, está o *big data*, endereçando o aumento exponencial de geração e disponibilidade de dados, com múltiplas oportunidades e desafios para a pesquisa científica; no outro extremo, a cauda longa, cujas descrições sublinham a variedade em termos de estrutura, domínio disciplinar, complexidade, contexto, formato, volume, localização e utilização na pesquisa (HEIDORN, 2008; BORGMAN, 2015, e-IRG Task Force, 2016).

A perspectiva sistêmica do espaço de dados é um foco rico de estudos e de atividades acadêmicas e gerenciais para pesquisadores, bibliotecários e cientistas de dados e se desdobra em iniciativas em âmbito nacional e internacional. Um exemplo ilustrativo vindo da comunidade europeia é o Grupo de Interesse em Dados de Pesquisa da Cauda Longa (*Long Tail of Research Data Interest Group*)[3] do RDA – Research Data Alliance[4]. No contexto norte-americano, o *Sustainable Environment/ Actionable Data* (SEAD)[5], um projeto do NSF Data Net, objetiva desenvolver ciberinfraestrutura e serviços sustentáveis para o acesso e a preservação de dados em comunidades científicas típicas da cauda longa.

A ciência da cauda longa da pesquisa e suas qualidades

A ciência da cauda longa não é um sinônimo de questões científicas menos relevantes ou de uma ciência menor, tampouco os dados que ela gera podem ser desconsiderados em um contexto cuja diversidade

3 <https://www.rd-alliance.org/long-tail-research-data-interest-group.html >
4 <https://www.rd-alliance.org/ >
5 <http://sead-data.net/>

e integração de dados são a chave para respostas para novas indagações da ciência. Ao contrário, é um território com altos coeficientes de autonomia, que favorece a inovação e a geração de conhecimentos multi e interdisciplinares.

Verdadeiramente, os resultados da imensa quantidade de projetos identificados como pertencentes a este segmento da ciência, se apropriadamente gerenciados, podem contribuir para um genuíno *big data* científico, visto que é ela que, coletivamente, produz a maior parte dos dados de pesquisa da ciência e tem o potencial de contribuir para grandes realizações e para um grande acúmulo de novos conhecimentos (HEIDORN, 2008). Isso fica mais claro, quando se considera que uma parte expressiva dos produtos de pesquisa publicados em artigos científicos, em forma de dados terciários, são de dados provenientes da cauda longa, que ainda se manifestam em outros produtos de pesquisa, como patentes e modelos, multimídias e muito mais. "De certa forma, eles [os dados da cauda longa] são peças de um lego que, se forem colocados juntos corretamente, têm o potencial de gerar novos conhecimentos", destaca (WYBORN; LEHNERT, 2016 p.1).

Alguns autores como Heidorn (2008) consideram a cauda longa da pesquisa como um espaço de transformações propício à independência criativa. "Parece mais provável que a ciência transformadora venha mais da cauda do que da cabeça" (p.2). Isso acontece, principalmente, porque as questões a serem respondidas pelos grandes projetos geradores de dados são relativamente bem compreendidas. Já a cauda longa é um terreno fértil para novas ideias e para uma ciência inédita, nunca antes tentada. Assim, quando submetidas a processos adequados de gestão e de agregação de valor proporcionado pela curadoria, as coleções de dados podem ser comparadas com os dados da cauda e da cabeça e integradas a eles - para revelar padrões temporais e espaciais em larga escala, que podem conduzir a novos *insights*

e descobertas científicas (WYBORN; LEHNERT, 2016). É preciso considerar, ainda, que o compartilhamento dos dados da cauda longa é considerado essencial para aumentar a transparência e a autocorreção no mundo científico e para verificar, revisar e validar a pesquisa e atenuar conhecidos vieses das publicações acadêmicas (FERGUSON, *et al.*, 2014).

Não obstante ao relativo grau de negligência e de opacidade da cauda longa da pesquisa, quando comparada com a grande ciência, várias iniciativas importantes, que incluem pesquisas, instalações, ferramentas, grupos de interesse e ações governamentais, concentram sua atenção na ciência produzida pelos pequenos laboratórios. Essa valorização dos produtos de pesquisa da cauda longa se realiza com maior nitidez no escopo do movimento mundial sobre os pressupostos da ciência aberta e da ciência cidadã e se realiza também com base na percepção de que a cauda longa é um catalisador importante das pesquisas interdisciplinares devido à sua diversidade intrínseca. Nessas instâncias, reconhecem-se as grandes possibilidades das coleções pequenas e médias produzidas nas universidades e institutos de pesquisa, que são dispersas, mas valorizadas "por sua cobertura, seu ciclo de vida, as observações e as variáveis que são únicos". (HEDSTROM; MYERS, 2014, p.1).

Por fim, é importante observar que, nem sempre, a grande ciência, com seus predicados definidos por padrões, homogeneidade e estabilidade, é o modelo adequado para algumas das mais sofisticadas áreas de pesquisa. Por exemplo, para a área de Neurociência é na cauda longa em que, de fato, reside a *mainstream* da pesquisa. "Compreender o cérebro requer um esforço cooperativo de integrar informações, [...] combinando dados gerados por diferentes técnicas praticadas por várias disciplinas da neurociência", nos informa Ferguson e seus colaboradores (2014, p. 1442). Muitas vezes, a integração entre os dados da cauda e da cabeça, como acontece em alguns segmentos da Astronomia, é o modelo mais eficiente.

Complexidade dos dados da cauda longa

Embora grande parte das diferenças entre a pequena e a grande ciência seja contundente, algumas características, no plano da gestão de dados, são comuns entre esses dois mundos. Não há dúvida de que a gestão e a curadoria devem sustentar a confiança nos dados. Isso é efetivado por várias camadas de gestão e de curadoria, como, por exemplo, pela adição de metadados e de documentação de qualidade que contextualizem e garantam a interpretação e a proveniência dos dados, e pelo arquivamento em infraestruturas baseadas em repositórios certificados que assegurem a persistência e a estabilidade das coleções de dados. Essas infraestruturas devem proporcionar aos dados a possibilidade de serem descobertos, acessados, interpretados e reusados e de estarem linkados a outros recursos. Mas a diversidade e a complexidade dos dados da cauda longa impõem alguns desafios importantes para sua gestão e curadoria que precisam de um grau a mais de reflexão.

A metáfora da cauda longa dos dados é um território "onde os dados são altamente heterogêneos e dispersos entre instituições, projetos, laboratórios, pequenos grupos de pesquisa e pesquisadores individuais". A cauda longa da pesquisa é constituída de dados, usuários, questões de pesquisa e metodologias que diferem fundamentalmente da forma como a "cabeça" se configura. No domínio da *big Science*, as pesquisas dependem de um número relativamente pequeno de grandes coleções de dados que são bem conhecidas, estão arquivados e preservados em grandes bancos de dados que oferecem vias que os tornam fáceis de ser descobertos e acessados. A dimensão grandiosa dessas coleções abrange o volume, a cobertura e uma base de usuários que gera uma demanda imediata para esses ativos. No outro extremo, os pesquisadores da cauda longa produzem e usam uma massiva quantidade de dados que está largamente dispersa

e mantida em milhões de arquivos de computador que são difíceis de descobrir e de usar (HEDSTROM; MYERS, 2017).

Os dados da cauda longa se distribuem em todos os domínios do conhecimento, o que significa dizer que sua gestão depende fortemente de contextos disciplinares e de fluxos de trabalho específicos de instituições, de laboratórios e de equipes de pesquisa que geram uma vasta gama de coleção de dados. A heterogeneidade dessas coleções varia em termos de volume, que vai de poucos bytes até a escala de exabytes; em termos de formato de arquivo e de uso de padrões proprietários provenientes da diversidade de instrumentos e das tecnologias geradoras dos dados; em termos de referência a modelos padronizados de dados e da complexidade dos objetos digitais, que vão de simples a compostos, e que podem ter diferentes versões e variar com o tempo.

As pequenas coleções de dados produzidas pelos projetos da cauda são, geralmente, muito pouco integradas e têm um baixo grau de compartilhamento. Esse segmento funciona tipicamente como uma "indústria artesanal", onde os dados são intercambiados e baseados nas relações profissionais e na comunicação pessoal (CRAGIN, 2010, p.1). Geralmente estão acompanhados de poucos metadados ou de nenhum e apresentam documentação pobre ou inexistente, que limita ainda mais o compartilhamento aos pesquisadores mais próximos e constitui um obstáculo decisivo para o reuso. Faltam ainda ferramentas e serviços apropriados para que os dados sejam explorados e minerados no contexto mais amplo da comunidade científica (WYBORN; LEHNERT, 2016). De forma diferente de campos como o da Física e o da Astronomia, que tendem a ter padrões nas práticas de geração de dados e nos meios de compartilhamento, no escopo das *small sciences*, isso não é comum ou esperado.

Esse rápido diagnóstico revela que a consistência, a estabilidade e a visibilidade dessas coleções são comprometidas pela falta de infraestruturas

gerenciais, tecnológicas, comunitárias e, possivelmente, culturais que assegurem o acesso persistente, o controle de qualidade e a padronização e a integração dos dados. Isso pode ser uma indicação de que a pesquisa da cauda longa precisa de sistemas mais próximos de suas idiossincrasias e de seus fluxos de trabalho para superar a opacidade de parte dos seus fluxos. O uso pela cauda longa de padrões específicos e não regulados implica que são necessários uma curadoria personalizada e o controle nos repositórios institucionais menores e, com maior ênfase, de repositórios disciplinares, como confirma Heidorn (2008). Além do mais, a importância da cauda longa para a ciência sugere que seus requisitos específicos devem ser contemplados durante o planejamento e a implementação de ciberinfraestruturas de pesquisa e na formulação das políticas científicas em âmbito nacional, internacional e local (e-IRG Task Force, 2016).

Infraestrutura de gestão de dados da cauda longa: alguns requisitos

A natureza fragmentada e multidisciplinar da pesquisa da cauda longa cria um espaço importante para novos *insights*, mas, ao mesmo tempo, impõe o desafio de criar infraestruturas capazes de tornar esses dados visíveis e reusáveis em ambientes em que não há, tradicionalmente, padrões nas práticas de geração e compartilhamento de dados. É preciso considerar, também, que as características das pesquisas da cauda longa sugerem que algumas das estratégias usadas para orientar os investimentos em infraestruturas para o *big data* podem não funcionar tão bem para a cauda longa, que pode se beneficiar mais de novas abordagens de desenvolvimento de infraestruturas que se ajustem aos diferentes conjuntos de especificidades de suas comunidades. São muitas as variáveis que devem ser consideradas na concepção de sistemas de gestão de dados: tecnológicas, gerenciais, econômicas, sociais e, sobretudo, disciplinares e culturais.

Hedstrom e Myers (2014) identificam algumas características das comunidades científicas da cauda longa que podem ser consideradas na concepção de infraestruturas para a gestão de dados.

- Foca em problemas que requerem dados, métodos, ferramentas e expertise provenientes de múltiplas disciplinas;
- Precisa de muitos e diferentes tipos de dados sobre fenômenos físicos, naturais e sociais, com o objetivo de entender as interações entre os sistemas naturais e os humanos;
- Usa uma combinação de dados (de campo) observacionais, dados experimentais, simulações e modelos.
- Conduz as pesquisas em laboratórios ou centros de pequeno e de médio portes que estão sob a direção de um único coordenador de pesquisa ou um diretor de Centro.

Essas características são usadas na concepção do Sustainable Environment/Ationable Data (SEAD) já referenciada. Essa plataforma se volta para a gestão durante todo o ciclo de vida dos dados da cauda longa e oferece ferramentas com as quais os pesquisadores podem gerenciar seus dados com mais facilidade, da geração até o arquivamento, e publicar em um repositório digital apropriado, além de apoiar a interpretação e o compartilhamento. Um requisito inicial considerado pelo SEAD é que as infraestruturas para gerir os dados da cauda longa precisam fazer uma transição suave entre o simples armazenamento de dados – posto que parte significativa deles é armazenada nos computadores pessoais dos pesquisadores ou em serviços comerciais como o Dropbox - para níveis mais elaborados de tratamento, como, por exemplo, a implantação de sistemas e fluxos de trabalho que permitem que os pesquisadores assinalem mais facilmente metadados disciplinares e documentação para suas coleções de dados, para que sejam publicados e, posteriormente, recuperados e

compartilhados. Os objetivos do programa explicitam os grandes desafios da gestão dos dados da cauda longa (HORSTMANN, 2015):

- Compreender bem mais a pesquisa da cauda longa;
- Equacionar os desafios envolvidos na gestão de diversos *datasets*;
- Compartilhar e desenvolver práticas para gestão de dados heterogêneos;
- Trabalhar para aumentar a interoperabilidade entre repositórios.

Wyborn e Lehnert (2016) argumentam que a dificuldade com as coleções da cauda longa é de que só podem existir uns poucos especialistas ou grupos de pesquisa desenvolvendo uma coleção particular de dados, porém há muitas centenas dessas coleções de dados especializadas. Não é economicamente viável desenvolver uma única solução para cada coleção específica. A economia de escala não vai funcionar dessa forma. Propõem, em seguida, uma abordagem em três instâncias:

- **Repositórios focados em domínios específicos** – Os repositórios multidisciplinares e institucionais não têm expertise e infraestrutura para atender aos múltiplos requisitos dos domínios específicos e assegurar a integração e a reusabilidade; os repositórios disciplinares podem oferecer serviços sofisticados de curadoria, agregação, análise, estatísticas, visualização e modelagem.
- **Uso de padrões** – Uso de conceitos do ISO Observation and Measurement Model, que oferece um modelo geral e um esquema para apoiar e empacotar observações de instrumentos de laboratório, sistemas de sensores e outros; e o uso de esquemas padronizados de metadados, taxonomias e ontologias para integrar e interoperabilidade dos dados armazenados em múltiplos sistemas de dados.
- **Padronização da saída de instrumentos** – Trabalhar em cooperação com os fabricantes de instrumentos, porquanto os dados

da cauda longa são frequentemente compartilhados em formatos proprietários e inacessíveis, que tornam a conversão difícil para formatos abertos e, às vezes, impossível.

Outras ações podem contribuir para a visibilidade e sustentabilidade dos dados da cauda longa, como, por exemplo:

- **Integrar os dados da cauda longa à *big science*** - A integração dos dados da cauda longa aos padrões da grande ciência é viável em alguns campos disciplinares. Por exemplo, na pesquisa em Astronomia, em que uma infraestrutura de dados - mundial, distribuída e interoperável - que pertencem à *big science* contém as observações de telescópios localizados na superfície da Terra ou no espaço são colocadas disponíveis pelos arquivos dos observatórios, também contém dados que pertencem à cauda longa, em particular, resultados de pesquisas vinculados a publicações em periódicos acadêmicos, curados em centros de dados disciplinares.
- **Integrar os sistemas de curadoria aos fluxos de trabalho dos laboratórios** - Projetar plataformas de gestão que possam ser adaptáveis aos fluxos de pesquisa, desde o projeto da pesquisa até a transferência dos dados para repositórios disciplinares.

À guisa de conclusão

Os estudos sobre os dados de pesquisa produzidos pela cauda longa e a atenção das organizações voltadas para esse propósito, como é o caso da RDA, sugerem que esse é o tempo de se voltar a dar atenção à ciência produzida pelos pequenos laboratórios, seja por causa dos dados que produzem, seja por inovar e pela capacidade disruptiva proporcionada

pelos graus de independência por parte dos seus pesquisadores de esquemas mais rígidos adotados pelos grandes projetos.

A mais óbvia barreira para o reuso dos dados de pesquisa da cauda longa é a falta de infraestruturas que possibilitem a descoberta de dados que sejam úteis para estudos específicos, modelos ou decisões. O equacionamento das muitas e diferentes variáveis envolvidas se inicia com o estabelecimento de políticas nacionais abrangentes, que envolvam questões como financiamento contínuo, políticas mandatórias e esquemas de reconhecimento e de recompensa aos pesquisadores que publicam seus dados. Essas políticas são um guarda-chuva para as políticas institucionais que, finalmente, refletem-se nas estruturas tecnológicas e gerenciais das comunidades específicas. É necessário, entretanto, como enfatizam Hedstrom e Myers (2014), que os sistemas sejam de baixa barreira de entrada para redes que possibilitem a descoberta de ações sobre os dados e que os pesquisadores e seus colaboradores façam melhorias pequenas e incrementais sobre os dados à medida que eles os usem e tornem sua curadoria uma parte integral do seu compartilhamento e de seu uso.

Por fim, as exigências da cauda longa da ciência devem ser tomadas em conta durante a implementação de ciberinfraestruturas e de formulação de políticas em âmbito nacional, internacional e local. Além disso, a cauda longa da ciência deve ser considerada e beneficiada por todos os desenvolvimentos e recursos investidos nas ciberinfraestruturas voltadas para a grande ciência.

Referências

BORGMAN, C. L.. The conundrum of sharing research data. *Journal of the Association for Information Science and Technology*, v. 63, n. 6, p. 1059-1078, June 2012. Disponível em: <https://dl.acm.org/citation.cfm?id=2222887>. Acesso em 23 nov. 2017.

BOYD, D.; CRAWFORD, K.. Critical questions for Big Data: provocations for a cultural, technological and scholarly phenomenon. *Information, Communication & Society,* v. 15, n. 5, p. 662-679. Disponível em: <https://www.microsoft.com/en-us/research/wp-content/uploads/2012/05/CriticalQuestionsForBigDataICS.pdf >

CRAGIN, M. et al. Data sharing, small science and institutional repositories. *Philsophical Transaction of the Royal Society*, v. 368, p. 4023-403, 2010.

E-IRG Task Force**.** *Long tail of data***.** Netherland: e-IRG, 2016. Relatório. Disponível em <http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf >. Acesso em: 23 nov. 2017.

FERGUSON, A. *et al*. Big data from small data: data-sharing in the 'long-tail' of neuroscience. *Nature Neuroscience*, v. 17, n. 11, Nov. 2014. Disponível em <https://www.nature.com/articles/nn.3838 >. Acesso em: 23 nov. 2017.

HEDSTROM, M.; MYERS, J.. *SEAD:* finding the long tail lost in big data. 2014 Disponível em <http://sead-data.net/sites/default/files/pubs/NDS-White-Paper-The-Long-Tail-Lost-in-Big-Data-2.pdf >. Acesso em: 23 nov. 2017.

HORSTMANN, W.. *Beyond big bata***:** the long tail of research. 2015. Apresentação em Power-point. Disponível em <http://e-irg.eu/documents/10920/288074/2+Wolfram+Horstmann.pdf >. Acesso em: 23 nov. 2017

HEIDORN, B. P. Shedding light on the dark data in the long tail. *Library trends*, v. 57, n. 2, p. 280-299, 2008. Disponível em: <https://www.ideals.illinois.edu/bitstream/handle/2142/9127/Heidorn_LongTail_PreprintwEdits.doc.pdf?sequence=7 >. Acesso em: 23 nov. 2017.

PAYETTE, S.; HEDSTROM, M.; PLALE, B. SEAD: Infrastructure for managing research data in the long tail. In: GRACE HOPPER CONFERENCE FOR WOMEN IN COMPUTING (GHC15), Houston, TX. oct. 2015. *Anais…* Houston, TX, 2015. Disponível em: <http://sead-data.net/sites/default/files/news/GHC15-SEAD-Datanet_0.pdf > Acesso em 24 nov. 2017.

THE ROYAL SOCIETY. Science as an open enterprise. London: The Royal Society Science Policy Centre, 2012. Disponível em: <https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf>. Acesso em: 23 nov. 2017.

WYBORN, L.; LEHNERT, K.. Exploiting the long tail of scientific data: making small data BIG. In: ERESEARCH AUSTRAULASIA CONFERENCE, Melbourne, Australia, 10-14 Oct. 2016. *Anais...* Melbourne, Australia, 2016. Disponível em: <https://eresearchau.files.wordpress.com/2016/03/eresau2016_paper_88.pdf>. Acesso em: 23 nov. 2017.

3

CAMADAS DE REPRESENTAÇÃO DE DADOS E SUAS ESPECIFICIDADES NO CENÁRIO CIENTÍFICO

Plácida Leopoldina Ventura da Costa Santos
Ricardo César Gonçalves Sant'Ana

> For all experience and for the possibility of experience, understanding is indispensable, and the first step which it takes in this sphere is not to render the representation of objects clear, but to render the representation of an object in general, possible. [1] (Kant, Critique of Pure Reason, 1999).

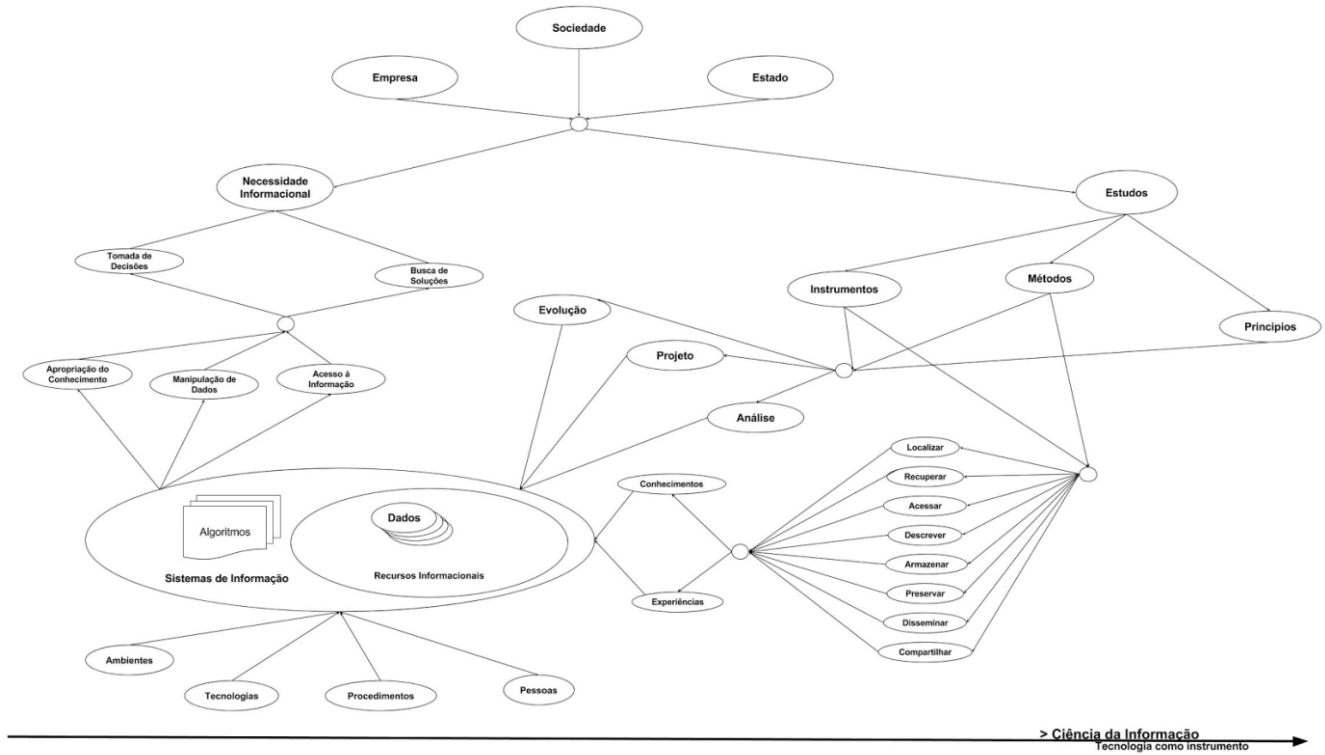
A busca por soluções para responder à necessidade informacional da sociedade ganha importância crescente em tempos de grandes quantidades de dados disponíveis. Por seu lado, o desenvolvimento e a implementação dessas soluções exigem estudos que possibilitem definir princípios, métodos e instrumentos que contemplem a análise, o projeto e a evolução dos sistemas de informação. Esses sistemas são constituídos dos seguintes elementos: ambientes, pessoas, recursos informacionais, tecnologias e procedimentos (LAUDON; LAUDON, 2014).

1 "A toda a experiência e à sua possibilidade pertence o entendimento, e a primeira coisa que ele faz para tanto não é tornar clara a representação dos objetos, mas tornar possível a representação de um objeto em geral" (KANT, 1999).

As metodologias e os instrumentos desenvolvidos ao longo do tempo para localizar, descrever, armazenar, preservar, acessar, recuperar, disseminar e compartilhar as experiências e os conhecimentos humanos convertidos em dados e corporificados na forma de recursos informacionais, necessitam de atenção em várias áreas do conhecimento e de modo especial na Ciência da Informação.

Quanto ao foco e ao direcionamento desses estudos, justifica-se direcioná-los aos aspectos relacionados à dimensão dos sistemas de informação baseados em tecnologia digital, devido a fatores como volume, variedade e velocidade dos dados que podem ser acessados, que tornam o uso das tecnologias essencial. Desse cenário, emerge o protagonismo técnico e científico na sustentação da capacidade para a busca de soluções e a tomada de decisões em processos tão amplos como os que envolvem a manipulação de dados, o acesso à informação e a apropriação do conhecimento.

Figura 1: Cenário de fatores e elementos envolvidos no acesso a dados



Fonte: elaborado pelo autor

Vale lembrar que as tecnologias de informação e comunicação (TIC) e os processos adjacentes envolvidos na administração de informações aumentaram o volume de dados e de informações disponíveis e só através deles é que se poderão encontrar soluções para gerir e controlar essa quantidade de dados que possa contribuir para um acesso mais simétrico entre todos os envolvidos, porque o tratamento de dados para a geração de conhecimentos é um recurso inesgotável que se autoalimenta do compartilhamento (SANTOS; SANT'ANA, 2002).

A escolha do termo 'através', no parágrafo anterior, ao se referir às TIC se justifica por demonstrar que se está falando não só de se usar o recurso, mas também de digeri-lo como parte da construção de novos caminhos, passando, necessariamente, por um processo de ampliação do alcance da 'datificação' (MAYER-SCHOENBERGER; CUKIER, 2013) composto, agora, de uma ressignificação do papel da tecnologia e dos dados na questão do acesso a conteúdos.

Acrescente-se a essa discussão a fluidez desses tempos de constantes mudanças proporcionadas pelo uso e pelo desenvolvimento acelerado de recursos tecnológicos fortemente empregados nos processos de acesso a dados, de uso da informação e de geração de conhecimentos (BAUMAN, 2001). Assim, é preciso ampliar o quadro referencial sobre as possibilidades de interpretar e de analisar o próprio conceito de dado.

Nesse sentido, propõe-se aprofundar a interpretação do conceito de dado, que Santos e Sant'ana (2015, p. 205) definem como

> [...] uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que,

> mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação.

O dado é concebido como um elemento que pode ser diretamente tratado por instrumentos digitais, considerando sua composição baseada na tríade **entidade - atributo - valo**r <e, a, v> (BOOCH, RUMBAUGH e JACOBSON, 1999), que é a base para modelagem de dados bastante utilizada no mapeamento de dados heterogêneos (NADKARNI et al., 1999), citada nos textos de inteligência artificial (WINSTON, 1992), e que se originou no conceito de listas de associação utilizadas na linguagem LISP.

A proposta deste texto é de integrar o contexto do usuário e a semântica mínima necessária ao contexto dos dados para que possam ser interpretados.

Desafios

A imensa quantidade de dados e de informações disponíveis não terá serventia se não houver meios para armazená-los e gerenciá-los de forma eficiente e viável. Já é possível, entretanto, afirmar que há uma capacidade de recursos tecnológicos para o armazenamento, que vem sendo eficiente em termos de velocidade e de custo, oportunizando o armazenamento de conteúdos em abundância antes inimaginável.

Assim, a atenção deve ser direcionada à recuperação dos novos patamares de informação que ficaram disponíveis. Nesse ponto, apesar de o assunto estar voltado para os dados, é possível citar Araújo (2009), quando diz que "a grande questão passou a ser não mais a aquisição de livros, mas sua organização, de forma a se conseguir promover a sua recuperação e o

seu uso". O grande e crescente volume de dados armazenados e ao alcance de acesso direto transformou-se em um desafio na busca por soluções para os problemas sobre como construir pontes entre os usuários e os bancos de dados na tarefa de encontrar os dados desejados e, principalmente, de interpretá-los.

Na busca por soluções e planejamento, é importante ressaltar que a estrutura de um dado é estabelecida pela identificação dos seguintes elementos <e> e <a>, que pertencem à tríade, entidade - atributo - valor <e,a,v>, que compõe o seu esquema, e é por meio da estrutura que se podem viabilizar a utilização e a reutilização de dados.

Dados estruturados tem sua estrutura interna explícita e seu conteúdo <v> interno é evidenciado por meio de uma semântica que permite que todo o seu conteúdo seja interpretado de forma autônoma por máquina. Dados não estruturados, por sua vez, exigem interpretador externo para identificar sua estrutura e semântica e interpretar seu conteúdo.

A interpretação de dados, especialmente os não estruturados, solicita a utilização de metadados para ser representados.

A representação dos dados em um sistema de informação se dá pela utilização de metadados, e sua relevância é percebida na complementação da estrutura semântica mínima da tríade <e, a, v> que compõe determinado dado.

Ao pensar no dado como um elemento puramente sintático, com uma baixa carga semântica, tem-se como consequência um esforço adicional para representá-lo.

> [...] já que não bastam elementos que o descreva como um todo e que propiciem sua recuperação. São necessários, ainda, elementos que permitam a sua interpretação por quem os acesse, com informações que detalhem sua

> estrutura e possibilitem a interpretação de cada atributo que os compõe. (SANT'ANA, 2017, p. 4).

Nesse momento, o foco já deve estar no usuário, que, independentemente de suas habilidades e competências, usará os dados vinculados a determinado contexto, que deve estar tangível em sua representação, de modo a favorecer sua visibilidade, utilização e reutilização (SANT'ANA, 2017).

Construção de pontes entre usuários e bases de dados

A interpretação dos dados por parte dos usuários pressupõe a interpretação dos elementos semânticos mínimos que o constituem (SANTOS; SANT'ANA, 2015; NADKARNI et al., 1998). Isso significa vincular a definição da entidade e do próprio atributo a conteúdos que possibilitem sua devida contextualização. Para que essa vinculação seja possível, é preciso viabilizar um processo de construção de camadas de representação, um processo que envolve fatores como tempo, atores, recursos e o próprio conteúdo.

Quando nos referimos ao conteúdo, estamos falando do próprio valor do dado. Assim, podemos considerar como primeiro fator envolvido a dimensão tempo, não a relacionada ao próprio conteúdo, ou seja, a data de referência daquele valor, mas a que faz parte dos ciclos de vida em que estará envolvido. Esse tempo envolve as fases de coleta, de armazenamento, de recuperação e de descarte, o que irá envolver atores e recursos específicos.

Quanto aos atores, temos os envolvidos no processo de obtenção/captação desses dados, ou seja, diretamente ligados à fase de coleta e que, portanto, têm contato direto com a fonte ou a forma de obtenção. A

eles cabe identificar e obter não somente o valor ou conteúdo nuclear do dado, mas também contextualizar esse valor de tal forma que ele possa ser utilizado. Esse processo de coleta requer planejamento e ações para que esses dados sejam obtidos e estejam disponíveis ao menos enquanto a coleta esteja em processo.

Nessa fase, os objetivos do acesso ao dado já podem ser satisfeitos, caso sejam imediatos. Para ilustrar, podemos imaginar uma câmera em um veículo autônomo que, ao coletar imagens, processa-as para definir as variáveis que propiciarão as decisões de agir na condução do veículo rumo ao destino, mas essas imagens não são necessariamente registradas já que seu uso é imediato. Nossa visão também é um exemplo desse processo, já que não memorizamos tudo o que vemos. Assim, esses dados são obtidos, utilizados e descartados. No entanto, parte dos dados coletados pode ser necessária para um acesso posterior, o que leva a uma nova fase do ciclo de vida dos dados, em que os dados são registrados. Nessa fase, novamente se deve levar em conta todo um processo de planejamento e de execução de ações que possibilitem que esses dados estejam disponíveis em um suporte para acesso futuro e que podem envolver outros atores, principalmente para planejar e executar, já que essa fase requer competências e conhecimentos distintos da fase anterior.

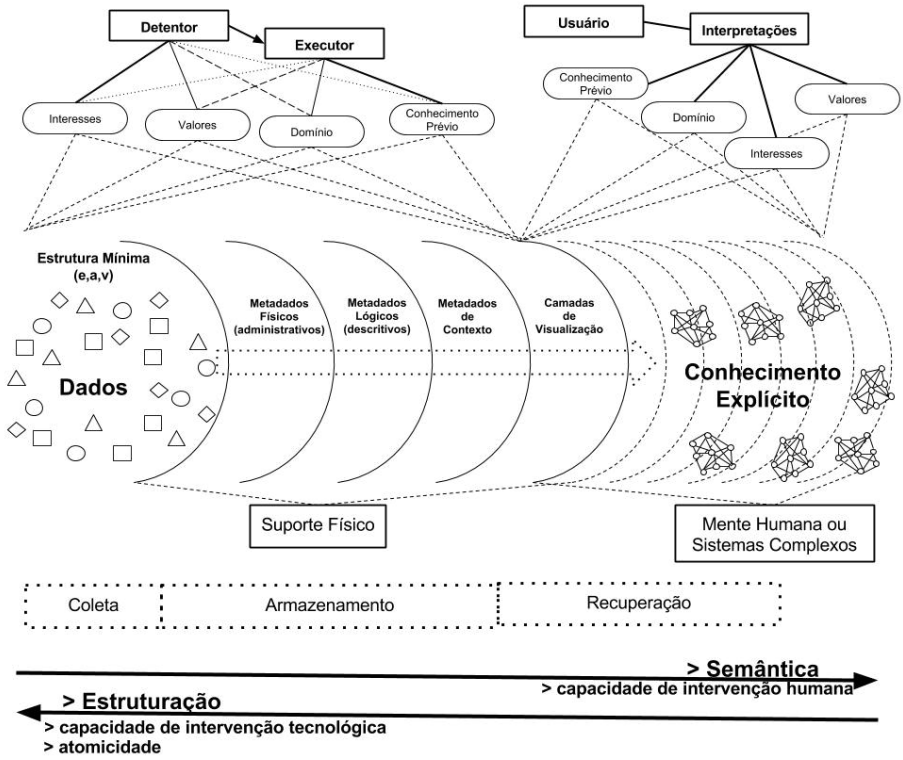
Esse acesso futuro implica um processo em que possam ocorrer o acesso e a interpretação do valor armazenado, o que se configura como um momento em que será preciso, novamente, planejar e executar ações específicas para que esse acesso seja possível. Essa é a fase de recuperação, que também pode envolver atores outros, e não, os envolvidos nas fases anteriores, também em função das necessidades específicas envolvidas. É nessa fase em que as competências estudadas e desenvolvidas na Ciência da Informação têm mais adesão entre todas as fases (SANT'ANA, 2016).

Assim, na fase de coleta e de armazenamento, configuram-se competências específicas para os que irão efetivar as ações necessárias - os 'executores' - que devem responder pelos conhecimentos técnicos requeridos pelo processo. No entanto, para realizar essas fases, é necessário empenhar recursos que vão desde a disponibilidade de dispositivos tecnológicos até o comprometimento de recursos humanos (executores). Adotamos a definição de 'detentores' para os que respondem pela posse desses bens de produção e que são parte importante desse processo, principalmente na definição de seus objetivos, o que impacta todo o modelo de representação desses dados.

Essa relação entre detentores e executores se faz por meio de fatores como interesses, valores, domínio e conhecimento prévio, presentes nos dois lados da relação, mas com intensidades distintas, conforme apresentado na Figura 2. De um lado, temos os interesses de forma mais intensa no flanco dos detentores, que podem ser individuais, coletivos e institucionais; e no lado dos executores, mais intensidade de relevância do conhecimento prévio.

Dados coletados e armazenados abrem a possibilidade de disponibilizar acesso a estes dados – é a fase de recuperação - que mantêm as mesmas relações entre os detentores e os executores, porém, a partir de então, precisam ser interpretados, o que requer a participação de novos atores - os usuários – que também os interpretam sob o impacto de fatores como conhecimento prévio, domínio, interesses e valores, que devem ser levados em consideração não só no planejamento e na execução das fases de coleta, de armazenamento e de recuperação, como também do estudo e da análise de cenários dos processos de acesso a dados.

Figura 2: Elementos, fatores e fases no cenário de acesso a dados



Fonte: elaborado pelo autor

Conforme demonstrado na Figura 2, pode-se perceber que o processo de acesso a dados apresenta momentos e atores distintos, o que torna o processo complexo e frágil. Podem ocorrer dissonâncias entre os valores coletados e as informações geradas por sua interpretação, na outra ponta do processo, e fragilidades na preservação e na proteção desses dados, o que configura um cenário de requisitos que vai além da estrutura mínima de composição de um dado (e, a, v) e justifica esforços na construção de camadas de representação que envolvem diferentes tipos de metadados.

Nessas camadas de representação, pode-se identificar uma camada mais próxima do contexto do suporte, uma camada de metadados físicos que respondem pelos metadados de gestão do próprio recurso que sustenta o acesso a um banco de dados, a uma planilha ou a uma arquitetura menos estruturada e que atende aos requisitos de integração e de integridade física dos dados. A NISO (RILEY, 2017) identifica uma parte desses tipos de metadados como administrativos, mais especificamente, como metadados técnicos e de preservação, e outra como descritivos, ao citar a identificação das entidades e dos atributos. A identificação desses metadados apresenta uma forte dependência dos executores que desempenham atividades mais próximas dos recursos tecnológicos.

No nível de abstração, inserem-se os metadados lógicos, que têm uma carga semântica maior de uso do que de gestão física dos dados, como na camada anterior, atendem aos requisitos de preservação, privacidade e integridade lógica dos dados, mantêm uma relação com a gestão e com o uso dos dados e incluem metadados que a NISO identifica (RILEY, 2017) como descritivos com os metadados sobre as relações entre os dados, administrativos com os metadados de preservação e de direitos autorais.

Finalizando a fase de armazenamento, temos a camada de metadados de contexto, que incluem elementos que atendem a requisitos como proveniência e fatores mais gerais relacionados à qualidade desses dados, que a NISO (RILEY, 2017) chama de estruturais, com os metadados sobre as relações entre as partes e os relacionados às linguagens de marcação, que incluem dados sobre elementos de formatação, destaques e quebras oriundos do processo de coleta e que contribuem para o contexto do dado armazenado.

Na fase de recuperação, os atores que irão interpretar os dados - os usuários - são considerados de forma mais direta (todos os atores estão envolvidos em todo o processo, em todas as fases, mas com intensidades

distintas em cada uma delas). Portanto, nessa camada, a localização, a disponibilidade e o acesso aos dados são considerados sob o ponto de vista da necessidade do usuário que se pretende atender, seja ele específico ou não, e fica vinculada aos processos de coleta de eventuais usuários.

Por fim, esses dados podem ser acessados e interpretados gerando resultados não só para o detentor como também para os usuários, que poderão instanciar essas interpretações em forma de relatórios ou de publicações, no caso do contexto científico, gerando novos ciclos de vida, em que a fase de recuperação alimenta fases de coleta para novos usos desses dados.

Nesse processo, a carga semântica dos dados cresce, a cada camada de representação, e tanto a adequação para uso de recursos tecnológicos quanto a atomicidade desses dados decrescem. Convém enfatizar que essas camadas são interdependentes, e as fases não são autoexcludentes, ou seja, elas se sucedem, mas podem ocorrer concomitantemente ou se retroalimentarem em um processo cíclico, que é a essência do próprio conceito de ciclo de vida dos dados.

Considerações

No contexto científico, esse é um cenário crítico, devido às suas especificidades, como os atores envolvidos, a relevância de fatores como conhecimento prévio e o domínio e os tempos próprios relacionados à questão da disponibilidade dos dados.

Em relação aos atores envolvidos, no cenário científico, os objetivos relacionados aos ciclos de vida dos dados tendem a se vincular às pesquisas, portanto, sob a responsabilidade do pesquisador, o que pode gerar dificuldades para definir quem seria o detentor, quando se consideram as instâncias envolvidas, como a instituição à qual ele está vinculado, as agências de fomento e, até, os governos e eventuais empresas parceiras.

A execução também pode ser compartilhada entre os colaboradores e o próprio pesquisador, que divide essa responsabilidade e atua mais intensamente no contexto do domínio em que os dados estão inseridos, mas pode receber aportes de colaboradores nas questões que requerem competências mais específicas, no que diz respeito ao uso e à adequação dos recursos de tecnologias da informação e comunicação e na operacionalização do processo.

Já os fatores envolvidos com os que respondem como detentores ganham complexidade, devido à influência adicional de interesses públicos, princípios éticos e autoria. Os interesses públicos tendem a justificar a abertura do acesso aos dados o quanto antes como parte da contrapartida de recursos investidos na pesquisa e em prol do bem comum. Os princípios éticos, por outro lado, reforçam a necessidade de preocupações como as voltadas para garantias de direitos como privacidade e de transparência na realização de pesquisas para garantir que os limites morais e éticos não sejam desrespeitados. Quanto às questões de autoria, dificultam a cessão de acesso aos dados, principalmente enquanto eles não tenham sido totalmente explorados pelo detentor, para garantir que os esforços investidos no processo tenham respaldo com a publicação de resultados inéditos.

Todas essas questões abrem novas e amplas possibilidades de discussão que extrapolam o escopo deste texto. Por isso esperamos que ele seja motivo para se ampliarem debates e reflexões sobre elas.

Referências

BAUMAN, Z. *Modernidade líquida.* Rio de Janeiro: Zahar, 2001.

BOOCH, G.; RUMBAUGH, J.; JACOBSON, I. *The unified modeling language user guide*. Reading, Mass.: Addison-Wesley, 1999

KANT, I. *Crítica da razão pura*. São Paulo: Nova Cultural. 1999 p.178.

LAUDON K.C.; LAUDON, J.P. *Sistemas de informação gerenciais.* 11. ed. São Paulo: Pearson, 2014.

MAYER-SCHOENBERGER, V.; CUKIER, K. *Big data*: a revolution that will transform how we live, work, and think. Londres: John Murray, 2013.

NADKARNI P.M, BRANDT C, FRAWLEY S. Managing attribute-value clinical trials data using the ACT/DB client-server database system. *Journal of the American Medical Informatics Association*. v.5, n.2, mar-abr, 1998. p.139-151.

NADKARNI P.M., MARENCO L, CHEN R, SKOUFOS E, SHEPHERD G, MILLER P. Organization of heterogeneous scientific data using the EAV/CR representation. *Journal of the American Medical Informatics Association*. v.6, n.6, nov-dez. 1999. p.478-493.

RILEY, J. *Understanding metadata*: what is metadata, and what is it for? Baltimore, National Information Standards Organization (NISO), 2017. NISO Primer Serie.

SANTOS, P. L.V. A.C.; SANT'ANA, R. C. G. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela Ciência da Informação. *Ciência da Informação*, Brasília, v. 42, n. 2, jan. 2015. ISSN 1518-8353. Disponível em: <http://revista.ibict.br/ciinf/article/view/1382>. Acesso em: 25 jan. 2017. doi:https://doi.org/10.18225/ci.inf..v42i2.1382.

SANTOS, P.L.V.A.C.; SANTANA, R.C.G. Transferência da informação: análise para valoração de unidades de conhecimento. *DataGramaZero* - Revista de Ciência da Informação - v.3 n.2 abr/2002. Disponível em: http://www.brapci.inf.br/index.php/v/a/1259//

WINSTON P.H. *Artificial intelligence*. 3. ed. [s.l.]: Addison Wesley, 1992.

4

A PRIVACIDADE E A QUESTÃO DOS DADOS

Tassyara Onofre de Oliveira
Bernardina Maria Juvenal Freire de Oliveira
Guilherme Ataíde Dias

A privacidade é um conceito amplo no que diz respeito à proteção da autonomia individual e da relação entre um indivíduo e a sociedade (incluindo governos, empresas e outros indivíduos). A definição de privacidade varia entre países e indivíduos, com base em experiências passadas e entendimentos culturais. O termo 'privacidade' originou-se na língua inglesa a partir da palavra *privacy.* Bastos e Martins conceituam a privacidade como a

> [...] faculdade que tem cada indivíduo de obstar a intromissão de estranhos em sua vida privada e familiar, assim como de impedir-lhes o acesso a informações sobre a privacidade de cada um, e também impedir que sejam divulgadas informações sobre essa área da manifestação existencial do ser humano. (BASTOS; MARTINS, 1989, p. 63).

Podemos assumir que a privacidade envolve tudo o que o indivíduo deseja ocultar sobre o conhecimento público. Atualmente, é entendida como um direito fundamental e expressão da dignidade

humana, pois as informações e os dados pessoais contribuem, de forma basilar, para construir a imagem do indivíduo perante o mundo. O direito à privacidade não diz respeito somente a "esconder" determinados dados, mas também envolve aspectos como o acesso a eles, seu controle, sua utilização e o processamento de todos os dados pessoais. De acordo com Doneda,

> [...] a sutil definição do que é exposto ou não sobre alguém, do que se quer tornar público ou o que se quer esconder, ou a quem se deseja revelar algo, mais do que meramente uma preferência ou capricho, define propriamente o que é um indivíduo. (p. 78, 2010).

Uma crescente preocupação com a tutela jurídica da privacidade é própria de nosso tempo. A ideia de privacidade não é recente, porquanto já existia em outras épocas e sociedades. Os povos antigos vivenciavam uma intensa vigilância em suas atividades, já que havia interdependência entre os integrantes da família. A vida privada era centrada nas dinâmicas relacionadas ao dia a dia da vida em família. Assim, devido à importância crescente dos dados na sociedade contemporânea, nada mais natural que haja, na mesma proporção, uma preocupação crescente em proteger os direitos de seu titular.

Díaz (2013) refere que, na Roma antiga, a casa (o lar doméstico) desempenhava um papel muito importante na família, porque era considerado um lugar sagrado. Durante a Idade Média, a intimidade era um privilégio atribuído a poucos, como os senhores feudais, que detinham a propriedade, inclusive, de seus vassalos. Nessa época, a intimidade era uma extensão da propriedade de praticamente tudo, ou melhor, a propriedade era uma condição para se ter privacidade. O entendimento filosófico da privacidade como um pressuposto fundamental do homem surgiu antes de sua concepção jurídica, a partir do pensamento cristão.

Santo Agostinho se reporta à intimidade como um momento em que o homem, sozinho, reflete sobre si mesmo em sua relação com Deus.

Ressalte-se, entretanto, que a privacidade só começou a se aproximar das características atuais no Século XIX e a se apresentar como nós a conhecemos no Século XX, principalmente depois do surgimento da *Internet*. Esse novo contorno e o uso da privacidade foram marcados a partir do influente artigo intitulado *The Right To Privacy,* escrito por Samuel D. Warren e Louis D. Brandeis em 1890. Nesse texto, surgiu a primeira manifestação individual do conceito de "ser deixado só". Os autores colocam em evidência a ocorrência de transformações sociais, políticas e econômicas e o surgimento de novos inventos, como a fotografia, por exemplo, que contribuíram para a ocorrência de violações da vida privada das pessoas.

Respaldando-se nesse panorama, os autores analisam determinado número de decisões em tribunais ingleses e americanos, considerando a existência de um princípio geral na *common law*, que é o *right to privacy* - o direito a privacidade. O artigo declarava que os fotógrafos e as empresas jornalísticas invadiam o recinto sagrado da vida privada e doméstica. A vítima teria sido o próprio Samuel Warren, que era inconformado com a intromissão da imprensa em sua vida familiar. Eis um trecho do artigo:

> 1. O direito a privacidade não proíbe qualquer publicação da matéria que é de interesse público ou geral. [...] 2. O direito à privacidade não proíbe a divulgação de qualquer matéria, apesar de sua natureza privada, quando a publicação é feita mediante circunstâncias que tornem uma informação privilegiada de acordo com a lei de calúnia e difamação. [...] 4. O direito a privacidade cessa após a publicação dos fatos pelo individuo ou com seu consentimento. [...] Os remédios para uma invasão do direito a privacidade também são sugeridos por aqueles administradores da lei de difamação e na lei da propriedade literária e artística, nomeadamente:

> 1. Uma ação de responsabilidade civil por dados em todos os casos. (WARREN E BRANDEIS, 1890, p. 13, Tradução nossa)[1].

Warren e Brandeis (1890) consideram, em seu artigo, que a invasão da privacidade se configura como uma profunda ofensa, que afeta a noção do ser humano em sua individualidade, dignidade, independência e honra. O fato de um artigo publicado no ano de 1890 ainda ser considerado uma obra relevante a respeito do tema é notável, especialmente se levarmos em conta a importância e a atualidade da matéria.

Avançando cronologicamente em relação ao caso de Warren e Brandeis (1890), podemos mencionar o caso recente divulgado sobre o *Facebook* e a empresa de publicidade *Cambridge Analytica*[2] que provocou diversas reações em todo o mundo. Em relação a esse caso, o governo dos Estados Unidos convocou Mark Zuckerberg, cofundador e CEO do *Facebook*, para uma extensa sabatina no Congresso Norte-americano a respeito da legalidade do seu modelo de negócios, no que se refere às questões de privacidade.

O Brasil, por receber influência do direito romano, adota a *civil law*, que também vigora em grande parte do sistema jurídico europeu. Países sensíveis ao direito de privacidade priorizam o direito individual

---

1 I. The right to privacy does not prohibit any publication of matter which is of public or general interest.[...] 2. The right to privacy does not prohibit the communication of any matter, though in its nature private, when the publication is made under circumstances which would render it a privileged communication according to the law of slander and libel.[...] 4. The right to privacy ceases upon the publication of the facts by the individual, or with his consent. The remedies for an invasion of the right of privacy are also suggested by those administered in the law of defamation, and in the law of literary and artistic property, namely :- I. An action of tort for damages in all cases.

2 Vide: https://www.theguardian.com/technology/2018/dec/19/facebook-cambridge-analytica-washington-dc-lawsuit-data

em defesa do interesse da pessoa, talvez por causa dos reflexos causados pelo uso da informação como uma estratégia na Segunda Guerra Mundial. Após o término do conflito, foi publicada a Declaração Universal dos Direitos Humanos (1948), cujo artigo 12º preconiza:

> Ninguém sofrerá intromissões arbitrárias na sua vida privada, na sua família, no seu domicílio ou na sua correspondência, nem ataques à sua honra e reputação. Contra tais intromissões ou ataques toda a pessoa tem direito a protecção da lei. (ONU, 1948, *online*).

Zagol e Tibiriçá (2011) interpretam que determinadas culturas enfatizam mais os direitos da comunidade do que os direitos individuais. Como exemplo, podemos citar os Estados Unidos da América e a Inglaterra, que utilizam a *Common Law*.

A tutela do direito à privacidade dos dados pessoais

Atualmente, o direito de ter os dados pessoais preservados encontra tutela em várias normas, o que inclui a Declaração Universal dos Direitos Humanos (1948), o Pacto Internacional de Direitos Civis e Políticos (1966), a Convenção Europeia dos Direitos do Homem (1953), a Convenção Americana de Direitos Humanos e Organismos Internacionais (1969) e a Declaração Americana dos Direitos e dos Deveres do Homem (1948). Existem também órgãos que visam garantir a proteção aos dados pessoais, como o Tribunal Europeu dos Direitos do Homem, a Organização das Nações Unidas (ONU) e o Conselho de Direitos Humanos, da Organização das Nações Unidas (ONU). No âmbito nacional, a Lei de Proteção aos Dados Pessoais, a Constituição Federal, o Marco Civil da Internet e outras leis almejam proteger a privacidade

no Brasil. No entanto, com exceção parcial da Lei de Proteção aos Dados Pessoais, entendemos que os diplomas não conseguem abarcar de forma satisfatória as necessidades associadas com a proteção da privacidade no contexto dos dados.

Legislação pátria e privacidade

O direito à privacidade tem um grau de suma importância na Constituição Brasileira. Considerado como um dos direitos de personalidade e, portanto, é revestido de característica própria de direito fundamental e cláusula pétrea. Em decorrência da fragilidade do objeto (privacidade), pode ser violado mais facilmente.

Nossa Constituição (1988) vigente coloca dois tipos de instâncias em relação à privacidade do indivíduo: a intimidade e a vida privada, o que resulta em uma multiplicidade de entendimentos de uma doutrina que procura formular definições para cada uma delas. O art. 5º, inciso X, de nossa Carta Magna tutela, de forma autônoma, o conceito de vida privada e o distingue de intimidade. Pressupõe-se que o constituinte utilizou a expressão vida privada em sentido estrito, como uma das esferas da intimidade.

Uma das fontes para se distinguir constitucionalmente a intimidade de vida privada é a prática jurídica francesa, grande influenciadora da doutrina civilista ocidental, que considera o direito de intimidade apenas como um aspecto mais restrito ao direito à vida privada. O art. 9º do Código Civil francês contribui para o entendimento destas ideias:

> Toda pessoa tem direito ao respeito de sua vida privada. Os Juízes podem, sem prejuízo da reparação do dano sofrido, ordenar todas as medidas, tais como sequestro, embargo e

> outras, aptas a impedir ou fazer cessar um atentado à intimidade da vida privada; essas medidas podem, se houver urgência, ser ordenadas em liminar (FRANÇA, 1994, *online*).

No âmbito da Internet, a proteção da privacidade do usuário evoluiu um pouco mais com a promulgação do Marco Civil da Internet (Lei n°12965/2014), especialmente em seu art. 11º:

> Art. 11º. Em qualquer operação de coleta, armazenamento, guarda e tratamento de registros, de dados pessoais ou de comunicações por provedores de conexão e de aplicações de internet em que pelo menos um desses atos ocorra em território nacional, deverão ser obrigatoriamente respeitados a legislação brasileira e os direitos à privacidade, à proteção dos dados pessoais e ao sigilo das comunicações privadas e dos registros (BRASIL, 2014, *online*).

De fato, o Marco Civil da Internet possibilitou avanços quanto à proteção da privacidade de dados pessoais. Mas, como fica claro, a proteção é exclusiva no ambiente da Internet. E a necessidade de segurança não é apenas no contexto da grande rede, embora vivamos em uma realidade em que quase tudo o que está codificado em formato digital trafega na Internet, ainda que existam situações de violação dos dados pessoais em outros contextos.

Apesar de a Constituição Brasileira diferenciar o direito a intimidade do direito a vida privada, os doutrinadores continuam entendendo que esses dois direitos devem ser tratados como sinônimos. A vida privada não é um conceito imutável no espaço e no tempo. Zagol e Tibiriçá (2011) comentam que esse conceito unívoco pode variar conforme a sociedade e depende da posição que cada indivíduo ocupa no momento específico, na sociedade específica. A vida exterior e a vida

profissional são componentes da vida privada. Assim, a privacidade detém um caráter mais subjetivo que, em muitas circunstâncias, impede uma resposta clara, harmônica e rápida para os problemas. Doneda (2011) explica que a proteção de dados pessoais tem, em seu campo de aplicação, um caráter mais objetivo, e sua finalidade é de proteger o dado e, por meio dele, a pessoa. Convém registrar que as regras de proteção de dados pessoais costumam ser muito mais concretas e específicas.

A proteção da personalidade corrobora a garantia de custódia, presente na cláusula geral da personalidade, que dispõe sobre o valor intangível da dignidade do ser humano, princípio consagrado constitucionalmente e integrador do ordenamento pátrio e que conduz as relações públicas e privadas para pôr fim à divisão que usualmente se aplica às relações jurídicas e pretende distinguir e definir os direitos de personalidade de outros direitos. Seja sua natureza de um direito fundamental, seja de um direito de personalidade, a privacidade demonstra o ponto em comum para o qual o ordenamento caminha: a preservação da dignidade humana.

Motivados pela necessidade de proteção de dados pessoais dos indivíduos, foi editada e promulgada a Lei n°13.709/2018 com o propósito de estabelecer regras para disciplinar à forma como os dados podem ser armazenados e utilizados por empresas ou mesmo por pessoas físicas. O objetivo da Lei n°13.709/2018 é proteger os direitos fundamentais de liberdade e privacidade e o livre desenvolvimento de personalidade da pessoa natural, independente do meio, do país de sua sede ou do país onde estejam localizados, desde que os dados tenham sido coletados no Brasil ou qualquer outra operação seja realizada na país, ou ainda que, a atividade tenha se realizado fora do Brasil, mas que tenha objetivo de ofertar serviços ou bens ou tratamento de dados dos indivíduos localizados no território nacional.

Legislações internacionais e privacidade dos dados

Não existe, até o momento, um tratado global que trate especificamente da proteção de dados. Contudo, como já exposto, a Declaração Universal dos Direito Humanos (1948), o Pacto Internacional Sobre os Direitos Civis e Políticos (1966) e a Convenção Americana sobre Direitos Humanos (Pacto de São José, em 1969) asseguram a não interferência na vida privada e familiar da pessoa, do seu lar e de sua correspondência.

A Declaração Universal dos Direito Humanos (1984) indica, em seu art. 18º, que ninguém será sujeito a interferências em sua vida privada, na de sua família, em seu lar ou em sua correspondência. O Pacto Internacional dos Direitos Civis e Políticos (1966), em seu art. 17º, proíbe, expressamente, ingerências arbitrárias ou abusivas na vida privada e familiar das pessoas e expressa e assegura a proteção à privacidade.

Quanto aos dados sensíveis, essas normas internacionais vedam a discriminação em diversas formas, como nos artigos 2º e 7º da Declaração Universal dos Direito Humanos, 1º, I, e art. 24º, do Pacto de São José, e o art. 25º do Pacto Internacional dos Direitos Civis e Políticos.

Teorias da privacidade

Duas teorias procuram identificar como se opera a privacidade: a Teoria dos Círculos Concêntricos (ou das esferas) e a Teoria do Mosaico.

Teoria dos Círculos Concêntricos (ou das esferas)

A Teoria dos Círculos Concêntricos ou das Esferas foi anunciada no ano de 1957 por Heirinch Henkel, durante o Fórum Jurídico Alemão. Essa teoria baseia-se no grau de sujeição da pessoa às ingerências externas.

> A esfera privada (o círculo da vida privada em sentido amplo) encerra três círculos concêntricos (camadas dentro de camadas): o círculo da vida privada em sentido restrito (a camada superficial), que contempla o círculo da intimidade (a camada intermediária), no qual se acomoda o mais denso desses três compartimentos, o círculo do segredo (núcleo) (FROTA, 2007, p. 461).

Observou-se, por meio dessa teoria, que a liberdade de informação e o direito à privacidade compõem círculos concêntricos de proteção e de conhecimento. No Brasil, o maior propagador dessa teoria foi Costa Jr. (2007, p.23), que faz uma distinção entre a esfera individual (proteção à honra) e a esfera privada (proteção contra a indiscrição).

Figura 1: Teoria dos Círculos Concêntricos

Fonte: Produzida pelos autores com base em Costa Jr. (1995)

Como contextualizado na Figura 1, de acordo com a interpretação de Costa Jr. (2007):

> Aqui, não se trata mais do cidadão no mundo, relacionado com os semelhantes, como na esfera individual. Trata-se, pelo contrário, do cidadão na intimidade ou no recato, em seu isolamento moral, convivendo com a própria individualidade (COSTA JR., 2007, p. 24).

A primeira esfera da privacidade corresponde ao estágio do anonimato, as condutas mais ocultas de uma pessoa estão na esfera privada. Abrangendo um grande número de relações interpessoais, inclusive aquelas mais superficiais. Pode-se cogitar em possível interesse público à informação de tais dados, na medida em que algumas circunstâncias do indivíduo podem ser consideradas relevantes para a sociedade. Como exemplo, os de fatos e informações que o indivíduo almeja, em uma primeira análise, excluir do conhecimento alheio, como a sua imagem, seus hábitos e costumes. Já no círculo intermediário, o estágio de intimidade onde são protegidos o sigilo profissional, sigilo domiciliar e algumas comunicações pessoais. São informações mais restritas, divididas com reduzido número de pessoas de confiança, como no ambiente familiar e amigos íntimos. O último estágio da intimidade pessoal, a solidão, seria um ponto dentro da esfera do segredo, o menor e mais oculto deles. São aqueles fatos ou informações cujo conteúdo o sujeito não deseja dividi-lo, apenas em restritas circunstâncias. Por exemplo, a opções sexual, religiosa e filosófica do indivíduo.

Teoria do Mosaico

Considerando que a Teoria das Esferas não era suficiente para enfrentar as mais modernas e sofisticadas formas de atacar a privacidade por intermédio das novas tecnologias, Conesa (1984) elaborou uma nova teoria mais integralizada, a Teoria do Mosaico:

> [...] existem prioridades, do ponto de vista do direito à privacidade, que, no entanto, conectadas com outros, talvez,

> e relevantes, também podem servir para tornar totalmente transparente a personalidade de um cidadão, como acontece com pequenas pedras que formam os mosaicos que, se não dizem nada, unidas, formam um conjunto cheio de significados (CONESA, 1984, p.45, tradução nossa)[3].

O autor já havia contemplado, anos atrás, o risco do tratamento dos dados pessoais, pois, isolados, não tinham nenhum caráter íntimo. Entretanto, ao submeter um tratamento individualizado e direcionado a determinado indivíduo, possibilita a elaboração de um perfil pessoal e detalhado. Para ilustrar a situação, o autor propôs a Teoria do Mosaico. Ele reuniu pequenas peças isoladas (dados) sem um significado que, ao serem estruturadas (tratadas) de forma esquematizada e organizada, resultou em uma figura (no caso dos dados pessoais) com total sentido, conforme demonstrado na Figura 2:

Figura 2: Mosaico rosa dos ventos

Fonte: Posenato (2016)

3 Existen prioridades desde el punto de vista del derecho a la privacidad y que, sin embargo, conectadas con otros tal vez y relevantes, también pueden servir para hacer totalmente transparente la personalidad de un ciudadano como sucede con pequeñas piedras que forman los mosaicos que si no dicen nada pero unidas forman un conjunto lleno de significados.

Não importa se os dados dizem respeito à privacidade, à intimidade ou ao segredo, mas o uso do que é feito dela. Os produtos de *software* usados para criar e gerir bancos de dados são capazes de criar um perfil do sujeito a partir da coleta dos dados dispersos que só passam a adquirir significado quando reunidos.

A teoria do mosaico é bastante útil para entender e explicar a invasão de privacidade com o uso das novas tecnologias, o que contribui significativamente para se compreender o problema da coleta e do armazenamento de dados pessoais por entidades públicas e privadas. Ante esse panorama, cabe analisar se o direito à intimidade evoluiu e adaptou-se a esse novo desafio, que consiste na coexistência pacífica do uso cada vez mais constante das novas tecnologias e o respeito à intimidade das pessoas.

Brasil, privacidade e proteção de dados

O mundo inteiro passa por uma mudança de paradigma em relação à privacidade e à proteção de dados pessoais. Essa mudança visa não só proteger o cidadão, mas também fomentar uma sociedade e um mercado movidos a dados. Uma lei sobre a proteção de dados possibilita que o cidadão saiba como eles são utilizados por organizações, por empresas e pelo governo. Seu objetivo é de estabelecer padrões mínimos a serem seguidos quando um dado pessoal for usado e como uma finalidade específica, a criação de um ambiente seguro e controlado para seu uso e o de outros, sempre garantindo ao cidadão protagonismo sobre seus dados pessoais e nas decisões fundamentais a respeito da utilização deles. O impacto maior de uma lei que verse sobre a proteção de dados seria possibilitar o equilíbrio das assimetrias de poder sobre a informação pessoal existente entre o titular dos dados pessoais e os que os usam e compartilham.

É instigante quando uma nova temática surge no Direito, mesmo que o tema seja relativamente novo no Brasil, mas já razoavelmente discutido em outros países. É sob esse prisma que a proteção de dados pessoais deve ser analisada. A discussão começou recentemente com, aproximadamente, 40 anos de atraso em relação a países europeus e aos Estados Unidos, que possuem legislação específica desde a década de 70 do Século XX.

Numa era em que a tecnologia se encontra em diversos setores, o proprietário dos dados encontra-se totalmente vulnerável, sem nenhuma certeza de quais dos seus dados foram armazenados, disponibilizados ou acessados sem seu consentimento, ou seja, tratados de alguma forma. Em decorrência desse fato, a sociedade deve se mobilizar e questionar as eventuais práticas abusivas em relação ao tratamento dos seus dados pessoais e discutir sobre o futuro das relações tecnológicas na ambiência digital, ou seja, do próprio ser humano e de seus dados pessoais associados.

Para Stefano Rodotà (2008), o uso de dados é uma fonte de poder para gerar novas situações de poder na sociedade. Os órgãos estatais e as corporações que já detinham poder político ou econômico agora detêm o poder dos dados e, consequentemente, das informações geradas. Se esse controle concentrado de poder não for regulado, poderá gerar desequilíbrios na democracia, pois aprofunda desigualdades e cria dois estamentos: o dos que não controlam os dados e as informações e os dos que as controlam.

No Brasil, a falta de uma lei específica para proteger os dados pessoais causou por muitos anos insegurança jurídica: "Os direitos não nascem todos de uma vez. Nascem quando devem ou podem nascer" (BOBBIO, 1992, p.6). A partir desse contexto, entendemos que a criação da Lei n° 13.709/2018 sobre a questão dos dados pessoais, foi uma ação fundamental para o cidadão brasileiro, para o mercado e para

toda a sociedade. Ao longo do processo de regulamentação do tema, toda sociedade ficou a mercê do comando de uma variedade de diplomas jurídicos, conforme apresentado anteriormente. Antes da promulgação da lei era possível fazer apenas alguns recortes em dispositivos legais distintos. A Lei n° 13.709/2018 trás consigo elementos da realidade da sociedade brasileira e os desafios trazidos essencialmente pela evolução e pelo uso dos dispositivos de tecnologia da informação.

Consequentemente, acompanhando as premências, o Brasil não poderia ser indiferente aos fatos expostos na mídia. O que desde 2013 era o Projeto de Lei nº 330/2013 tornou-se após a sua promulgação a Lei n° 13.709/2018, que inclui questões discutidas pela sociedade civil, governos e empresas há pelo menos seis anos, em um processo que contou com diversas consultas e audiências públicas.

O texto trás diversos conceitos e sujeitos sobre a dinâmica do tratamento de dados pessoais. Como exemplo:

> Art. 5º Para os fins desta Lei considera-se:
> I - **dado pessoal**: informação relacionada à pessoa natural identificada ou identificável;
> II - **dado pessoal sensível**: dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural;
> III - **dado anonimizado:** dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento;
> IV - **banco de dados**: conjunto estruturado de dados pessoais, estabelecido em um ou em vários locais, em suporte eletrônico ou físico;
> V - titular: pessoa natural a quem se referem os dados pessoais que são objetos de tratamento;
> VI - **controlador:** pessoa natural ou jurídica, de direito público ou privado, a quem competem as decisões referentes ao tratamento de dados pessoais;

VII - **operador:** pessoa natural ou jurídica, de direito público ou privado, que realiza o tratamento de dados pessoais em nome do controlador;
VIII - **encarregado:** pessoa natural, indicada pelo controlador, que atua como canal de comunicação entre o controlador e os titulares e a autoridade nacional;
IX - **agentes de tratamento:** o controlador e o operador;
X - **tratamento:** toda operação realizada com dados pessoais, como as que se referem a coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração;
XI - **anonimização**: utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo;
XII - **consentimento:** manifestação livre, informada e inequívoca pela qual o titular concorda com o tratamento de seus dados pessoais para uma finalidade determinada;
XIII - **bloqueio**: suspensão temporária de qualquer operação de tratamento, mediante guarda do dado pessoal ou do banco de dados;
XIV - **eliminação:** exclusão de dado ou de conjunto de dados armazenados em banco de dados, independentemente do procedimento empregado;
XV - **transferência internacional de dados:** transferência de dados pessoais para país estrangeiro ou organismo internacional do qual o país seja membro;
XVI - **uso compartilhado de dados:** comunicação, difusão, transferência internacional, interconexão de dados pessoais ou tratamento compartilhado de bancos de dados pessoais por órgãos e entidades públicos no cumprimento de suas competências legais, ou entre esses e entes privados, reciprocamente, com autorização específica, para uma ou mais modalidades de tratamento permitidas por esses entes públicos, ou entre entes privados;
XVII - **relatório de impacto à proteção de dados pessoais:** documentação do controlador que contém a descrição dos processos de tratamento de dados pessoais que podem gerar riscos às liberdades civis e aos direitos fundamentais,

> bem como medidas, salvaguardas e mecanismos de mitigação de risco;
> XVIII - **órgão de pesquisa:** órgão ou entidade da administração pública direta ou indireta ou pessoa jurídica de direito privado sem fins lucrativos legalmente constituída sob as leis brasileiras, com sede e foro no País, que inclua em sua missão institucional ou em seu objetivo social ou estatutário a pesquisa básica ou aplicada de caráter histórico, científico, tecnológico ou estatístico;
> XIX - **autoridade nacional:** órgão da administração pública indireta responsável por zelar, implementar e fiscalizar o cumprimento desta Lei. (BRASIL, 2018, *online*, **grifos nossos**)

O objetivo precípuo desta lei é estabeler regras de como as empresas e o poder público tratam os dados pessoais, ou seja, como coletam, como armazenam, como comercializam e como fixam limites para que estas ações indicadas ocorram. A lei apresenta os fundamentos que regulam a proteção de dados:

> Art. 2º A disciplina da proteção de dados pessoais tem como fundamentos:
> **I - o respeito à privacidade;**
> II - a autodeterminação informativa;
> III - a liberdade de expressão, de informação, de comunicação e de opinião;
> IV - a inviolabilidade da intimidade, da honra e da imagem;
> V - o desenvolvimento econômico e tecnológico e a inovação;
> VI - a livre iniciativa, a livre concorrência e a defesa do consumidor; e
> VII - os direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais. (BRASIL, 2018, *online*, **grifo nosso**)

Como também os princípios que norteiam as condutas sobre a proteção e o tratamento:

> Art. 6º As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios:
> I - **finalidade:** realização do tratamento para propósitos legítimos, específicos, explícitos e informados ao titular, sem possibilidade de tratamento posterior de forma incompatível com essas finalidades;
> II - **adequação:** compatibilidade do tratamento com as finalidades informadas ao titular, de acordo com o contexto do tratamento;
> III - **necessidade:** limitação do tratamento ao mínimo necessário para a realização de suas finalidades, com abrangência dos dados pertinentes, proporcionais e não excessivos em relação às finalidades do tratamento de dados;
> IV - **livre acesso:** garantia, aos titulares, de consulta facilitada e gratuita sobre a forma e a duração do tratamento, bem como sobre a integralidade de seus dados pessoais;
> V - **qualidade dos dados**: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento;
> VI - **transparência:** garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial;
> VII - **segurança:** utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão;
> VIII - **prevenção:** adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais;
> IX - **não discriminação:** impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos;
> X - **responsabilização e prestação de contas:** demonstração, pelo agente, da adoção de medidas eficazes e capazes de comprovar a observância e o cumprimento das normas de proteção de dados pessoais e, inclusive, da eficácia dessas medidas. (BRASIL, 2018, *online*, **grifos nossos**)

A Lei n° 13.709/2018 foi promulgada em 14 de agosto de 2018, mas visto o período de *vacatio legis*, entrará efetivamente em vigor a

partir de 15 de janeiro de 2020, até lá será necessário se amparar nas leis já existentes no país. Em tempo, as empresas, órgãos públicos e pessoas físicas irão se adequar para acolher o novo ordenamento. Espera-se que a legislação receba a devida importância e atenção, pois é uma lei primordial para a sociedade brasileira, diante desses novos cenários e relações que emergem diante dos produtos e serviços oferecidos pelas tecnologias de informação e comunicação no contexto dos dados.

Referências

BASTOS, C. R.; MARTINS, I. G. *Comentários à Constituição do Brasil.* vol. 2, São Paulo: Saraiva, 1989.

BOBBIO, N.. *A era dos direitos*. Rio de Janeiro: Campus, 1992.

BRASIL. Constituição. Constituição da República Federativa do Brasil. Diário Oficial da União, 5 out. 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 06 mai. 2018

_____. Relatório inicial relativo ao Pacto Internacional dos Direitos Civis e Políticos de 1966. Brasília. 1994.

_____. Lei nº 12.965, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm>. Acesso em: 06 mai. 2018

_____. Lei n°13.709, de 14 de agosto de 2018. Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet). Brasília. 2018. Disponível em: < http://www.planalto.gov.br/CCIVIL_03/_Ato2015-2018/2018/Lei/L13709.htm> Acesso em: 15. ago. 2018.

CONESA, F. M.. Derecho a la intimidad, informática y Estado de Derecho, Universidad de Valencia, Valencia, 1984.

COSTA JR., P. J. da. *O direito de estar só*: tutela penal da intimidade. 4. ed. São Paulo: Revista dos Tribunais, 2007.

DECLARAÇÃO UNIVERSAL DOS DIREITOS HUMANOS. Assembleia Geral das Nações Unidas em Paris. 10 dez. 1948. Disponível em: < https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=por >Acesso em: 06 de mai. 2018.

DÍAZ, E. L.. *El derecho al honor y el derecho a la intimidad*. Madrid. 2013.

DONEDA, D. Privacidade e transparência no aceso à informação pública. Zaragoza: Prensas Universitárias de Zaragoza, 2010. Disponível em: <http://www.egov.ufsc.br/portal/sites/default/files/lefis11-09.pdf>. Acesso em: 06 mai. 2018.

FRANÇA. Código Civil Francês, de 30 de julho de 1994. Disponível em: < https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006070721&idArticle=LEGIARTI000006419288>. Acesso em: 20 nov. 2017

FROTA, H. A. A proteção da vida privada, da intimidade e do segredo no Direito Brasileiro Comparado. *Revista Jurídica Unijus*. V.9, n. 11, Uberaba. 2007.

ORGANIZAÇÃO DOS ESTADOS AMERICANOS, Convenção Americana de Direitos Humanos ("Pacto de San José de Costa Rica"), 1969.

PEREZ LUÑO, A.. *Ensayos de Informática Jurídica*. México: Biblioteca de Ética, Filosofia del Derecho y Política, 1996, p. 35.

RODOTÀ, Stefano. A vida na sociedade da vigilância: a privacidade hoje. Organização, seleção e apresentação de Maria Celina Bodin de Moraes. Tradução: Danilo Doneda e Luciana Cabral Doneda. Rio de Janeiro: Renovar, 2008.

WARREN, S. D. BRANDEIS, L. D. S. The right to privacy. *Harvard Law Review*, Vol. 4, No. 5 (Dec. 15, 1890). Disponível em:<http://www.english.illinois.edu/-people/faculty/debaron/582/582%20readings/right%20to%20privacy.pdf> Acesso em: 06 mai. 2018.

ZAGOL, C.; TIBIRIÇÁ, S. A.. Direito à informação e privacidade: equilíbrio, gestão e conflitos. ETIC - Encontro de Iniciação Científica. Vol. 7, No7. 2011. Disponível em:<http://intertemas.unitoledo.br/revista/index.php/ETIC/article/view/3860/3620> Acesso em: 06 mai.2017.

5

REPOSITÓRIOS DE DADOS CIENTÍFICOS: um panorama teórico-prático

Laerte Pereira da Silva Júnior
Thais Helen do Nascimento Santos

Introdução

O processo de construção do conhecimento científico tem como uma das suas principais características a capacidade de ser comunicável. Marconi e Lakatos (2011) assinalam a comunicabibilidade do conhecimento científico por meio de três aspectos: 1) a linguagem científica deve informar todos os seres humanos; 2) o conhecimento deve ser formulado de modo que outros pesquisadores possam verificar os dados e as hipóteses da investigação, a fim de multiplicar a possibilidade de confirmá-las ou refutá-las; e 3) o conhecimento científico deve ser considerado como propriedade de toda a humanidade, porque "[…] a divulgação do conhecimento é mola propulsora do progresso da Ciência" (ibidem, p. 35).

Uma vez que a comunicação é imprescindível ao fazer científico, surgem iniciativas para ampliar a divulgação de dados e publicações resultantes de investigações e reiterar a responsabilidade social dessa

prática. Essa missão se encontra com o movimento de ciência aberta, cujos pilares são o acesso aberto, os dados abertos, a investigação/inovação aberta, as redes abertas de ciência e a ciência cidadã (MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR, 2016).

No âmbito das iniciativas da ciência aberta, ressalta-se a divulgação não apenas dos resultados, mas também dos dados coletados e sistematizados para a investigação. Sobre os dados científicos, Curty (2015, p. 3, tradução nossa) advoga: "[…] tornaram-se a principal moeda da ciência […]". Isso ocorre em virtude de os dados científicos abertos se desvelarem como elementos que aumentam o nível de transparência, de reprodutibilidade e de eficiência na ciência, o que, consequentemente, beneficia toda a sociedade.

O novo modo de compreender os dados científicos advém da evolução tecnológica, dos movimentos *Open Access* e da ciência aberta (estes últimos interligados entre si, porém distintos em seus objetivos centrais). Os cientistas estão conscientes que é pertinente divulgar os dados científicos e ofertar ferramentas tecnológicas com soluções para melhorar a aquisição, o arquivamento, a manipulação e a transmissão de grandes volumes de dados. Como efeito, assistimos a uma adesão maior dos cientistas às práticas de comunicação, gerenciamento e compartilhamento de dados (CURTY, 2015). Nesse contexto, o objetivo deste estudo é o de identificar e explorar os principais *softwares* para a criação de repositórios e de serviços de dados científicos no prisma da ciência aberta. Para isso, a metodologia aplicada consiste da revisão da literatura e da pesquisa documental em *websites* e portais de *software* para criar repositórios e serviços de dados científicos.

Além desta introdução, o estudo encontra-se estruturado por uma seção teórica dedicada a contornar as convergências entre a comunicação científica, o movimento *Open Access* e a ciência aberta.

Em seguida, abordamos os repositórios de dados científicos, com ênfase em suas características e funções. Na quarta seção, apresentamos a caracterização das soluções de gerenciamento dos dados científicos, sejam iniciativas nacionais ou internacionais. Depois, dispomos as considerações finais, em que indicamos as perspectivas futuras para investigar essa temática.

Ciência aberta: convergência entre a comunicação científica e o movimento *Open Access*

O termo 'comunicação científica' foi cunhado, na década de 1940, pelo físico irlandês e historiador da ciência, John Bernal (CHRISTOVÃO; BRAGA, 1997). Apesar disso, historicamente, a comunicação científica tem como exórdio o período da Guerra Fria, quando se demandava muito tempo buscando informações, que, quando eram encontradas, configurava-se de baixa qualidade. Outra dificuldade se manifestava no atraso da produção científica, devido às barreiras de acesso à informação relevante e adequada para o desenvolvimento das pesquisas. Diante desse cenário, foi necessário ressignificar o valor da informação, entendida, especialmente, como uma estratégia de disputa. E como é um recurso essencial para a produtividade científica, os cientistas precisam obtê-la com rapidez, qualidade e exatidão.

> No pós-guerra, estabeleceu-se o fenômeno da Guerra Fria, o conflito entre os EUA e a URSS, que se estendeu pelos mais variados campos, da influência política às medalhas olímpicas, da ostentação bíblica à corrida espacial. Nesse contexto de competição, o desenvolvimento científico e tecnológico torna-se central, estratégico. E, para o aumento da produtividade e da velocidade de produção de novos conhecimentos científicos, percebeu-se a importância da informação. (ARAÚJO, 2009, p. 198).

Embora o termo tenha uma ampla gama de significados e de interpretações (AUTRAN, 2014), é consenso que, de modo geral, a comunicação científica incorpora os processos de produção e de desenvolvimento da ciência. Caribé (2015) afirma que a comunicação científica pode decorrer de forma interna ou externa. A comunicação interna desvela-se no âmago da comunidade científica, e a externa, na comunidade da educação científica e na popularização da ciência. Todavia, desde que foi criado o conceito ou quando sua necessidade foi reconhecida para estabelecer estratégias científicas e/ou governamentais, a comunidade científica pautou-se em revistas de acesso restrito, ou seja, em revistas pagas. Devido à demanda de informações precisas e de boa qualidade, aplicou-se um custo para o acesso a elas. Consequentemente, como uma parcela de pesquisadores não tinha acesso à produção científica, o sistema de comunicação perdia sua eficiência, ao limitar o reconhecimento e o impacto dos resultados obtidos pelos pesquisadores de diversas instituições pelo mundo.

Na observância dessas implicações, surgiu um movimento que visava democratizar a comunicação dos resultados de estudos e que foi enunciado com a declaração da *Budapeste Open Access Initiative* (BOAI), no ano de 2002. O movimento *Open Access* eclodiu de um novo contexto de comunicação, que ocorre, essencialmente, pela Internet. A antiga tradição prezava a disposição dos cientistas para publicarem seus estudos sem remuneração, enquanto a tecnologia, elemento basilar da nova tradição, possibilitava a distribuição da literatura acadêmica por todo o mundo, de forma gratuita, sem restrições. Assim, para a BOAI (2002, *online*), o acesso aberto à literatura diz respeito à

> […] sua disponibilidade gratuita na internet, permitindo a qualquer usuário ler, baixar, copiar, distribuir, imprimir, buscar ou usar dessa literatura com qualquer propósito le-

> gal, sem nenhuma barreira financeira, legal ou técnica que não o simples acesso à internet. A única limitação quanto à reprodução e distribuição, e o único papel do *copyright* nesse domínio sendo o controle por parte dos autores sobre a integridade de seu trabalho e o direito de ser propriamente reconhecido e citado.

A literatura em acesso aberto é produzida em plataforma digital ou digitalizada. Envolve os artigos publicados em periódicos, avaliados por pares e outras publicações não revisadas que sejam de interesse da comunidade acadêmica. A disponibilização desses materiais não tem custo para o autor ou para os usuários.

A declaração da BOAI (2002) indica, ainda, duas formas de operacionalizar as iniciativas de acesso aberto: pela via verde ou pela via dourada. A via verde envolve o compartilhamento dos materiais por meio dos repositórios institucionais, com o auto-arquivamento, e a via dourada é o acesso por meio de periódicos abertos. Embora tenham características diferentes, ambas as estratégias podem ser aplicadas, de forma integrada, por pesquisadores, acadêmicos, professores, estudantes ou qualquer cidadão interessado em conhecer a produção científica de uma ou mais comunidades. Desfazer as barreiras do acesso à produção acadêmica é o principal objetivo. Como consequência, a difusão do conhecimento fortalecerá a educação e poderá acelerar o desenvolvimento de pesquisas. Em outras palavras, "[…] a disponibilização da literatura científica em acesso aberto é imprescindível a um sistema de comunicação científica para que esse impulsione o progresso da ciência e a sua eficácia" (SILVA JÚNIOR, 2017, p. 32).

As transformações ocorridas nos processos de produção e de comunicação científica mobilizaram não só o movimento *Open Access*. Em um contexto mais amplo, pesquisadores e instituições de fomento à pesquisa de todo o mundo estão refletindo sobre os novos desafios

propostos pela evolução tecnológica na comunicação científica. Um desses novos desafios cristaliza-se com a ciência aberta.

Oliveira e Silva (2016, p. 6) explicam que a ciência aberta "[…] é o fio condutor de investigações científicas apoiadas por uma ciberinfraestrutura tecnológica e metodológica que permite o uso, reúso e reprodutibilidade de dados de pesquisa". Tendo a colaboração como um elemento fundador, a ciência aberta prioriza o gerenciamento correto dos dados de pesquisa, o compartilhamento do conhecimento, o estímulo à inovação, dentre outros. Curty (2015) afirma que a ciência aberta inclui os movimentos de *open data*, *open access* e *open software*. Albagli (2015) corrobora essa assertiva, porém adiciona às iniciativas de ciência aberta o *hardware* científico aberto, os cadernos científicos abertos, a *wikipesquisa*, a ciência cidadã e a educação aberta. Por essa razão, a ciência aberta vai além do acesso aberto aos resultados de pesquisas, porquanto se pauta na abertura de todo o processo científico em razão da responsabilidade social do fazer acadêmico.

O novo modo de pensar e de produzir a ciência de forma aberta apresenta diversas vantagens. O Ministério da Ciência, Tecnologia e Ensino Superior (2016) afirma que são benefícios da ciência aberta: a) aumentar a eficiência da pesquisa; b) aumentar o conhecimento do processo de trabalho científico; c) promover o rigor acadêmico e melhorar a qualidade das pesquisas; d) acelerar a criação de novas temáticas de estudo; e) promover o desenvolvimento da sociedade e da cultura; f) fomentar o impacto econômico e social da ciência; g) valorizar a propriedade intelectual; e h) promover o retorno dos resultados científicos às instituições.

Novas estratégicas no fazer científico requerem uma nova agenda de atividades para se compreender, praticar e divulgar a ciência aberta. Internacionalmente, os pilares da ciência aberta estão bem firmados dentro e fora da Academia: agências de fomento, instituições de pesquisa

e universidade estão alinhadas para a prática da ciência aberta. Para tanto, os projetos de pesquisa submetidos à apreciação para financiamentos já estabelecem estratégias para a abertura de dados, ferramentas científicas, educação aberta etc. Ademais, há um apoio ao desenvolvimento de novas práticas de políticas, diretrizes e estrutura tecnológica. No cenário brasileiro, a ciência aberta ainda se encontra em estágio embrionário. Por essa razão, ainda são escassas as produções acadêmicas sobre o tema e faltam diretrizes e iniciativas práticas (OLIVEIRA; SILVA, 2016).

Um dos focos principais nas práticas de ciência aberta está nos dados científicos, que, como são eixos centrais do processo científico (CURTY, 2015; OLIVEIRA, SILVA, 2016), diferentes ações têm privilegiado seu gerenciamento e compartilhamento de forma correta, isto é, o uso, o reúso e a reprodução dos dados científicos. Além de os pesquisadores terem coletado e usado os dados dentro dos propósitos dos seus estudos, a partilha deles possibilitará que outros pesquisadores os reusem, estimulando o progresso e a inovação científica por meio da economia de tempo, de dinheiro e de esforços de pesquisa. Sobre isso, Curty (2015, p. 3, tradução nossa) complementa:

> A maior amplitude e acessibilidade dos dados de pesquisa é um item fundamental na agenda da ciência aberta, a qual objetiva maximizar o custo-eficácia dos recursos socioeconômicos, melhorar a utilidade e aplicação dos dados, além do foco ou restrições de tempo dos coletores dos dados originais e promover melhor escrutínio e transparência na ciência […].

Nessa perspectiva, dedicamos os próximos capítulos a descrever e a explorar os principais *software* para a criação de repositórios e serviços de dados científicos. Todas as ferramentas seguem a concepção da ciência aberta e incentivam os pesquisadores a armazenarem e a gerenciarem

corretamente os dados para serem usados e funcionam como um meio para compartilhá-los, que envolve o reúso e a reprodução efetivados por outros pesquisadores.

Repositórios de dados científicos: características e funções

> Um repositório de acesso aberto é uma base de dados ou um arquivo virtual criado para coletar, disseminar e preservar a produção científica, como artigos científicos e conjuntos de dados, tornando-os disponíveis livremente. A ação de depositar material em um repositório é denominada de arquivamento (ou autoarquivamento). A depender das preferências pessoais ou das políticas do editor, o autor pode disponibilizar seu trabalho em acesso aberto ou, temporariamente, restringir seu acesso. Um repositório pode fazer parte de uma instituição, departamento, campo de pesquisa ou tema. Nesse sentido, eles são tipificados como repositório institucional ou repositório temático. (OpenAIRE, 2017, *online*, tradução nossa).

Há três tipos de repositórios de acesso aberto: os temáticos ou disciplinares, os institucionais e os de dados científicos. Este último coleta, preserva e compartilha os dados de pesquisa (SDUM, 2017) e é definido como

> um arquivo digital para coletar e expor conjuntos de dados e seus respectivos metadados. Muitos repositórios de dados também aceitam publicações e possibilitam a conexão dessas publicações com os dados subjacentes. Por exemplo: Zenodo, DRYAD, Figshare. (OpenAIRE, 2017b, *online*, tradução nossa).

Nas publicações, os metadados podem ser descritos com precisão pelos bibliotecários. Já sua qualidade para os dados de pesquisa depende da contribuição dos pesquisadores envolvidos

em sua produção. Os pesquisadores, como dominam seu campo de conhecimento, podem contribuir para uma adequada descrição de um conjunto de dados no contexto de sua produção, que possibilita que esses dados sejam reusados por outros estudiosos (AMORIN *et al.*, 2015). Embora os cientistas costumem armazenar e compartilhar seus dados em variados meios, os repositórios de dados são considerados os locais mais adequados para armazená-los e apresentá-los com alta qualidade e disponibilizá-los a um número maior de pessoas. Nesse tipo de repositório, as coleções de dados adquirem uma ampla visibilidade, por isso se pode assegurar que estão prontos para ser reusados (RIN; JISC, 2011).

Existem diversas plataformas disponíveis sob uma licença *open-source*[1] para se construir um repositório de dados científicos, como o *Invenio,* o *DSpace e o Dataverse*. O *Invenio* é um *framework*[2] para bibliotecas digitais que possibilita a criação de um repositório digital com capacidade de armazenamento em larga escala. A tecnologia oferecida por esse *framework* abrange todos os requisitos de gerenciamento de bibliotecas digitais: ingestão, classificação, indexação, recuperação e divulgação dos documentos. Ele é compatível com os padrões *Open Archives Initiative*[3] e com o MARC 21 (Opendata CERN, 2015), além de ser flexível e compatível com variados esquemas de metadados e disponibilizado com uma API completa. Entretanto, a migração dos dados em uma possível desativação de uma plataforma no futuro poderia ser muito difícil, porquanto seu modelo relacional é complexo e muito amarrado ao próprio código do *framework* (AMORIN *et al.*, 2015). A seção *Getting Started* do *website* do *Invenio* informa que ele

1 https://opensource.org/licenses
2 https://pt.wikipedia.org/wiki/Framework
3 https://www.openarchives.org/

é um *framework* modular de pacotes de componentes colaborativos e independentes e que esses pacotes estão localizados em três repositórios do *GitHub*:

> • *no Inveniosoftware*, que é o principal repositório e consiste em uma coleção de pacotes-base e pacotes principais, que são mantidos de forma coerente e homogênea pela equipe de projeto *Invenio*.
> • no *Inveniosoftware-contrib*, uma coleção de pacotes de terceiros que estendem as funcionalidades do *Invenio*. Eles são mantidos por equipes de colaboradores e podem seguir práticas diferenciadas. Os pacotes também podem incubar um recurso experimental ou não comprovado que pode amadurecer mais tarde no repositório principal.
> • no *Inveniosoftware-attic*, uma coleção de pacotes descontinuados ou obsoletos que não são mais mantidos (INVENIO, 2016, online, tradução nossa).

Cinco repositórios construídos com o *Invenio* estão cadastrados no *Registry of Research Data Repositories* (RE3DATA.ORG, 2017): High Energy Physics Database Repository, Zenodo, B2SHARE, CERN Open Data e Inspire-HEP. Na seção de documentação do *website* do *Invenio*, constatamos mais três repositórios com essa tecnologia: *Cern Document Server*, *Caltech Library Catalog* e *DESY Publication Database* (INVENIO, 2015). Apesar dessas poucas instalações do *Invenio*, o *Zenodo* tem se destacado na União Europeia por ter o projeto OpenAIRE[4] como um dos seus financiadores.

O *DSpace* é uma plataforma que foi lançada no ano de 2002 pela *Hewlett-Packard Company*, em parceria com o *Massachusetts Institute of Technology* (MIT) *Libraries*. Foi criada para atender a uma demanda da comunidade do MIT, que precisava gerenciar o armazenamento

4 https://www.openaire.eu/

de publicações acadêmicas e o material de pesquisa produzido em formatos cada vez mais complexos. O sistema *DSpace* contempla todas as funcionalidades de que uma instituição de pesquisa necessita para fazer funcionar um serviço de repositório digital da maneira mais simples possível (SMITH *et al.*, 2003). Esse sistema, predominantemente, tem sido usado como uma plataforma para repositórios institucionais, a fim de se divulgar a produção da literatura científica e, mais recentemente, como repositório de dados científicos.[5] Quando é comparado com o CKAN, com o *Figshare*, com o *Zenodo*, com o *ePrints* e com o EUDAT, o *DSpace* é uma exceção no que concerne ao uso de diferenciados esquemas de metadados. Esses esquemas podem ser adequados para diferentes áreas do conhecimento. Consequentemente, otimiza-se o reúso de dados (AMORIN *et al.*, 2015).

Quanto ao funcionamento, o *DSpace* tem as seguintes características:

1. Um depositante utiliza uma interface *web* para submeter um item, ao qual será associado um *handle*.
2. Os arquivos de dados (*bitstreams*) têm um formato técnico e informações técnicas que serão mantidas com o arquivo de dados para preservação em longo prazo.
3. Um item é uma unidade de arquivamento constituída por um conteúdo agrupado, relacionado e descrito pelos metadados. Os itens são organizados em coleções.

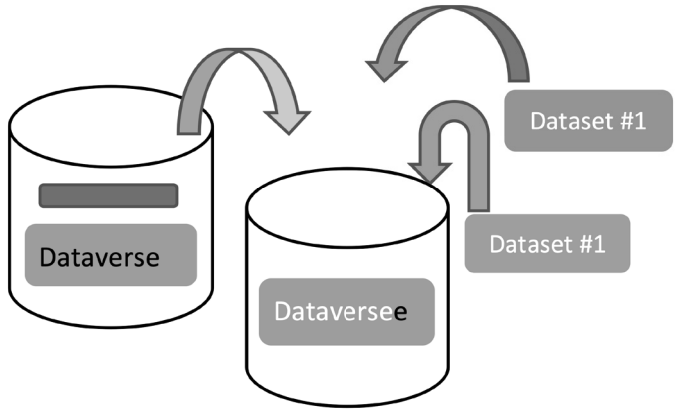
5 Um exemplo de repositório de dados científicos com o sistema *DSpace* é o da Universidade de Cambridge. Disponível em: <https://github.com/inveniosoftware>. Acesso em: 21 nov. 2017.

4. As coleções são organizadas em comunidades, que constituem o nível mais alto na hierarquia do sistema e correspondem às unidades administrativas e acadêmicas.
5. A arquitetura modular do *DSpace* possibilita a criação de um repositório multidisciplinar e de larga escala de armazenamento.
6. O *DSpace* tem um compromisso com a preservação funcional, por isso mantém os itens acessíveis, enquanto as tecnologias de formatos, de mídias e os paradigmas vão evoluindo no transcurso do tempo.
7. A interface de recuperação dos itens possibilita a exibição dos que podem ser exibidos em um *browser*, enquanto os demais podem ser baixados para ser exibidos em uma aplicação específica (DSPACE, 2017).

O *Dataverse* é um *software* de código aberto (*open-source software*[6]) para compartilhar, preservar, citar, explorar e analisar dados de pesquisa. Um repositório *Dataverse* hospeda inúmeros *dataverses*, cada um com variados conjuntos de dados. Esses conjuntos contêm metadados descritivos e o arquivo de dados acompanhado da respectiva documentação e do código. Um *dataverse* pode ser um contêiner para outros *dataverses* (DATAVERSE PROJECT, 2017a). A figura 1 ilustra o esquema de organização de um *dataverse*.

6 Para entender o conceito de plataforma *open source,* consultar, na *Wikipedia*, o verbete *open-source-software*. Disponível em: <https://en.wikipedia.org/wiki/Open-source_software>. Acesso em: 22 nov. 2016.

Figura 1 – Diagrama esquemático do *Dataverse* na versão 4.0



Fonte - Dataverse management (2017): adaptação

O desenvolvimento do *software Dataverse* é liderado pelo *Institute for Quantitative Social Science* da *Harvard University,* que financia o projeto em parceria com a *National Science Foundation*, com o *National Institute for Health* e com outros. O *Dataverse* é uma aplicação *web Java Enterprise Edition.* Os principais componentes da arquitetura do sistema são:

1. Linux: RHEL/CentOS;
2. Glassfish: um servidor de aplicações Java Enterprise Edition;
3. PostgreSQL: um banco de dados relacional;
4. Solr: um mecanismo de busca.
5. Servidor SMTP: usado para enviar *e-mail* para recuperar senhas e outras notificações.
6. Serviço de identificador persistente: suporta o DOI e o Handle. Esse serviço requer um DOI registrado ou uma autoridade Handle.net (DATAVERSE PROJECT, 2017b).

A interface *web* do *Dataverse* é composta de quatro menus (*About, Community, Best Practices, Software*) e um formulário para contato. Dentre esses menus, destacamos os submenus *Getting Started,* que apresenta o serviço para os pesquisadores, os periódicos e as instituições, e o *Guides,* que orienta os usuários, os administradores e os desenvolvedores a operacionalizarem o serviço. A qualidade, a simplicidade e a abrangência das informações encontradas nas referidas opções são um forte estímulo para se adotar essa plataforma.

Os serviços de repositórios de dados científicos abertos de uso geral

Os serviços de repositórios de dados científicos abertos podem ser restritos a uma comunidade de uma instituição (*Apollo - University of Cambridge Repository*) ou ser ofertados a um pesquisador ou instituição de qualquer parte do mundo (*Zenodo, Dryad, Dataverse*).

O *Zenodo* é hospedado no *European Council for Nuclear Research* (CERN) e integra a missão do CERN de divulgar seu trabalho. Um dos financiadores desse repositório é a *European Commission*, através dos projetos OpenAIRE, e é tecnicamente mantido pelo CERN *Data Centre* e pelo *Invenio digital library framework*. Seus metadados e identificadores persistentes são armazenados em uma instância *PostgreSQL.* No que concerne aos aspectos de segurança, o repositório assume os seguintes compromissos:

> Centro de Processamento de Dados do CERN - Nossos centros de processamento de dados estão localizados nas instalações do CERN, e todo o acesso físico é restrito a um número limitado de funcionários com treinamento apropriado e que tiverem obtido acesso de acordo com suas funções profissionais. Por exemplo: a equipe do Zenodo não tem acesso físico ao Centro de Processamento de Dados do CERN.

Servidores - Nossos servidores são gerenciados de acordo com as diretrizes de segurança do CERN. Por exemplo: o acesso remoto aos nossos servidores é restrito ao pessoal treinado do Zenodo. O sistema operacional e os aplicativos instalados são mantidos e atualizados com os últimos *patches* de segurança através do sistema de gerenciamento de configuração automático *Puppet*.

Rede - A equipe de segurança do CERN administra os sistemas de detecção de intrusão baseados em *host* e em rede e monitora o fluxo, o padrão e o conteúdo, dentro e fora das redes do CERN, para detectar ataques. Todo o acesso ao zenodo.org acontece por meio de HTTPS, exceto o acesso a páginas de documentação estática hospedadas no portal GitHub.

Dados de acesso - O Zenodo armazena senhas de usuários usando algoritmos de *hash* na criptografia de senhas fortes (atualmente PBKDF2 + SHA512). Os *tokens* de acesso dos usuários ao GitHub e ao ORCID são armazenados criptograficamente e só podem ser decriptografados com a chave secreta do aplicativo.

Aplicativo - Empregamos um conjunto de técnicas para proteger sua sessão contra furto, quando se estiver conectado ao aplicativo, e executamos varreduras em busca de vulnerabilidades contra o aplicativo.

Staff - A equipe do CERN que pode acessar os dados dos usuários opera sob as regras estabelecidas pela circular CERN Operational Circular no. 5, que, dentre outros significados, determina:

> • O *staff* não deve trocar informações entre si a menos que elas sejam expressamente requeridas para executar as rotinas inerentes ao trabalho.
>
> • O acesso aos dados dos usuários deve ser sempre consistente com as obrigações profissionais e só deve ser permitido para resolver problemas e detectar questões de segurança, monitoramento de recursos e coisas semelhantes.
>
> • Os membros do *Staff* são responsáveis por danos resultantes de qualquer infração e podem ter o acesso retirado e estar sujeito a processos disciplinares ou legais, a depender da gravidade da infração (ZENODO, 2017a, *online*, tradução nossa).

Esses compromissos são fundamentais para conquistar a confiança dos depositantes, porque, nos termos de uso do *Zenodo,* afirma-se o compromisso de preservar o conteúdo educacional, informacional e de pesquisa de todos os depositantes de qualquer parte do mundo, que não tiverem acesso a um *data center* na organização ou instituição a que estiverem vinculados. O armazenamento desses conteúdos é gratuito, e se podem hospedar vários conjuntos de dados de até 50 GB. Embora o CERN não tenha estabelecido um limite para a quantidade desses conjuntos, na seção *FAQ* no *website* do *Zenodo*, é dito que não será possível hospedar uma quantidade demasiadamente grande sem que seja requerida uma doação monetária (ZENODO, 2017b).

O *Dryad Digital Repository* é um repositório que publica dados atrelados à literatura médica e científica internacional. Não tem fins lucrativos e recebe investimentos dos seguintes órgãos: *National Science Foundation* (quadriênio 2016 a 2019), *European Commission* (quadriênio 2015 a 2018) e *Center for Open Science* (desde 2015). Entretanto, os custos com a curadoria e a preservação são sustentados por meio de pagamento dos depositantes dos dados e membros associados ao repositório. Por outro lado, há isenção para os pesquisadores dos países classificados como de baixa renda pelo Banco Mundial.

O *Dryad* visa aos seguintes objetivos por meio dos termos de serviço:

> • aprimorar o registro acadêmico, disponibilizando os dados livremente para serem reusados em pesquisa e em educação;
> • permitir que o conteúdo do repositório seja amplamente indexado e descoberto;
> • atribuir e fornecer identificadores de objetos digitais (*Digital Object Identifier*) para o conteúdo do repositório;
> • rastrear o uso de conteúdo e promover a reutilização de dados como uma métrica de crédito acadêmico;
> • gerenciar embargos de publicação de dados conforme foi estipulado pelas políticas do editor;

> • tornar os dados acessíveis no presente e preservados para serem acessados no futuro, também;
> • respeitar e proteger a privacidade dos usuários;
> • atender à necessidade dos *stakeholders* (as pessoas interessadas no serviço) por um repositório digital confiável (DRYAD, 2015, online, tradução nossa).

O *Dryad* foi criado com o *software* de código aberto *DSpace*. Porém a equipe do repositório customizou o código original e o disponibilizou no repositório *Github*[7]. Para receber artigos e metadados provenientes das editoras, o repositório se comunica por *e-mail* e disponibiliza uma *REST API* para os que desejam ter mais controle sobre o processo de depósito. O *DataCite* é um serviço que fornece o identificador persistente *Digital Object Identifier* (DOI). A licença *Creative Commons* é utilizada para subsidiar o reúso dos dados. O sistema pode ser acessível também por meio de APIs, conforme as instruções disponibilizadas na *wiki*[8] do Dryad (DRYAD, 2016).

No repositório *Harvard Dataverse* (2017), os pesquisadores e as instituições de qualquer parte do mundo podem depositar seus dados. Já os periódicos podem utilizar também o UNC *Dataverse*, o *Scholar Portal Dataverse* e os repositórios encontrados no mapa disponível na *homepage* do projeto *Dataverse*. Todavia, a seção *Getting Started* e os termos de uso[9] do repositório não fazem qualquer menção a limites dos *dataverses* e *datasets*, tampouco a taxas ou doações monetárias. Para saber se haveria algum custo, enviamos uma solicitação para o *e-mail* support@dataverse.org e fomos informados de que o armazenamento de mais de 1TB de

7 Repositório do código do *Dryad*. Disponível em: <https://github.com/datadryad/dryad-repo>. Acesso em: 01 dez. 2017.

8 Wiki do Dryad. Disponível em: <http://wiki.datadryad.org/Data_Access>.

9 Harvard Dataverse Policies. Disponível em: <http://best-practices.dataverse.org/harvard-policies/index.html>.

dados está sujeito a uma cobrança de taxa. Esse repositório foi construído com o *software* de código aberto *Dataverse,* cujas características foram resumidamente descritas na seção anterior deste trabalho, por isso não é preciso customizá-lo, já que foram instalados os componentes principais da arquitetura do sistema.

No Brasil, a comunidade científica das instituições parceiras da Rede de Serviços de Preservação Digital (Rede Cariniana[10]) dispõe gratuitamente do repositório IBICT Dataverse[11], que foi instalado no ano de 2017. Até o dia 16 de novembro desse ano, contava com 23 *dataverses*, 150 conjuntos de dados, 438 arquivos e 435 *downloads*. Entretanto, os termos de uso não haviam sido publicados até a data referida. Outro repositório registrado no mapa do projeto *Dataverse* é o UnBraL Fronteiras[12], do Portal de Acesso Aberto das Universidades Brasileiras sobre Fronteiras e Limites.[13] Porém, ao acessar as opções do menu 'Ferramentas' do portal, em 16 de novembro de 2017, não encontramos qualquer referência ao *Dataverse*. Além disso, o repositório estava inacessível.

Considerações finais

O movimento pela ciência aberta fomenta um novo modo de refletir sobre a ciência e de realizá-la. Esse cenário tem o apoio de diferentes plataformas infocomunicacionais para produzir e fazer circularem a informação e o conhecimento construído no ambiente acadêmico (ALBAGLI, 2015). Portanto, para viabilizar o armazenamento, o uso, o

10 Rede Cariniana. Disponível em: <http://cariniana.ibict.br/>.

11 IBICT Dataverse. Disponível em: <https://repositoriopesquisas.ibict.br/dataverse/ibict>.

12 Repositório UnBraL Fronteiras. Disponível em: <http://unbral.nuvem.ufrgs.br/dvn/>.

13 Portal UnBraL Fronteiras. Disponível em: <http://unbral.nuvem.ufrgs.br/site/>.

reúso e a reprodução dos dados científicos, no prisma da ciência aberta, é preciso utilizar ferramentas adequadas, como os repositórios de dados científicos.

> Os cientistas podem escolher compartilhar e armazenar seus dados de várias maneiras, incluindo métodos mais informais e *ad hoc* (...). No entanto, os repositórios de dados são considerados a melhor opção para garantir o gerenciamento, a visibilidade e a disponibilidade de dados para um grande público (…) (CURTY, 2015, p. 5, tradução nossa).

Neste estudo, ocupamo-nos em caracterizar os repositórios de dados científicos, por meio de suas particularidades, em face das demais tipologias. Asseguramos que a utilização desses repositórios é a estratégia mais adequada para armazenar, gerenciar, usar e reusar os dados científicos. Em termos práticos, os principais *software* para a criação de repositórios e serviços de dados científicos são o Invenio, o DSpace e o Dataverse. A seleção dessas ferramentas tem como base sua primazia sobre as demais. O *Invenio*, por exemplo, refere-se ao *software* do repositório Zenodo, que, por sua vez, é financiado pelo projeto OpenAIRE (o mais importante projeto de acesso aberto da União Europeia); o *DSpace* é o *software* para repositório de acesso aberto mais utilizado em todo o mundo (OpenDOAR, 2017); e o Dataverse, o *software* do repositório de pesquisas do IBICT, que visa atender às instituições que participam da Rede Cariniana.

Assim, alcançamos o objetivo proposto: o de identificar e explorar os principais *software* para a criação de repositórios e serviços de dados científicos no âmago da ciência aberta. Com esses resultados, apresentamos à comunidade acadêmica os atuais e os mais pertinentes *software* que podem ser aplicados no tratamento e no compartilhamento dos dados científicos. Entretanto, conhecer e selecionar uma ferramenta para criar um repositório de dados científicos é o primeiro passo para

o estabelecimento de um programa de gestão de dados científicos. Por essa razão, novas pesquisas sobre essa temática podem voltar-se para os elementos para a constituição de um plano de gestão de dados, por meio das plataformas Mantra[14], DMPonline[15] ou DMPTool[16].

Referências

ALBAGLI, S.. Ciência aberta em questão. In: ALBAGLI, Sarita; MACIEL, Maria Lúcia; ABDO, Alexandre Hannud (Orgs.). *Ciência aberta, questões abertas*. Brasília: IBICT; Rio de Janeiro: UNIRIO, 2015, p. 9-25.

AMORIN, R. C. *et al*. A comparison of research data management platforms: architecture, flexible metadata and interoperability. In: WORLD CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES, s.n., 2015, Açores. *Proceedings*. Açores: WCIST, 2015. Disponível em: <http://dendro.fe.up.pt/blog/wp-content/uploads/2016/11/UAIS2016posprint.pdf>. Acesso em: 20 nov. 2017.

ARAÚJO, C. A. Á. Correntes teóricas da Ciência da Informação. *Ciência da Informação,* Brasília, v. 38, n. 3, p. 192-204, set./dez., 2009.

AUTRAN, M. de M. M.. *Comunicação da ciência, produção científica e rede de colaboração acadêmica:* análise dos programas brasileiros de Pós-graduação em Ciência da Informação. 2014. Tese (Programa Doutoral em Informação e Comunicação em Plataformas Digitais - FLUP) - Porto, 2014.

BUDAPESTE open access initiative. Portuguese translation: iniciativa de Budapeste pelo acesso aberto. Budapest, 2002. Disponível em:

14 Disponível em: <http://mantra.edina.ac.uk>.
15 Disponível em: <https://dmponline.dcc.ac.uk>.
16 Disponível em: <https://dmptool.org>.

<http://www.budapestopenaccessinitiative.org/translations/portuguese-translation>. Acesso em: 20 nov. 2017.

CARIBÉ, R. de C. do V.. Comunicação científica: reflexões sobre o conceito. *Informação & Sociedade: Estudos*, João Pessoa, v. 25, n. 3, p. 89-104, set./dez., 2015.

CHRISTOVÃO, H. T.; BRAGA, G. M.. Ciência da Informação e sociologia do conhecimento científico: a intermaticidade plural. *Transinformação*, Campinas, v. 9, n. 3, p. 33-45, 1997.

CURTY, R. G.. *Beyond "Data Thrifting"*: an investigation of factors influencing research data reuse in the social sciences. 2015. Dissertação (Syracuse University) - Nova Iorque, 2015.

DATAVERSE Project. *About.* S.l., 2017a. Disponível em: <https://dataverse.org/about>. Acesso em: 21 nov. 2017.

_____. *Installation Guide*. S.l., 2017b. Disponível em: <http://guides.dataverse.org/en/latest/installation/index.html>. Acesso em: 22 nov. 2017.

DSPACE. *About DSpace*. S.l., 2017. Disponível em: <http://www.dspace.org/introducing>. Acesso em: 21 nov. 2017.

DRYAD. *About* (Policies). Durham, 2016. Disponível em: <http://datadryad.org/pages/policies>. Acesso em: 23 nov. 2017.

_____. *About* (Repository features and technology). Durham, 2015. Disponível em: <http://datadryad.org/pages/repository>. Acesso em: 23 nov. 2017.

HARVARD Dataverse. *Harvard Dataverse*. Disponível em: <https://dataverse.harvard.edu/>. Acesso em: 23 nov. 2017.

INVENIO. *Getting started with Invenio*. S.l., 2016. Disponível em: <http://invenio-software.org/gettingstarted>. Acesso em: 21 nov. 2017.

_____. *Invenio digital Library framework*. S.l., 2015. Disponível em: <https://invenio.readthedocs.io/en/latest/>. Acesso em: 21 nov. 2017.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. *Metodologia científica*. 6. ed. São Paulo: Atlas, 2011.

MINISTÉRIO da Ciência, Tecnologia e Ensino Superior (Portugal). *Sobre ciência aberta.* Lisboa, 2016. Disponível em: <http://www.ciencia-aberta.pt/sobre-ciencia-aberta>. Acesso em: 20 nov. 2017.

OLIVEIRA, A. C. S. de; SILVA, E. M. da. Ciência aberta: dimensões para um novo fazer científico. *Informação & Informação*, Londrina, v. 21, n. 2, p. 5-39, maio/ago., 2016.

OpenAIRE – Open Access Infrastructure for Research in Europe. *FAQ (What are repositories?*). [S.l.:s.n.], 2017a. Disponível em: <https://www.openaire.eu/support/faq#ifaqCat-21>. Acesso em: 19 nov. 2017.

OPENDATA CERN. *About.* Genebra, 2015. Disponível em: <http://opendata.cern.ch/about>. Acesso em: 21 nov. 2017.

OpenDOAR. *Search a browse for repositories.* Notthinghan, 2017. Disponível em: <http://www.opendoar.org/find.php?format=charts>. Acesso em: 24 nov. 2017.

RE3DATA.ORG – Registry of research data repositories. *Re3data*. Disponível em: <https://www.re3data.org/>. Acesso em: 21 nov. 2017.

RIN – Reserach Information Network; JISC – Joint Information Systems Committee. *Data Centres*: their use, value and impact. Londres, 2011.

Disponível em: <http://www.rin.ac.uk/system/files/attachments/Data_Centres_Report.pdf>. Acesso em: 21 nov. 2017.

SDUM – Serviços de Documentação da Universidade do Minho. *Sobre repositórios OA*. Braga: s.n., 2017. Disponível em: <https://openaccess.sdum.uminho.pt/?page_id=348>. Acesso em 19 nov. 2017.

SILVA JÚNIOR, L. P. da. *Os repositórios institucionais das universidades federais do Brasil*: um modelo de política de preservação digital. 2017. Tese (Programa doutoral em Informação e Comunicação em Plataformas Digitais - FLUP) - Porto, 2017.

Smith, M. *et al*. DSpace: An open source dynamic digital repository. *D-Lib Magazine*, v. 9, n. 1, 2003.

ZENODO. *About – Infrastructure*. Genebra, 2017a. Disponível em: <http://about.zenodo.org/infrastructure/>. Acesso em: 23 nov. 2017.

_____. *Help – FAQ*. Genebra, 2017b. Disponível em: <http://help.zenodo.org/>. Acesso em: 23 nov. 2017.

6

CURADORIA E CICLO DE VIDA DOS DADOS

Sanderli José da Silva Segundo
Wagner Junqueira de Araújo

Introdução

O fenômeno *Big Data*, o qual chamaremos de megadados[1], é considerado uma nova fonte de capital (LAU et al., 2016). Dados são, de certa forma, um tipo comum de "pagamento" cobrado por empresas que atuam na *web*. Por exemplo, o usuário sempre depositará algum dado, consciente ou não, enquanto estiver navegando pela internet. Se ele não deixa "um rastro de dados", é porque não se conectou. Isso pode ser verificado nas redes sociais, cujos serviços destinados a usuários comuns são, geralmente, "gratuitos", quando se trata de pagamentos financeiros. Contudo, é necessário perguntar: até onde essa gratuidade é real? O usuário não sabe, mas existem mecanismos furtivos que capturam seus dados enquanto os sistemas são utilizados. Variáveis como localização, preferência, escolhas, compras ou vendas, investimento de tempo e

1 Aqui utilizaremos o termo 'megadados' ao invés de 'grande dado', porque grande, na língua portuguesa, remete basicamente à quantidade, e mega denota tanto quantidade quanto qualidade, ou seja, não necessariamente quantidade é qualidade.

histórico de acesso e de busca constam sempre na lista de monitoramento. Se essas variáveis forem analisadas adequadamente, poderão subsidiar empresas em suas estratégias de captação, vendas e fidelização. "[...] a ideia básica é a percepção de que praticamente tudo o que a humanidade faz no dia a dia vai gerar um rastro digital, que poderá ser eventualmente analisado" (ANTONIUTTI, 2015, p. 63). Quem nunca foi alvo de alguma propaganda? Basta procurar um simples produto no buscador Google que você visualizará anúncios referentes a ele no *Facebook*, na caixa de entrada do *e-mail* pessoal e em *banners* comerciais espalhados em *sites* de notícia e de compartilhamento de vídeos. Alguém coletou e repassou ou vendeu esses dados para fins comerciais! "Os dados fornecem percepções comportamentais de consumidores, e comerciantes traduzem essas percepções em vantagem de mercado" (EREVELLES; FUNAKA; SWAYNE, 2015, p. 897, tradução nossa[2]). Dados interpretados se transformam em informações estratégicas, articuladas, que têm o potencial de promover o desenvolvimento de produtos e de serviços personalizados e, consequentemente, mais atrativos.

> [...] na última década, permitimos que as máquinas atuassem como intermediárias em quase todos os aspectos de nossa existência. Quando nos comunicamos com amigos, nos divertimos, dirigimos, exercitamos, vamos ao médico ou lemos um livro, temos próximo algum computador transmissor de dados. Deixamos para trás uma vasta nuvem de *bits* e *bytes*. (SEIFE, 2015, p. 481, tradução nossa[3]).

2 Texto original: Data provide behavioral insights about consumers; marketers translate those insights into market advantage (EREVELLES; FUNAKA; SWAYNE, 2015, p. 897).

3 Texto original: In the past decade, we have allowed machines to act as intermediaries in almost every aspect of our existence. When we communicate with friends, entertain ourselves, drive, exercise, go to the doctor, read a book - a computer transmitting data is there. We leave behind a vast cloud of bits and bytes (SEIFE, 2015, p. 481).

Para Azucar, Marengo e Settanni (2018), os rastros, trilhas ou pegadas digitais deixadas pelos internautas fornecem dados que possibilitam prever traços da personalidade deles. Os autores afirmam que "a capacidade de usar pegadas digitais, para prever com precisão as características de personalidade, pode representar uma alternativa rápida e econômica para pesquisas e atingir populações maiores" (AZUCAR; MARENGO; SETTANNI, 2018, p. 151, tradução nossa[4]). Segundo Golder e Macy (2014), a internet é uma espécie de "telescópio social", e os rastros produzidos nesse terreno têm potencial para traçar o perfil comportamental de seus adeptos. "Sites de namoro online proporcionam uma oportunidade sem precedentes para estudar os efeitos das preferências raciais e étnicas nas escolhas de seleção de parceiros amorosos" (GOLDER; MACY, 2014, p. 133, tradução nossa[5]). Mesmo sendo considerado como uma nova fonte para gerar recursos financeiros, o megadados é difícil de ser gerido. E ainda que em nível micro em âmbito organizacional, as empresas geralmente não conseguem aproveitar sua capacidade porque têm formas limitadas de gerenciá-lo (LAU et al., 2016).

A IBM[6] é uma das empresas que apostam no gerenciamento dos dados e, por essa razão, criou, para fins comerciais, uma plataforma destinada a promover esse serviço, o *IBM Big Data*. Seu posicionamento é de que o conhecimento extraído deles é capaz de instruir os funcionários em suas decisões e "[...] aprofunda o engajamento do cliente, otimiza as operações, previne ameaças e fraudes e capitaliza novas fontes de receita"

4 Texto original: The ability to use digital footprints to accurately predict personality traits may represent a rapid, cost-effective alternative to surveys and reach larger populations (AZUCAR; MARENGO; SETTANNI, 2018, p. 51).

5 Texto original: For example, online dating sites provide an unprecedented opportunity to study the effects of racial and ethnic preferences on mate selection choices (GOLDER; MACY, 2014, p. 133).

6 Site: https://www.ibm.com/big-data/

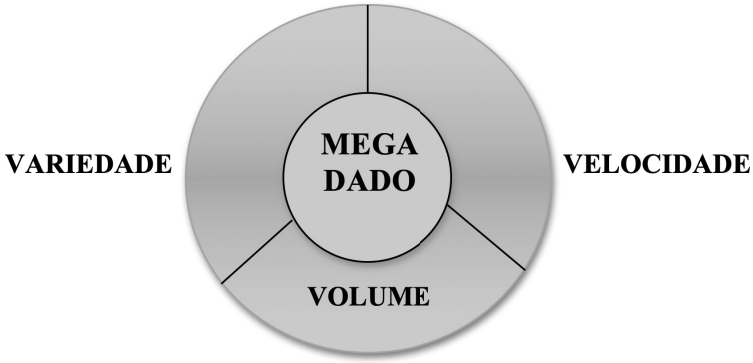
(IBM, 2017, *online*, tradução nossa[7]). Conhecer o comportamento de seus potenciais clientes possibilita saber onde e como atuar, e a Google pratica isso muito bem. Faça um teste! Acesse o computador ou dispositivo móvel que regularmente você utiliza para executar tarefas pessoais. Ao abrir o navegador, procure o termo "restaurante" e perceba que os resultados de busca exibem sugestões de ambientes em sua cidade. E se você for apaixonado por comida japonesa, possivelmente constará nos primeiros lugares do *ranking* algo relativo ao seu gosto culinário.

> [...] praticamente todas as ações cotidianas deixam rastros, e não se trata apenas das informações colhidas a partir das postagens no *Facebook*, no *Instagram* ou das buscas feitas no *Google*. Atitudes banais que passariam despercebidas para a maioria dos indivíduos são, agora, captadas e processadas de forma inteligente, permitindo a identificação de padrões [...] o *smartphone* que mesmo desligado registra a localização do indivíduo, ou o cartão de crédito, que indica as compras realizadas, são apenas alguns exemplos de rastros digitais que abrem caminhos de monitoramento, controle e vigilância (ABREU, 2015, p. 22).

Contudo o megadados ou *big data* tem terminologias que ainda estão em desenvolvimento. O que constitui *big data*, *little data* ou *no data* são seus desdobramento. O potencial a ser extraído deles, sua instrumentalidade (BORGMAN, 2015). Também devem ser considerados os 3Vs que compõem o megadados indicados por Laney (2001): volume, variedade e velocidade:

7 Texto original: […] deepening customer engagement, optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue (IBM, 2017, *online*).

Figura 1 – Os 3Vs do megadados



Fonte: Adaptado de Laney (2001)

Percebemos a dualidade quanti-quali, enquanto o volume aponta para o constante crescimento do megadados, alimentado pela produção ininterrupta de dados por todos os internautas. A velocidade indica a capacidade de processar dados, e a variedade discorre sobre os diferentes tipos, os formatos, as origens e os propósitos.

O aspecto qualitativo dos dados é relativo, porquanto depende de quem necessita deles. Talvez o setor voltado para o *design* mobiliário de uma livraria esteja mais preocupado com características físicas, dimensões e pesos dos livros, enquanto ao marketing cabe conhecer as preferências e as necessidades da clientela. Ao monitorar o comportamento de um grupo animal, os biólogos, possivelmente, chamam de megadados os que têm potencial para definir padrões comportamentais.

Já o quantitativo está diretamente ligado à adoção do termo *big data*, quando a primeira década do Século XXI assistiu à explosão da produção e da coleta de dados pessoais na internet. Por exemplo, ao contratar a publicidade patrocinada oferecida pelo *Facebook*, a empresa

Americanas.com, provavelmente, teve que manipular filtros e definir o público que verá seus anúncios promocionais de artigos infantis, vendidos via *e-commerce*. Quanto mais específicos forem os filtros, maior quantidade de dados será necessária para definir o público-alvo. Portanto, quantidade e qualidade são elementos construtores de megadados.

Há, de fato, um crescimento na produção de dados tanto em ambientes físicos quanto digitais. As diferenças fundamentais residem na forma de acessar, de armazenar, de compartilhar e de proporcionar longevidade, pois, como ainda é um fenômeno extremamente novo, não há parâmetros comparativos entre os dados digitais e os dados físicos, pois estes resistiram por séculos (BORGMAN, 2015).

Ao utilizar um dispositivo do tipo *smartphone*, por exemplo, o indivíduo cria e compartilha textos, áudios, vídeos e imagens, emite dados de seu posicionamento ao sistema de posicionamento global (GPS) e registra o horário em que cada atividade foi executada. O pesquisador Abid Mehmood afirma que "90% dos dados no mundo foram gerados nos últimos anos. O *Facebook* sozinho está gerando 25 *terabytes* de novos dados todos os dias" (MEHMOOD, et al., 2016, p. 1823, tradução nossa[8]). Essa explosão de dados alimenta outra característica dos negócios e da pesquisa científica, que possibilita vislumbrar novas tendências e identificar padrões.

Dados estão por toda parte, e nós produzimos e consumimos sem perceber. Aliás, a ascensão e o consequente barateamento da tecnologia *mobile* proporcionam uma constante conectividade. Quem desliga seu *smartphone* depois de utilizá-lo? Quem se desconecta da internet? Isso implica dizer que nossos dados são coletados continuamente. Mesmo

8 Texto original: 90 percent of the data in the world is generated in the past few years. *Facebook*, a social networking site alone is generating 25TB of new data everyday (MEHMOOD, et al., 2016, p. 1823).

quando transportamos o aparelho no bolso, por exemplo, os aplicativos que usam GPS ilustram bem essa afirmação, porque são capazes de exibir a posição precisa do dispositivo móvel e de seu portador. Estamos sendo vigiados e percebidos, e as redes sociais são canais comuns para essa exposição, porque

> [...] congregam indivíduos que a utilizam como base de apresentação, comunicação e ambiente de relacionamento; classificam-nos em grupos e perfis combinando variáveis diversas; rastreiam deslocamentos coletivos ou individuais de interesse; indexam, a qualquer tempo, qualquer um para entregar-lhe mensagens em geral publicitárias, particularmente significativas (ANTONIUTTI, 2015, p. 99).

A Pesquisa Nacional por Amostras de Domicílios Contínua, divulgada pelo IBGE em fevereiro de 2018, referente ao consumo de internet em 2016, afirma que, no Brasil, há cerca de 179 milhões de pessoas com mais de dez anos de idade, das quais 64% utilizam a internet, aproximadamente 116 milhões de internautas. Desses brasileiros conectados à internet, 109 milhões (94,6%) utilizam dispositivos móveis, enquanto 73 milhões (63%) utilizam computadores *desktop* e *notebooks* (G1, 2018). Há uma clara tendência à inserção de *smartphones* e de *tablets* na rotina dos brasileiros. O uso de dispositivos móveis em ambientes de consumo alimentar, como restaurantes e cafés, subiu nove pontos percentuais entre 2013 e 2014, e passou de 29% para 38%. É crescente o uso em outros ambientes, como escola, trabalho, shopping, carro, transporte público, reuniões com amigos, casa de familiares e na própria residência do usuário (IAB BRASIL, 2014).

Produzimos muitos dados! Mas, quem coleta e gere esse material? E nós, deveríamos nos preocupar com a gestão deles? Dados identificadores tendenciais, comportamentais e de perfis interessam tanto ao sistema comercial quanto à ciência.

Dados e pesquisa científica

Borgman (2015) faz uma distinção entre os termos '*big data*' (megadados), '*little data*' (pequenos dados) e '*no data*' (sem dados). Ela diz que há uma discussão semelhante acerca dos termos '*big science*' (mega ciência) e '*little science*' (pequena ciência). Assim como o termo '*big science*' foi utilizado para dar nome às áreas que empreendiam esforços para descobrir os segredos do universo, '*big data*' se refere aos "tesouros enterrados no fluxo de vida dos bits" (BORGMAN, 2015, p. 21, tradução nossa[9]). Conforme já referido, o vocábulo '*big*' (mega) e empregado aqui para se referir também à qualidade, e não, apenas, à quantidade ou ao volume. E esse valor de ser "mega" equivale a potencial disponibilidade para uso.

Já a mega ciência é caracterizada pelo esforço internacional colaborativo de colégios invisíveis na troca de informações formais ou informais em prol da evolução de determinado campo científico (PRICE, 1963). "São ciências maduras, tipificadas por grandes equipes distribuídas, colaborações internacionais, coletas de megadados e grandes instalações" (BORGMAN, 2015, p. 65, tradução nossa[10]). Analogicamente, respeitando-se as devidas proporções, os megadados, na área administrativa e na governamental, são assim caracterizados por causa dos conhecimentos em grande escala que podem ser extraídos deles (MAYER-SCHONBERGER; CUKIER, 2013 *apud* BORGMAN, 2015).

---

9 Texto original: Just a big science was to reveal the secrets of the universe, big data is expected to reveal the buried treasures in the bit stream of life (BORGMAN, 2015, p. 21).

10 Texto original: Big science, in Price's (1963) terms, are mature sciences typified by large, distributed teams, international collaborations, large data collections, and large instrumentation facilities (BORGMAN, 2015, p. 65).

As comunidades científicas, como qualquer outro segmento que utiliza tecnologias computacionais, produzem megadados constantemente.

Considerando que seria necessário tratar urgentemente a respeito do tema, o Conselho Nacional de Ciência Norte-americano (*National Science Board - NSB*) publicou, em 2005, o relatório intitulado 'Longa Vida das Coleções de Dados Digitais' (*Long-Lived Digital Data Collections*), em que são estabelecidas diretrizes e recomendações para o gerenciamento e o compartilhamento de dados entre a comunidade científica. O relatório é fruto de dois *workshops* realizados para capturar a opinião dos cientistas sobre o ciclo de vida dos dados de pesquisa. Quatro reflexões foram traçadas pelo *NSB*:

Quadro 1 – Reflexões da *NSB* acerca do ciclo de vida dos dados

Nº	Reflexão
1º	As coleções de dados digitais de longa duração são poderosos catalisadores do progresso e da democratização da ciência e da educação. Uma administração adequada da pesquisa requer política eficaz para maximizar seu potencial;
2º	A necessidade de coleções digitais está aumentando rapidamente, impulsionada pelo aumento exponencial do volume de informações digitais. O número de diferentes coleções apoiadas pela Fundação Nacional de Ciência (*National Science Foundation – NSF*) também está aumentando rapidamente. É preciso racionalizar a ação e o investimento nas comunidades científicas e na *NSF*;
3º	O *NSB* e a *NSF* estão posicionados de forma única, para assumir papéis de liderança no desenvolvimento de uma estratégia abrangente para coletas de dados digitais de longa duração, traduzindo essa estratégia em um quadro de políticas consistentes para guiar tais coleções;
4º	As políticas e as estratégias que são desenvolvidas para facilitar a gestão, a preservação e a partilha de dados digitais terão de abraçar completamente a heterogeneidade essencial em aspectos técnicos, científicos e outros encontrados em todo o espectro de coleções de dados digitais.

Fonte: Coletado do relatório *Long-Lived Digital Data Collections*, *NSB* (2015, p. 10, tradução nossa)

Para a ciência, a existência de megadados e de pequenos dados implica dizer que existem situações em que não há dados. Segundo Borgman (2015), os pesquisadores mais recompensados ainda são aqueles que se dedicam a criar novos dados, isto é, trabalham para produzir, e não, para reaproveitar dados. Essa é a situação em que "não há dados disponíveis". Outra ocasião é quando "os dados ainda não foram publicados", ou seja, o pesquisador não pretende, por exemplo, divulgar dados antes que seu trabalho tenha sido aceito para ser publicado, o que é muito comum, ou quando "os dados não são utilizáveis" e, apesar de divulgados, são inacessíveis, porque dependem dos conhecimentos pré-adquiridos e de recursos que os acessem. Seguindo esse raciocínio, a tendência para megadados abertos comprometeria ou colocaria em risco o financiamento de pesquisas.

Teorias, métodos, culturas e perguntas são tão diversos quanto os dados em uma pesquisa científica, porquanto cada campo científico tem diferentes tipos de dados e formas de compartilhá-los e de reutilizá-los e critérios para atribuir autoria e responsabilidade. Essa particularidade exige "uma análise mais cuidadosa das práticas envolvendo os dados, coleta, gestão, uso, interpretação, liberação, reutilização, depósito [...]" (BORGMAN, 2015, p. 63, tradução nossa[11]). Por exemplo, os campos das ciências humanas e das ciências sociais, por sua abordagem interpretativa, tendem a lidar com mais incertezas, e seus pesquisadores, em geral, trabalham sozinhos, fora das cadeias de compartilhamento de dados. Os objetivos da pesquisa científica influenciam a coleta de dados, e uma pesquisa exploratória pode ser feita com um reduzido número de dados. Já

11 A close analysis of data practices-what and how data are collected, managed, used, interpreted, released, reused, deposited, curated, and so on-can identify the range of variance, approaches that are common across fields, approaches that require extensive local adaptation, and opportunities for transfer (BORGMAN, 2015, p. 63).

uma pesquisa descritiva talvez precise contar com um número maior, e as previsões meteorológicas trabalham com dados em curto prazo, enquanto a Meteorologia, ciência do clima, estuda longos períodos climáticos (BORGMAN, 2015).

A quantidade e o tipo variam conforme os objetivos da pesquisa e divergem em função dos instrumentos de coleta. Os dados podem ser coletados tanto por máquinas quanto por pessoas. Contudo, este último requer do pesquisador mais conhecimento para manipular os instrumentos e os métodos de coleta. Por isso, outra variável é o sujeito da coleta, a perspectiva, seu grau de controle/domínio sobre os dados, condicionados aos conhecimentos prévios, que são apreendidos ou programados por meio da linguagem de computador. "A viagem de campo de um dia pode produzir dados suficientes para vários artigos. Em outros casos, um artigo de jornal pode representar muitos anos de coleta de dados" (BORGMAN, 2015, p. 67, tradução nossa[12]).

A fase de análise dos dados também resguarda diferentes performances, que dependem da clareza, da consistência ou da fragilidade dos dados e de hábitos dos pesquisadores responsáveis. "Enquanto alguns organizam os dados de forma rápida e discreta o mais rápido possível, outros acumulam por dias, meses ou anos, antes de abordar a pilha" (BORGMAN, 2015, p. 68, tradução nossa[13]).

Mas afinal, quando existem dados ou onde existem dados? Para Borgman (2015, p. 69, tradução nossa[14]), os dados existem quando o

12 Texto original: A one-day field trip may yield enough data for several papers. In other cases, one journal article may represent many years of data collection (BORGMAN, 2015, p. 67).

13 Texto original: Some organize their data quickly, discretely, and fully as early as possible. Others let data accumulate for days, months, or years before tackling the pile (BORGMAN, 2015, p. 68).

14 Rarely can a magic moment be established when things become data. Typically it

pesquisador "reconhece que uma observação, objeto, registro ou outra entidade poderia ser usada como evidência de fenômeno e depois coleta, adquire, representa, analisa e interpreta essas entidades como dados". Diversos dados podem ser coletados por intermédio de diferentes instrumentos e métodos em um mesmo campo de pesquisa. Recursos como câmeras, microfones e outros sensores que mapeiem padrões de movimento, efeitos físicos e biológicos conseguem captar grandes quantidades de dados. Além disso, figura a perspectiva do observador, ou observadores, e suas condicionantes, como conhecimentos tácitos, ideologias, costumes e habilidade para transcrever os acontecimentos observados, dentre outras variáveis.

Ressalte-se, porém, que tão importante quanto a fase da coleta é a descrição, pois a performance na utilização ou na reutilização será maximizada se a descrição estiver presente em todo o ciclo de vida dos dados. "Os metadados e a documentação adicional facilitam a descoberta de dados através de ferramentas de busca e informam outros pesquisadores quanto ao formato, variáveis, fonte, metodologia e a análise aplicada aos dados" (CROSAS, et al., 2015, p. 265, tradução nossa[15]). Assim, considerando os complexos formatos, os tipos e a natureza dos dados, os descritores ou metadados e os sistemas de classificação são primordiais para as políticas de compartilhamento, porque proporcionam condições de recuperação.

---

involves a process in which a scholar recognizes that an observation, object, record, or other entity could be used as evidence of phenomena and then collects, acquires, represents, analyzes, and interprets those entities as data (BORGMAN, 2015, p. 69).

15 Texto original: metadata and additional documentation facilitates data discovery through search tools, and informs other researchers about the format, variables, source, methodology, and analysis applied to the data (CROSAS, et al., 2015, p. 265).

Metadados

Os metadados estão ao nosso redor. É assim que Pomerantz (2015) introduz seu livro e cita o caso de Edward Snowden, em que ficou evidenciado que os Estados Unidos da América mantêm um programa de captura de metadados na *National Security Agency* (NSA). Um grande montante de metadados está ligado aos telefonemas, como o número de quem disca e de quem recebe, o tempo, a localização, dentre outros. Isso tem gerado desconforto entre os ativistas pró-privacidade. Metadados são elementos descritores, quer dizer, seu papel é de descrever dados. "Na era moderna da computação ubíqua, metadados tornou-se uma infraestrutura, assim como a rede elétrica e o sistema rodoviário" (POMERANTZ, 2015, p. 09, tradução nossa[16]). O mundo está estruturado a partir de um alicerce de metadados! Descritores estão por toda parte, vitrines, cardápios, catálogos, fichas técnicas, tabelas nutricionais, análises financeiras etc.

> [...] metadados explicitam os diferentes aspectos do recurso que descreve: sua estrutura, conteúdo, qualidade, contexto, origem, propriedade e condição. E auxiliam na organização, favorecem a interatividade, validam as identificações e asseguram a preservação e principalmente, otimizam o fluxo informacional melhorando o acesso aos dados. (SANTOS; SIMIONATO; ARAKAKI, 2014, p. 150).

Pomerantz (2015) classifica os metadados como descritivos – os que descrevem detalhadamente o objeto; administrativos – que fornecem informações sobre a origem e os requisitos de manutenção; estruturais

16 Texto original: In the modern era of ubiquitous computing, metadata has become infrastructural, like the electrical grid or the highway system (POMERANTZ, 2015, p. 09).

– que indicam como o objeto está organizado; e de preservação – que orientam as etapas que tratam da segurança e da preservação.

Para o autor, os metadados também fornecem dados que falam que o modo de usar "o editor de um livro eletrônico pode controlar quantas transferências o livro recebeu, em que datas e quais páginas foram baixadas no perfil do usuário" (POMERANTZ, 2015, p. 14, tradução nossa[17]). Conforme Lima, Santos e Santarém Segundo (2016, p. 52), os metadados são vitais para se entender bem "o recurso armazenado, pois descrevem informações semânticas e sintáticas sobre o recurso e podem ser comparados com um sistema de rotulagem".

A questão crucial para esse tema é: como atestar a confiabilidade dos metadados, já que eles afirmam algo a respeito de alguém ou de algum objeto? É aí em que entram os metadados de proveniência. O vocábulo 'proveniência', ou procedência, significa a origem de algo. Neste estudo, entendemos esse termo como a descrição do histórico referente à criação e à edição de algum dado digital. Para Borgman (2015, p. 76, tradução nossa[18]), "proveniência na *world wide web* inclui aspectos, tais como as atribuições de um objeto, quem são responsáveis, sua origem, os processos relativos a ele ao longo do tempo e seu controle de versões".

A qualidade dos descritores de proveniência pode influenciar o nível de confiabilidade e de credibilidade do dado, ao registrar de quem são os créditos (BORGMAN, 2015). Os metadados de proveniência, de acordo com Pomerantz (2015), é uma das subcategorias que constituem os metadados

17 Texto original: Finally, use metadata provides information about how an object has been used: for example, the publisher of an electronic book might track how many downloads the book has received, on what dates, and profile data about the users who downloaded it (POMERANTZ, 2015, p. 14).

18 Texto original: Provenance on the World Wide Web includes aspects such as the attribution of an object, who takes responsibility for it, its origin, processes applied to the object over time, and version control (BORGMAN, 2015, p. 76).

administrativos e informam a origem dos dados, quem os criou e editou, para facilitar a identificação e a consequente distinção entre as cópias e o original. "Metadados de proveniência é uma maneira de situar um recurso/objeto em uma rede social, para fornecer o contexto que o usuário precisa para avaliar o recurso/objeto" (POMERANTZ, 2015, p. 45, tradução nossa[19]).

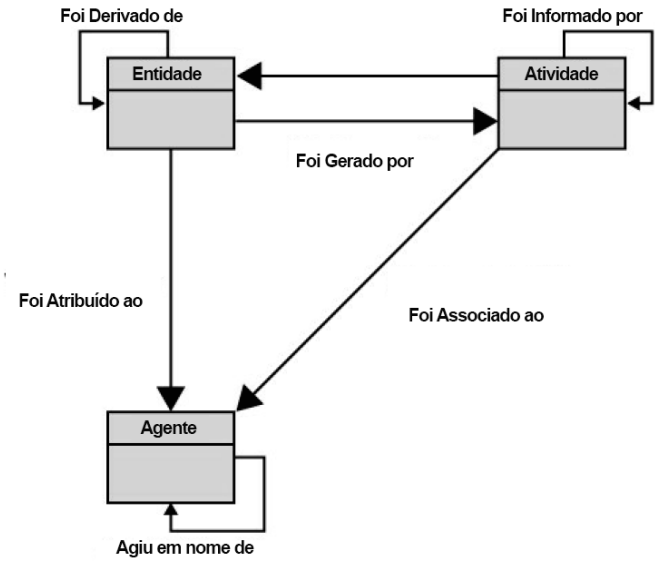
Estabelecido o contexto, o pesquisador conseguirá, entre outros aspectos, discernir o valor dos dados estudados para determinado grupo. Por exemplo, um texto de Fernando Pessoa tem mais valor cultural para países de língua portuguesa, sobretudo Portugal, que é a pátria desse escritor. Quando a proveniência é desconhecida, os resultados analíticos que envolvem um conjunto de dados têm valor questionável (CROSAS, et al., 2015). Uma das características mais evidentes do arquivo digital é sua potencial capacidade de ser duplicado sem perder elementos e de criar cópias idênticas ao original. Algo não tão fácil se o arquivo, item ou objeto fosse físico, pois as tecnologias empregadas na duplicação de objetos físicos são caras e, geralmente, disponíveis em indústrias e empreendimentos comerciais. Criar cópias idênticas, no âmbito físico, demanda altos custos relativos ao armazenamento da cópia.

Diferente do físico, a função 'cópia', no digital, é acessível a todos os níveis de usuários de computador, particular ou comercial. Por isso, os metadados que apontam a origem são importantes no mundo digital. Porém, saber a história e a origem de um objeto digital não é suficiente quando estão em jogo a validade e a confiabilidade dos dados, pois "proveniência não significa apenas a história de um recurso, mas as relações entre esse recurso e outras entidades que influenciaram sua história" (POMERANTZ, 2015, p. 45). Os metadados de proveniência (Figura 2) precisam deixar claro quem participou

19 Texto original: In short, provenance metadata is a way of situating a resource in a social network, to provide context that a user might need to evaluate a resource (POMERANTZ, 2015, p. 45).

dessa história, quais os agentes responsáveis por eventuais modificações dos dados, os indivíduos que tiveram acesso aos dados e os possíveis efeitos de suas interações. Essas descrições são relevantes para controlar as versões. Os padrões de metadados de proveniência "categorizam as relações entre recursos e entidades [...] as três estruturas fundamentais nesses modelos de metadados são entidade, agente e atividade" (POMERANTZ, 2015, p. 45, tradução nossa[20]). Ou seja, não basta conhecer o agente originário, o intermediário ou o final, mas também quais as atividades que foram exercidas por eles durante o fluxo de interação com os dados.

Figura 2: Estrutura básica dos metadados de proveniência



Fonte: Pomerantz (2015, tradução nossa)

20 Texto original: The three "core structures" in this data model are entity, agent, and activity, consistent with the W3C Provenance Incubator Group's definition: an entity is a resource, an agent is an entity that has influenced the life cycle of that resource, and an activity is the nature of that influence (POMERANTZ, 2015, p. 45)

A entidade é o recurso, o objeto, o dado. O agente é o ator, o indivíduo, a instituição que influenciou o ciclo de vida dos dados. A atividade é a descrição dessa influência sofrida. Borgman (2015) afirma que é possível verificar, nos registros de proveniência, como os dados foram coletados, as alterações sofridas durante sua manipulação, os critérios adotados para limpá-los e sintetizá-los, os produtos de *software* utilizados, as rotinas para manipulá-los, as informações necessárias para replicá-los ou interpretá-los. Também se podem analisar os níveis de relevância entre os agentes e os dados em questão. "Quanto mais distante o pesquisador é da fonte original dos dados, maior será a necessidade de informações sobre proveniência para a reutilização dos dados" (BORGMAN, 2015, p. 76, tradução nossa[21]). Os pesquisadores precisam confiar nos dados que alimentam seus projetos, por isso é necessário transparência quanto às pessoas, aos instrumentos de coleta e aos softwares que, de alguma forma, tocaram, influenciaram, modificaram ou manipularam os dados ou seus fluxos (BORGMAN, 2015).

Borgman (2015) enuncia que, quanto mais detalhada for a proveniência acerca de um indivíduo e os métodos que emprega para atuar, maior também será a dificuldade de liberar o conjunto de dados gerados por essas interações, o que impede que uma eventual reutilização seja autorizada. Por exemplo, se o objeto da pesquisa é uma postagem polêmica, que circula em uma rede social, a proveniência detalhada descreverá a origem, o produtor da postagem, seus dados pessoais, as condições e os métodos de postagem, os receptores que interagiram diretamente com o conteúdo e manipularam-no, modificaram-no ou o compartilharam efetuando cópias. O horário e a localização referentes

21 Texto original: The more distant the researcher is from the original source of data, the more that reuse depends upon the availability of provenance information (POMERANTZ, 2015, p. 76).

à produção da postagem original e à efetivação das cópias em forma de compartilhamentos são informações potencialmente coletáveis. Talvez, essas descrições particulares de atores investigados e seus dados pessoais resultem em uma barreira para o livre compartilhamento dos dados e sua irrestrita reutilização.

Todo esse movimento de criação, rastreamento, coleta e processamento de dados e metadados gera outras necessidades, como manter, preservar e disponibilizar tanto os dados quanto seus metadados para uso futuro.

Curadoria de dados

Alguns movimentos impulsionaram o apelo ao livre acesso a dados, entre eles, o *software* livre, há anos alardeado por Richard Stallman[22], embora, bem antes dessa recente militância, tenha sido fundado em 1950 o primeiro *Data Center* mundial, com o intuito de arquivar e distribuir dados (SHAPLEY; HART, 1982; KORSMO, 2010 *apud* BORGMAN, 2015) e, em 1966, tenha surgido a CODATA[23] (*Committee on Data for Science and Technology*), criada pelo *International Council for Science*, para promover a cooperação em gestão e utilização de dados (LITE; WOOD, 2012 *apud* BORGMAN, 2015). Essas iniciativas não só fizeram germinar

22 Sobre Richard Stallman: Richard Matthew Stallman, ou simplesmente "RMS" (Manhattan, 16 de março de 1953), é um ativista, fundador do movimento software livre, do projeto GNU e da FSF. Um aclamado programador e hacker, cujos maiores feitos incluem Emacs (e o GNU Emacs, mais tarde), o GNU Compiler Collection e o GNU Debugger. É também autor da GNU General Public License (GNU GPL ou GPL), a licença livre mais usada no mundo, que consolidou o conceito de copyleft. Desde a metade dos anos 1990, Stallman tem dedicado a maior parte de seu tempo ao ativismo político, defendendo *software* livre. Fonte: Wikipédia.com

23 Site: https://www.codata.org/

políticas relativas aos acessos livres, como também promoveram tensões entre interesses, pois existem dados de pesquisa acadêmica que têm valor comercial e os provenientes das atividades comerciais com potencial valor para investigações acadêmicas.

Problemas ainda maiores são enfrentados por dados digitais, porque o aparato tecnológico focado na coleta, na manipulação, na circulação e no compartilhamento dos dados contrasta com suas vulnerabilidades, que não os resguardam de todos os possíveis riscos concernentes à integridade, à sabotagem, ao plágio e aos demais ataques à propriedade. Fora isso, consecutivas versões de *hardware* e *software*, periodicamente lançadas no mercado, comprometem a leitura de dados antigos e geram incompatibilidade de extensão e consequente perda de dados.

Por isso, considerando que as tecnologias da informação e da comunicação possibilitam ampla produção, acesso e distribuição de dados digitais, gerir dados, objetivando acesso livre, é mais do que simplesmente adotar qualquer plano de gestão e instalar um *software* de processamento e de análise. O gestor, aqui tratado como curador, tomará decisões baseadas no ciclo de vida dos dados digitais, objetivando utilizá-lo adequadamente ao longo dos anos, para prevenir perdas e contaminações.

São essenciais, por exemplo, ferramentas de controle de qualidade que verifiquem a validade dos resultados das pesquisas e analisem a consistência e a integridade de seus dados. Na perspectiva *e-Science*, gerir dados é prepará-los para serem compartilhados e reusados em possíveis replicações, integrações ou novas abordagens.

> A expansão do conceito de acesso livre, incorporando agora coleções de dados de pesquisa, vem se consolidando amparada por várias ações cultivadas no próprio seio das comunidades científicas, que reconhecem esses estoques de informação como uma parte do patrimônio da ciência universal e um pilar imprescindível para o seu avanço. (SAYÃO; SALES, 2012, p. 181).

Afinal, quais os benefícios advindos do gerenciamento e do compartilhamento dos dados de pesquisa científica? De acordo com Hesse, Moser e Riley (2015), um dos proveitos do livre acesso aos dados consiste em aumentar o rigor e potencializar a reprodutibilidade das pesquisas mediante uma ciência transparente e colaborativa, para acelerar a criação de novos conhecimentos.

Para Giambrone et al. (2015), o uso de megadados na área de saúde favorece a identificação de doenças ainda em seus primeiros indícios, melhora a gestão e facilita o reconhecimento de fraudes no sistema de saúde. Segundo os autores, geraria para o sistema de saúde norte-americano uma economia de, aproximadamente, 300 bilhões de dólares por ano, ao evitar desperdícios durante os processos de investigação científica e desenvolvimento de equipamentos, como também ao evitar gastos na construção de políticas e operações clínicas desnecessárias.

A iniciativa norte-americana *DataOne*[24] (2016) diz que há vantagens decorrentes que atingem diretamente a comunidade, o patrocinador ou financiador da pesquisa científica e a própria comunidade científica. Na comunidade, o compartilhamento de dados surtiria efeito parecido com o apontado pela IBM (2017) - de subsidiar o processo decisório e de possibilitar respostas ágeis para os temas complexos como segurança, saúde e educação. Portanto, serviria de insumo na elaboração das políticas públicas.

Para o financiador da pesquisa, a disseminação dos dados reduziria gastos com duplicidade, como disseram Giambrone et al. (2015), e evitaria o financiamento de duas equipes de pesquisa produtoras de dados similares. Essa economia possibilitaria novas pesquisas, contribuiria para o avanço e os propósitos do órgão patrocinador e aumentaria a transparência referente aos gastos, às alocações de recursos e aos investimentos.

24 Site: https://www.dataone.org/

Assim, a comunidade científica seria contemplada com os seguintes benefícios: a expansão da compreensão fenomenológica; a construção de ciência a partir do reuso de dados já coletados; a diminuição de esforços duplos em coletas de dados; a possibilidade de comparar e de reproduzir resultados e a criação de uma relevante rede de colaboração entre cientistas. Sobre isso, Hesse, Moser e Riley (2015) asseveram:

> [...] quando um pesquisador deposita seus dados brutos, ele abre a possibilidade dos seus pares replicá-los e, dessa forma, verificar o que está sendo defendido na publicação científica; isto possibilita também que outros pesquisadores reusem os dados, os comparem e os combinem com outros dados, de forma que novas pesquisas podem ser geradas. (SAYÃO; SALES, 2012, p. 183).

Só uma gestão adequada dos dados digitais proporciona as condições necessárias para reduzir os riscos de integridade, perda ou inacessibilidade dos dados. Porém, existem diferentes formatos de dados, centenas de extensões, milhares de intencionalidades, várias formas de reprodutibilidade e de adaptabilidade e divergentes níveis de processamento. Por isso, a curadoria assegura a manutenção dos dados em curto prazo e sua preservação em longo prazo. A *US National Science Board*[25] (*NSB*, 2005) divide os dados em três categorias: observacional, computacional e experimental. Os dados observacionais são os extraídos de observações, de gravações de fatos ou de ocorrências de fenômenos, como, por exemplo, "observações diretas da temperatura do oceano em uma data específica, atitude dos eleitores antes de uma eleição ou fotografias de uma supernova" (*NSB*, 2005, p. 19, tradução nossa[26]); computacionais são os

25 Site: https://www.nsf.gov/nsb/

26 Texto original: Observational data, such as direct observations of ocean temperature on a specific date, the attitude of voters before an election, or photographs of a supernova (NBS, 2005, p. 19).

que necessitam de sistema computacional para ser executados, "resultados da execução de modelos ou simulações computacionais" (*NSB*, 2005, p. 19, tradução nossa[27]), e os experimentais, resultados da aplicação de procedimentos em condições controladas, "como mediações de padrões de expressão gênica, taxas de reações químicas ou o desempenho complexo de um motor" (*NSB*, 2005, p. 19, tradução nossa[28]) "para testar ou estabelecer hipóteses ou descobrir ou testar novas leis."

Borgman (2015) se encarregou de adicionar uma quarta categoria, a dos 'registros', em que são incluídos os dados que não se encaixam adequadamente nas três categorias anteriores, válidos para registrar fenômenos ou atividades humanas, "documentação de governo, empresas públicas e atividades privadas, livros e outros textos, materiais de arquivo, documentação sob a forma de registro de áudio e vídeo, vidro, placas [...]" (BORGMAN, 2015, p. 39, tradução nossa[29]).

A curadoria envolve o gerenciamento, desde a origem, assegura a recuperação, o compartilhamento e a possível reutilização dos dados, portanto, ajuda a maximizar seu potencial (LORD; MACDONALD, 2003). Sayão e Sales (2013, p. 4) afirmam que "a curadoria de dados de pesquisa permite que os dados possam ser tratados, arquivados em ambienteis digitais confiáveis, preservados e reconfigurados de forma que possam ser aplicados em novos contextos científicos.". Segundo Ball

27 Texto original: results from executing a computer model or simulation (NBS, 2005, p. 19).

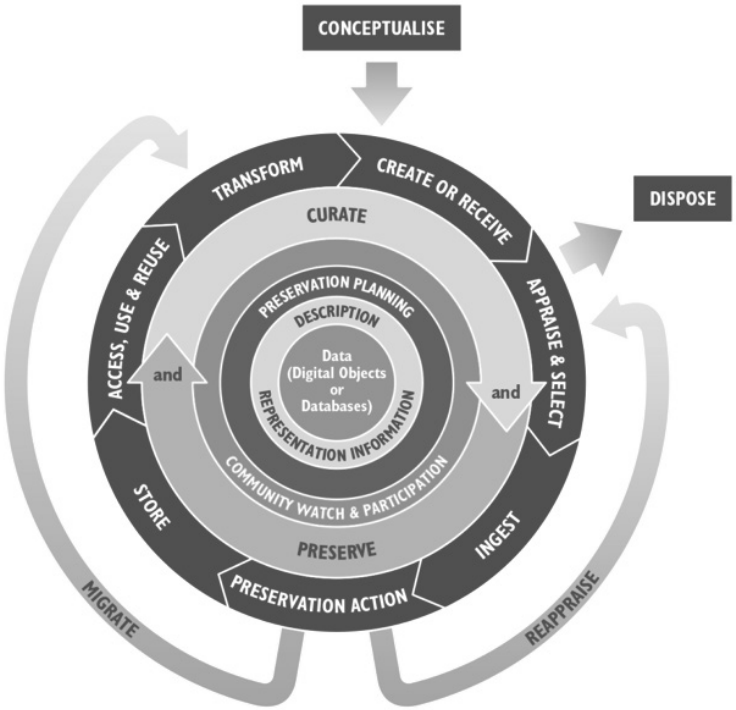
28 Texto original: Experimental data such as measurements of patterns of gene expression, chemical reaction rates, or engine performance present a more complex picture (NBS, 2005, p. 19).

29 Texto original: They can include documentation of government, business, public, and private activities; books and other texts; archival materials; documentation in the form of audio and video recordings, glass plates, papyri, cuneiforms, bamboo, and so on (BORGMAN, 2015, p. 39).

(2010), os ciclos de vida de dados digitais padronizam o processo de curadoria digital.

O *Digital Curation Centre* – DCC, criado em 2004 pela *University of Edinburgh*, em parceria com o *UK Office for Library and Information Networking* - UKOLN (*University of Bath*) e com o *Humanities Advanced Technology and Information Institute* - HATII (*University of Glasgow*), desenvolveu um modelo de ciclo de vida, que propicia uma visão gráfica dos estágios exigidos para que os dados sejam preservados e cuidados adequadamente, conforme ilustra a Figura 3:

Figura 3: Ciclo de vida de dados digitais do DCC



Fonte: DCC Curation Lifecycle Model (2018)

Esse modelo funciona através de camadas, e os estágios são iniciados a partir do núcleo, em que estão dispostos os dados. *Description Representation Information* (Descrição e Representação da Informação) é o primeiro estágio, em que são atribuídos metadados administrativos, descritivos, técnicos, estruturais e de preservação. O segundo, *Preservation Planning*, consiste em planejar e em preservar. Nele, são elaborados planos estratégicos que nortearão e gerenciarão todas as etapas do ciclo de vida. O terceiro, *Community Watch & Participation* (Vigilância e Participação Comunitária), tem o objetivo de manter informada a gestão da curadoria sobre as mudanças ocorridas na comunidade geradora dos dados e de fomentar o desenvolvimento de padrões, ferramentas e softwares adequados para acompanhar essas modificações. O quarto estágio - *Curate and Preserve* (Cure e Preserve) - abrange ações gerenciais, planejadas para favorecer a curadoria e a preservação durante o ciclo de vida (BALL, 2010; DIGITAL CURATION CENTRE, 2018).

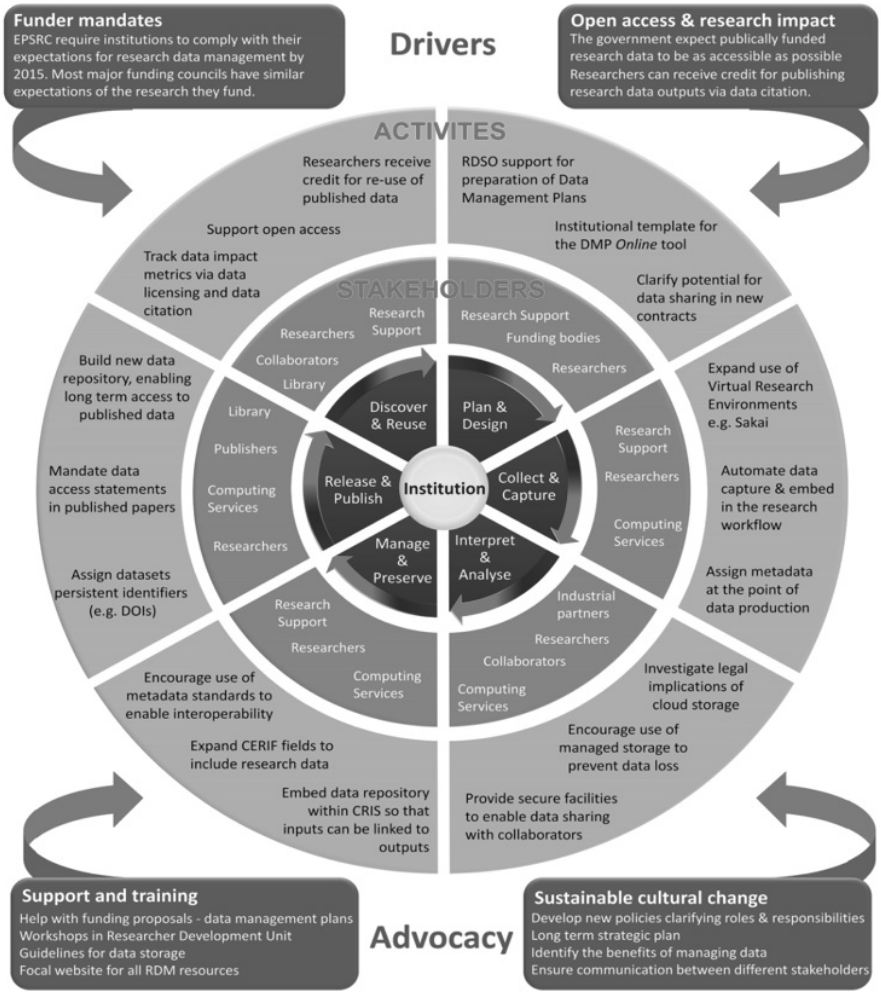
Em volta dos quatro estágios e do núcleo, encontram-se as 'Ações sequenciais'. O ciclo de vida inicia em *Conceptualise* (Conceituação), ato de planejar a criação dos dados, seu método de captura, opções de armazenamento e automatizações e simplificações de etapas. Também aborda questões orçamentárias para garantir recursos financeiros e evitar que o processo seja interrompido. Logo depois, vem o *Create or receive* (Criar ou receber), em que o curador gera dados e seus metadados ou os recebe de outro indivíduo ou repositório, e, em seguida, o *Appraise & select* (Avaliar e selecionar), em que será avaliada a relevância dos dados, conforme os objetivos da curadoria. Portanto, eventualmente, alguns dados serão descartados ou transferidos para outros curadores na fase *Ingest* (Ingerir). Depois de avaliados, selecionados e/ou descartados ou transferidos, iniciam-

se as *Preservation action* (Ações de preservação), que garantem a confiabilidade, a autenticidade e a usabilidade dos dados. As ações incluem limpeza de dados, validação, atribuição de metadados de preservação, atribuição de informações de representação e estrutura e compatibilização com os formatos adequados. Terminadas tais análises e aplicações, vem a fase *Store* (Armazenamento). Os dados precisam ser armazenados de forma segura, seguindo os padrões pré-estabelecidos, em mídias confiáveis e atualizadas. Além disso, é necessário produzir *backups* regulares e garantir sua integridade. Uma vez armazenados, o processo entra na fase de *Access, user & reuse* (Acesso, uso e reutilização). As ações aplicadas nessa etapa têm a finalidade de assegurar que os dados armazenados sejam encontráveis e acessíveis aos usuários que pretendem reutilizá-los. A última ação sequencial é a *Transform* (Transforme), que se fundamenta na produção de novos dados a partir dos originais. Isso ocorre quando, por exemplo, o hardware e o software tornam-se obsoletos e é necessário migrar os dados para um novo formato, ou quando subconjuntos dos dados são criados para criar resultados derivados (BALL, 2010; DIGITAL CURATION CENTRE, 2018).

Outro modelo de ciclo de vida de dados digitais, ligado ao UKOLN da *University of Bath* e ao DCC da *University of Edinburgh*, foi originado do Projeto *Research360*[30], cuja finalidade foi de construir mecanismos que possibilitassem que o gerenciamento de dados fosse incorporado a todo o ciclo de vida da pesquisa científica (MCKEN; et al., 2012). A Figura 4 apresenta o modelo proposto:

30 Sobre o projeto: http://www.ukoln.ac.uk/projects/research360/

Figura 4: Ciclo de Vida do Research360 Project



Fonte: Mcken et al. (2012)

Nesse modelo, o principal objetivo é de envolver a instituição durante o gerenciamento dos dados. Por isso, observa-se que o elemento *Instituition* aparece no núcleo do esquema. Consequentemente, explana, em cada fase da curadoria, quais são as partes interessadas e suas atividades

concernentes. A primeira fase - *Plan & Design* (Plano e Design) - dedicada ao planejamento do processo, interessa aos pesquisadores, aos órgãos de financiamento e ao suporte de pesquisa. Essa etapa é essencial para formatar o plano de gerenciamento de dados e esclarecer e definir as diretrizes que nortearão o compartilhamento dos dados. O segundo passo, *Collect & capture* (Colete e capture), em que são coletados e selecionados os dados relevantes para a pesquisa, concerne aos pesquisadores, ao suporte de pesquisa e aos serviços de computação. De acordo com o modelo, constam, nessa etapa, as seguintes atividades: expandir o uso de ambientes virtuais de pesquisa, automatizar a captura de dados, incorporar fluxos de trabalho na pesquisa e atribuir metadados aos dados que forem gerados. A terceira fase, *Interpret & analyse* (Interprete e analise), destinada à análise e à interpretação dos dados de pesquisa, interessa aos parceiros industriais (ou ao campo gerador desses dados), aos pesquisadores, aos colaboradores e aos serviços computacionais. É essencial, nessa etapa, investigar as implicações legais do armazenamento em nuvens, incentivar a gestão adequada do armazenamento, para evitar a perda de dados, e fornecer equipamentos e aplicativos seguros que possibilitem o compartilhamento de dados entre colaboradores. Em quarto lugar, vem o *Manage & preserve* (Gerencie e preserve). O gerenciamento e a preservação se relacionam com o suporte de pesquisa, os pesquisadores e os serviços computacionais. Esses atores têm o papel de incentivar o uso de padrões de metadados para garantir a interoperabilidade, incluir os dados nos formatos-padrão estabelecidos pela área que rege a pesquisa e incorporar os repositórios de dados a um sistema unificado. A penúltima fase - *Release & publish* (Lance e publique), envolve a publicação dos dados, por essa razão, importa à biblioteca, aos publicadores, aos serviços computacionais e aos pesquisadores. Aqui os dados originais geram novos dados. É preciso implementar autorizações e/ou embargos relacionados ao acesso aos dados publicados e incorporar, ao

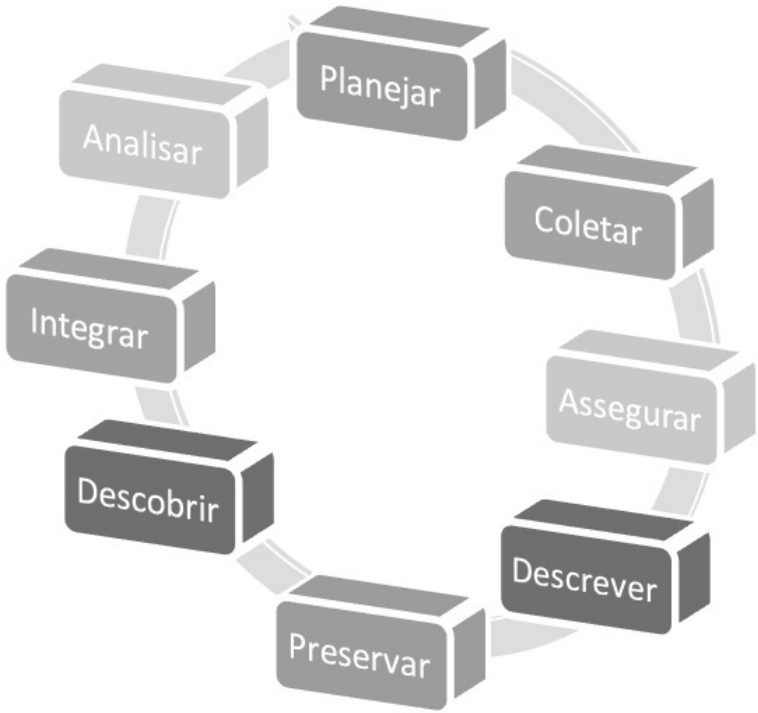
conjunto de dados, identificadores persistentes únicos, que possibilitem recuperá-los. A última etapa - *Discover & reuse* (Descubra e reutilize), é referente à reutilização dos dados e, para atingir o propósito, compromete a biblioteca, os colaboradores, os pesquisadores e o suporte de pesquisa. Nesse estágio, quem reutilizar dados deverá referenciar o pesquisador que os gerou. O suporte garantirá o acesso aberto adequado aos dados publicados. E os pesquisadores e os gestores do ciclo acompanharão as métricas e os impactos provocados pelo reuso dos dados, tendo como base as citações e as consequentes referenciações.

O terceiro ciclo de vida analisado neste capítulo foi desenvolvido por intermédio da Fundação *Data Observation Network for Earth – DataOne*[31]. O projeto é apoiado pela *National Science Fundation - NSF* e possibilita o acesso aberto a dados dispostos em vários repositórios de redes científicas e organizacionais, que chegam a um montante de 46,1 *terabytes* de dados e 28,0 *gigabytes* de metadados, até maio de 2018. Sua missão é de "permitir a criação de novas ciências e conhecimentos através do acesso universal a dados sobre a vida e a terra e o ambiente que a sustenta" (*DATAONE*, 2018, *online*, tradução nossa[32]). A iniciativa mantém um forte braço educacional, que empenha esforços para fornecer materiais que instruem pesquisadores para gerenciarem os dados adequadamente, desde sua coleta até a publicação. Para nortear seu empenho pedagógico, a *DataOne* estabeleceu o próprio ciclo de vida dos dados (Figura 5), com as seguintes fases: planejar, coletar, assegurar, descrever, preservar, descobrir, integrar e analisar.

31 Website do DataOne: https://www.dataone.org/

32 Texto original: Enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it (*DATAONE*, 2017, *online*).

Figura 5: Ciclo de vida dos dados



Fonte: Adaptado e traduzido de *DataOne* (2017)

Na primeira etapa, 'Planejar', cria-se o plano de gestão de dados (*Data Management Planning - DMP*), responsável por definir os tipos de dados que serão produzidos, os métodos de preservação e de manutenção da integridade, os *hardware* e os *software* necessários, as políticas de acesso e de compartilhamento, os requisitos de confidencialidade, privacidade e propriedade intelectual, as regras de citação e de referenciação, os planos para eventuais transições, integrações ou arquivamento de dados, dentre outros (*DATAONE*, 2017). A elaboração desse plano evita a duplicação de esforços e, por conseguinte, economiza tempo e contribui para que a equipe

não decida de última hora como tratar os dados. *NSB* indica os objetivos que precisam ser levados em consideração durante o planejamento:

> [...] garantir que todas as obrigações legais e as expectativas da comunidade na proteção da privacidade, segurança e propriedade intelectual sejam plenamente cumpridos; participar no desenvolvimento de padrões da comunidade para a coleta de dados, armazenamento, utilização, manutenção e migração; trabalhar no sentido da interoperabilidade entre as comunidades científicas e incentivar a integração multi-disciplinar de dados; assegurar que as decisões comunitárias sobre a coleta de dados tem em conta as necessidades dos usuários externos; incentivar o acesso livre e aberto sempre que possível e; oferecer incentivos, recompensas e reconhecimentos para cientistas que partilham e arquivam dados (*NSB*, 2005, p. 25, tradução nossa[33]).

Segundo o *DataOne* (2016), nessa fase, é elementar descrever como e quais dados serão geridos em curto e em longo prazos, compatibilizados em suas diferentes versões e submetidos a *backups*. Tão importante quanto é definir os equipamentos e os programas computacionais que serão utilizados durante a coleta, a manipulação, a análise, a preservação e o compartilhamento e quais medidas de segurança e de acesso serão implantadas. Também são definidos papéis e responsabilidades para a gestão, quem irá coletar os dados, gerar metadados, analisar e liderar

33 Ensure that all legal obligations and community expectations for protecting privacy, security, and intellectual property are fully met; participate in the development of community standards for data collection, deposition, use, maintenance, and migration; work towards interoperability between communities and encourage cross disciplinary data integration; ensure that community decisions about data collections take into account the needs of users outside the community; encourage free and open access wherever feasible; and provide incentives, rewards, and recognition for scientists who share and archive data (*NSB*, 2005, p. 25).

o projeto, fazer parte do suporte técnico e gerenciar os *backups* e os arquivamentos e o controle de versões.

A segunda fase, 'Coletar', estabelece padrões para a manipulação inicial dos dados. Entradas e valores foras do padrão dificultam a recuperação. Por exemplo, determinado conjunto de dados está sendo tabulado em planilha de *excel*, e a coluna referente à data contém linhas que se alteram entre aa/bb/cccc (dia/mês/ano) e aa/bbbb (mês/ano). Esse conflito quer dizer que usuário e *software* terão que procurar o mesmo tipo de dado utilizando dois métodos diferentes e alongando o tempo de recuperação. Também há leituras irregulares, quando são misturados textos e números em uma mesma coluna. *DataOne* (2017) aconselha que o pesquisador evite inserir ou corrigir o conjunto de dados, a fim de preservar a integridade e garantir sua credibilidade. É preferível que ele altere os atributos só para leitura, visando minimizar as chances de terceiros, não autorizados, manipularem qualquer parte do conteúdo. Durante a tabulação, deve-se adotar a organização por colunas e linhas padronizadas. Cada coluna única é portadora de entrada/campo específico, e cada linha tem seu valor inserido sob a diretriz de um formato previamente acordado. Sugere, ainda, o *DataOne* (2017) que a planilha seja salva com um formato legível pelos diversos sistemas operacionais e pelos banco de dados, porque, geralmente, formatos específicos, tipo "xls" e "xlsx", são atualizados frequentemente ou substituídos, mas formatos genéricos como o "csv" são básicos e, por isso, lidos na maioria dos sistemas. A ideia é de evitar que os dados coletados se tornem obsoletos ou não interpretáveis em longo prazo por causa de erros na coleta.

'Assegurar', conforme o *DataOne* (2016), é garantir que o conjunto de dados não sofra contaminações ocasionadas por erros de comissão e/ou omissão nem gere efeitos negativos resultantes de sua divulgação. Erros de comissão acontecem quando dados incorretos ou imprecisos são

coletados e integrados ao conjunto, e erros de omissão, quando os dados são documentados inadequadamente ou omitidos, empurrados para baixo do tapete. Quando existem dois ou mais responsáveis por alimentar um conjunto de dados, aumenta a probabilidade de inserções duplicadas. Dados contaminados, viciados ou enviesados comprometem o resultado das análises, sua consistência e veracidade. Outro cuidado necessário é em relação à sensibilidade dos dados, ou seja, os conteúdos são passíveis de receber restrições de acesso e divulgação por causa de fatores como segredo de mercado, embargo estabelecido por pesquisa científica, violência ou eroticidade explícita, privacidade, sigilo judiciário, dentre outros. O *DataOne* (2017) recomenda que se deve assegurar se os dados têm propriedades confidenciais, analisar possíveis problemas decorrentes da liberação dos dados e executar e reavaliar as políticas de divulgação e de acesso a eles.

A etapa 'Descrever' fundamenta-se na implantação de metadados, que representarão, de forma detalhada, um conjunto ou coleção de dados, que facilitam a compreensão e a avaliação de aplicabilidade em possíveis reutilizações. Assim, quando forem compartilhados, seus usuários potenciais saberão de antemão, sem, necessariamente, ter acesso à completude dos dados, seu conteúdo e o público-alvo, o formato, o tipo, onde e quando foram criados, em quais políticas de uso estão envolvidos, quem são seus responsáveis e quais os softwares e os *hardware* que são necessários para garantir seu uso, sua segurança e sua preservação. Essa fase também define se os metadados serão incorporados aos dados e/ou disponibilizados separadamente (*DATAONE*, 2017). Geralmente os arquivos digitais já têm metadados integrados preenchíveis ao dispor do usuário, localizados na opção 'propriedades' e divididos nas abas 'geral', 'segurança', 'detalhes' e 'versões anteriores'. Com todas essas informações, o pesquisador avaliará se o conjunto de dados é aplicável às suas necessidades. Os metadados são uma espécie de relatório, cujo registro

> [...] evitará a duplicação de dados porque os investigadores podem determinar se os dados já existem. Os cientistas são capazes de compartilhar informações confiáveis sobre um conjunto de dados através de seus metadados. Os cientistas que desejam reutilizar um conjunto de dados podem ter certeza de sua origem e qualidade (*DATAONE*, 2016, *online*, tradução nossa[34]).

A fase 'Preservar' inclui temas como *backup* e arquivamento, que, embora semelhantes, têm funções diferentes. Os *backups* protegem a integridade dos dados, previnem perdas parciais e totais e envolvem cópias periódicas de segurança, para minimizar a dependência aos dados originais e seus consequentes prejuízos decorrentes do uso contínuo. Uma boa prática indicada pelo *DataOne* (2016) consiste em manter várias versões de *backups* em locais diversos. Imagine que um incêndio toma o terceiro e o quarto andares do prédio administrativo de uma empresa de médio porte, justamente nos ambientes onde eram armazenados os *backups* e todos os dados originais produzidos. A perda não seria total, se essa empresa mantivesse *backups* em lugares distintos simultaneamente, por exemplo, em um serviço remoto de armazenamento em nuvens e em computadores e *Hard Disk Drive* - HDs externos localizados em prédios diferentes.

Já o arquivamento é direcionado para um conjunto de dados que pararam de ser acrescidos e alimentados. Não há mais trabalho baseado nele, e sua utilidade é tão somente para consulta de referência. Seu conteúdo serve como parâmetro para outras atividades, mas não é relevante para ser reutilizado ou incorporado a novos projetos. É hora de armazenar

34 Texto original: Metadata records will help avoid data duplication because researchers can determine if data already exists. Scientists are able to share reliable information about a dataset by creating metadata and passing it along with the dataset. Scientists wishing to reuse a dataset can be confident of its origins, data quality (*DATAONE*, 2010, *online*)

definitivamente. Alguns cuidados são necessários para garantir a consulta aos dados em longo prazo, como convertê-los em formatos padrões, que podem ser lidos em qualquer computador, independentemente das versões do sistema operacional, e adotar o uso das extensões lidas em softwares não proprietários (*DATAONE*, 2017). Isso minimizará problemas concernentes à compatibilidade e à desatualização. Também é importante manter uma agenda para fazer *backups* e indicar os responsáveis por criar essas cópias de segurança.

'Descobrir' é o período do ciclo em que os dados são encontrados por terceiros. Todavia, para que isso ocorra, é necessário publicá-los, pois, se não estiverem hospedados em repositório apropriado, destinado ao compartilhamento, provavelmente serão esquecidos, engavetados ou descartados. É essencial comunicar, com clareza, se os dados estão disponíveis para serem reutilizados ou somente visualizados. Também é preferível utilizar repositórios de fonte aberta. Sayão e Sales (2015) afirmam que a atribuição de identificadores persistentes contribui para identificar dados, como o *Digital Object Identifier* (DOI), o *Uniform Resource Identifier* (URI), o *Persistent Uniform Resource Locator* (PURL), o *The Handle System* (HDL) e o InChi (*IUPAC International Chemical Identifier*). Ainda conforme Sayão e Sales (2015), há vantagens para o próprio pesquisador quando ele publica dados em repositórios, como o reconhecimento por ser uma fonte transparente e confiável, ser citado e referenciado e melhorar a qualidade e a apresentação de seus dados através das críticas recebidas pela comunidade científica.

Em seguida, vem a fase 'Integrar', em que ocorre a reutilização. No entanto, esse período depende do êxito nas etapas anteriores. Para isso, é imprescindível gerenciar o ciclo de vida, objetivando uma iminente integração do conjunto de dados publicados. Ao se tornar acessíveis, os

dados serão integrados em outros projetos, a depender de sua relevância. O responsável pela integração precisa manter um modelo conceitual que descreva as relações entre os conjuntos de dados criados e reaproveitados de diferentes fontes. Além disso, se novos dados nascerem dessa integração, será preciso continuar transparente, sinalizar qual a contribuição dos dados reaproveitados e manter seus metadados citando suas fontes.

Por fim, 'Analisar', que é, de fato, o momento de avaliar se os objetivos traçados estão sendo atingidos. Para isso, o curador examinará, periodicamente, se o plano de gestão irá garantir a segurança e a preservação dos dados, comparando os padrões de metadados existentes e verificando se o adotado atende às expectativas; comparará o desempenho de diferentes repositórios de dados, para eleger qual é o adequado para seu projeto; fará, regularmente, testes de segurança nos *backups* e nos arquivos para evitar dados corrompidos ou desatualizados; assegurará a qualidade de seus dados, apurando possíveis contaminações ocasionadas por eventuais erros na hora de colhê-los e de manipulá-los e avaliará se continuam atendendo às necessidades do projeto e de suas políticas que permitem, proíbem ou inibem o compartilhamento total ou parcial de dados.

O Ciclo de Vida, da Curadoria Digital, deve fazer parte de um projeto maior, responsável por criar estruturas tecnológicas que suportem o volume de dados gerados, incumbido também de captar recursos humanos especializados na utilização de ferramentas de *software* que processam, analisam e compartilham dados e seus metadados. Da mesma forma, é preciso formatar políticas públicas para lidar com o fenômeno do megadados, a fim de prover um aparato legal e condições para que as organizações consigam captar, processar, armazenar, distribuir e preservar os dados, sem ferir os direitos dos autores. E, para o projeto funcionar adequadamente, evitando interrupções ou desatualizações, é essencial garantir recursos financeiros, os quais manterão estruturas, tecnologias e

pessoas voltadas ao desempenho, a curto, médio e/ou longo prazo, das fases de curadoria digital descritas neste capítulo.

Referências

ABREU, G. O. L.. *A soberania dos dados versus a autonomia do usuário*: big data, internet das coisas e as estratégias afirmativas do anonimato. João Pessoa: UFPB, 2015. Dissertação de mestrado defendida em Comunicação, Universidade Federal da Paraíba, 2015. Disponível em: <http://bdtd.ibict.br/vufind/Record/UFPB_e0883f5a020b2ccfb280eb51ac67db4e>.

ANTONIUTTI, C. L.. *Usos do big data em campanhas eleitorais*. Rio de Janeiro: UFRJ, 2015. Tese de Doutorado defendida em Ciência da Informação, Universidade Federal do Rio de Janeiro, 2015. Disponível em: <http://bdtd.ibict.br/vufind/Record/IBCT_30edf379ab0daa0d84d6d58da44a03e0>.

AZUCAR, D.; MARENGO, D. SETTANNI, Michele. Predicting the big 5 personality traits from digital footprints on social media: a meta-analysis. *Personality and Individual Differences*, v. 124, n. 1, p. 150-159, 2018. Disponível em: < https://www.sciencedirect.com/science/article/pii/S0191886917307328 >.

BALL, A. *Review of the state of the art of the digital curation of research data*. Bath: University of Bath, 2010. Disponível em: < http://opus.bath.ac.uk/19022/>.

BORGMAN, C. L. *Big data, little data, no data*: scholarship in the networked world. Cambridge: MIT Press, 2015.

CROSAS, M. et al. Automating open science for big data. *ANNALS*, American Academy of Political and Social Science, n. 659, p. 260-

273, 2015. Disponível em: <https://gking.harvard.edu/publications/automating-open-science-big-data>.

DATAONE. *Current member nodes summary*. 2017. Disponível em: <https://www.dataone.org/>.

DATAONE. *Best practices with the tag*: discover. 2017. Disponível em: <https://www.dataone.org/best-practices/discover>.

DATAONE. *Education modules*. 2016. Disponível em: <https://www.dataone.org/education-modules>.

DIGITAL CURATION CENTRE. *The DCC Curation Lifecycle Model*. 2018. Disponível em: <http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf>.

EREVELLES, S.; FUKAWA, N.; SWAYNE, L.. Big data consumer analytics and the transformation of marketing. *Journal of Business Research*, n. 69, p. 897-904, 2015. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0148296315002842>.

G1. *Brasil tem 116 milhões de pessoas conectadas à internet, diz IBGE*. 2018. Disponível em: <https://g1.globo.com/economia/tecnologia/noticia/brasil-tem-116-milhoes-de-pessoas-conectadas-a-internet-diz-ibge.ghtml>.

GIAMBRONE, G. P.. Information technology innovation: the power and perils of big data. *British Journal of Anaesthesia*, v. 115, n. 3, p. 339-342, 2015. Disponível em: <https://academic.oup.com/bja/article/115/3/339/312431>.

GOLDER, S. A.; MACY, M. W.. Digital footprints: opportunities and challenges for online social research. *Annual Review of Sociology*, v. 40, p. 129-152, 2014. Disponível em: <10.1146/annurev-soc-071913-043145>.

HESSE, B. W.; MOSER, R. P.; RILEY, W. T. From big data to knowledge in the social sciences. *ANNALS*, American Academy of Political and Social Science, n. 659, p. 16-32, 2015. Disponível em: <http://journals.sagepub.com/doi/abs/10.1177/0002716215570007>.

IBM. *What is changing in the realm of big data*? Disponível em: <https://www.ibm.com/big-data/us/en/>.

LANEY, Doug. 3D data management: controlling data volume, velocity, and variety. *Application delivery strategies*, n. 949, 2001. Disponível em: <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.

LAU, R. Y. K. Big data commerce. *Information & Management*, n. 53, p. 929-933, 2016. Disponível em: <https://scholars.cityu.edu.hk/en/publications/publication(2b95e41e-d199-4585-ad09-462062fb1189).html>.

LIMA, F. R. B.; SANTOS, P. L. V. A. da C.; SANTARÉM SEGUNDO, J. E. Padrão de metadados no domínio museológico. *Perspectivas em Ciência da Informação*, v. 21, n. 3, p. 50-69, jul./set., 2016. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2639>.

LORD, P.; MACDONALD, A. *E-science curation report*: data curation for e-science in the UK: an audit to establish requirements for future curation and provision. Twickenham: JISC: JCSR, 2003. Disponível em: <https://www.cs.york.ac.uk/ftpdir/pub/leo/york-msc-2007/information/vsr-curation/science-dc-report.pdf>.

MCKEN, K. et al. *Research360*: the research lifecycle. Bath: University of Bath: JISC: UKOLN, 2012. Disponível em: <http://opus.bath.ac.uk/32292/1/Bath_Mcken_research_data.pdf>.

MEHMOOD, A. et al. Protection of big data privacy. *IEEE Access*, v. 4, p. 1821-1834, 2016. Disponível em: <https://ieeexplore.ieee.org/document/7460114/>.

NATIONAL SCIENCE BOARD. *Long-lived digital data collections*: enabling research and education in the 21 century. Arlington: National Science Foundation, 2005. Disponível em: <https://www.nsf.gov/geo/geo-data-policies/nsb-0540-1.pdf>.

POMERANTZ, J.. *Metadata*. Cambridge: MIT Press, 2015.

PRICE, D. J. de S.. *Little science, big science*. New York: Columbia University Press, 1963. Disponível em: <http://www.andreasaltelli.eu/file/repository/Little_science_big_science_and_beyond.pdf>.

SAYÃO, L. F.; SALES, L. F.. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. *Informação & Sociedade:* estudos, João Pessoa, v. 22, n. 3, p. 179-191, set./dez., 2012. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/12224>.

SAYÃO, L. F.; SALES, L. F.. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. *Tendências da Pesquisa Brasileira em Ciência da Informação,* v. 6, n. 1, 2013. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/102>.

SAYÃO, L. F.; SALES, L. F.. *Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores*. Rio de Janeiro: CNEM, 2015. Disponível em: <http://carpedien.ien.gov.br/bitstream/ien/1624/1/GUIA_DE_DADOS_DE_PESQUISA.pdf>.

SANTOS, P. L. V. A. da C.; SIMIONATO, A. C.; ARAKAKI, F. A.. Definição de metadados para recursos informacionais: apresentação da metodologia BEAM. *Informação & Informação*, v. 19, n. 1, p. 146-163,

jan./abr., 2014. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/15251>.

SEIFE, C.. Big data: the revolution is digitized. *Comment Books & Arts*, v. 518, p. 480-481, 2015. Disponível em: <https://www.nature.com/articles/518480a>.

7

MODELO ORGANIZACIONAL PARA GESTÃO INTEGRADA DE DADOS DA BIODIVERSIDADE BRASILEIRA

Pedro Luiz Pizzigatti Corrêa

A Gestão de Dados é a disciplina responsável por definir, planejar, implantar e executar estratégias, procedimentos e práticas necessárias para o gerenciamento, de forma efetiva, dos recursos de dados e informações nas organizações. (BOHLE, 2013)

Com a expansão de grandes repositórios de dados e seu uso nas atividades científicas, uma nova área de pesquisa foi estabelecida, denominada de *Data Science*, que visa elaborar modelos para a descoberta de conhecimentos em um grande volume de dados, provenientes de múltiplas fontes, incluindo sensores, aplicativos, instrumentos de coleta, mídias sociais e notícias.

Um dos grandes desafios para viabilizar a nova ciência baseada em dados científicos abertos é o volume e a diversidade de dados desestruturados que a pesquisa científica gera e que envolve a procura de padrões e de comportamentos e estabelece novos processos para gerar conhecimentos.

O principal desafio que se impõe, em especial, às organizações de pesquisa brasileiras, como as Universidades, os Centros de Pesquisas e as Agências Financiadoras de Pesquisas, é de criar infraestruturas integradas de alcance local/

institucional, agregadas nacional ou regionalmente, e baseadas nos padrões de interoperabilidade disponíveis nas comunidades científicas internacionais.

Esse desafio foi abordado para a gestão de dados científicos da biodiversidade brasileira em Corrêa (2012a, b), que definiu um modelo para a gestão integrada de dados científicos de biodiversidade brasileira, por meio de um estudo de caso aplicado às demandas do Ministério do Meio Ambiente. O modelo proposto estabeleceu: (i) as diretrizes para a gestão de dados nas instituições de pesquisa vinculadas ao Ministério do Meio Ambiente (MMA); (ii) as diretrizes de política de dados institucional; e (iii) uma infraestrutura computacional para integrar dados científicos de biodiversidade (DA SILVA, 2014). Esse modelo foi aplicado na integração de dados do Ministério do Meio Ambiente a partir de 2012, o que gerou o Portal da Biodiversidade Brasileira, disponibilizado em novembro de 2015. O Sistema está operacional na infraestrutura de Tecnologia da Informação do MMA disponível em: <portaldabiodiversidade.icmbio.gov.br>.

Este capítulo apresenta o modelo organizacional e as diretrizes de política de dados institucional definido em Corrêa (2012a, b) para a gestão de dados científicos de biodiversidade, estabelecendo os papéis, os processos e os relacionamentos intra e inter organizacionais.

Infraestruturas de dados de biodiversidade

Uma infraestrutura organizacional para integração de dados de biodiversidade, também chamada de rede (*network*), é um conjunto de instituições, sistemas e serviços, interligados de forma estruturada a fim de possibilitar a interoperabilidade e o acesso unificado a diversos recursos computacionais distribuídos como 'nós' em uma rede (CAMPBELL, 2003). Algumas dessas infraestruturas consideradas são:

- *Global Biodiversity Information Facility*[1] – GBIF (GBIF, 2016);
- *Data Observation Network for Earth*[2] – DataONE (DATAONE, 2016);
- *Inter American Biodiversity Information Network – Pollinators Thematic Network*[3] – IABIN PTN (IABIN_PTN, 2016).

Essas infraestruturas se apoiam em normas, padrões e protocolos que garantem a interoperabilidade dos diversos ambientes, sistemas e dados heterogêneos.

O GBIF foi criado a partir de uma iniciativa global, com o objetivo de promover a utilização eficiente do conhecimento sobre a diversidade biológica do planeta, um dos grandes desafios do Século XXI. O GBIF apresenta "um mundo onde a informação sobre biodiversidade é gratuita e universalmente disponível para a ciência, para a sociedade e para um futuro sustentável". Sua missão é de se tornar a mais importante fonte global de informações sobre biodiversidade e ser um gerador de soluções inteligentes para o meio ambiente e o bem-estar. Para alcançar esses objetivos, o GBIF encoraja uma grande variedade de fontes/publicadores de dados de todo o mundo a descobrir, a organizar, a digitalizar e a publicar seus dados por meio de sua rede de biodiversidade.

O DataONE é um programa patrocinado pela *National Science Foundation* (NSF) dentro do programa *DataNet*. Seu objetivo principal é o de armazenar dados ecológicos e ambientais produzidos por cientistas do mundo inteiro. A meta do DataONE é de preservar e dar acesso a dados multiescalas, multidisciplinares e multinacionais. O DataONE interliga uma infraestrutura computacional existente para disponibilizar um

1 Portal de dados GBIF: <www.gbif.org>
2 Portal do DataONE: <www.dataone.org>
3 Portal da IABIN: <http://www.oas.org/en/sedi/dsd/iabin/>

modelo distribuído, que utiliza tecnologias robustas para gerir os dados, com foco na preservação dos dados de biodiversidade em longo prazo.

O IABIN PTN é uma rede temática de polinizadores que compõe uma das seis redes temáticas estabelecidas no âmbito da *Inter-American Biodiversity Information Network* (IABIN). Seu objetivo principal é de desenvolver um banco de dados interligado e integrado entre as principais fontes de dados de polinizadores e os membros IABIN, que compartilham conteúdo crítico relacionado a polinizadores, por meio de um conjunto comum de padrões e protocolos de compartilhamento de dados.

O IABIN PTN disponibiliza um catálogo *online* dinâmico e interligado de polinizadores do hemisfério ocidental, que inclui dados sobre:

- nomes (*checklists*) de abelhas, beija-flores, morcegos e outras espécies polinizadoras de grande importância;
- exemplares das principais coleções;
- contato dos polinizadores e dos especialistas;
- associações polinizadores-plantas;
- literatura sobre polinizadores;
- demais informações relacionadas (como, por exemplo, dados geográficos, códigos de barras genéticos etc.).

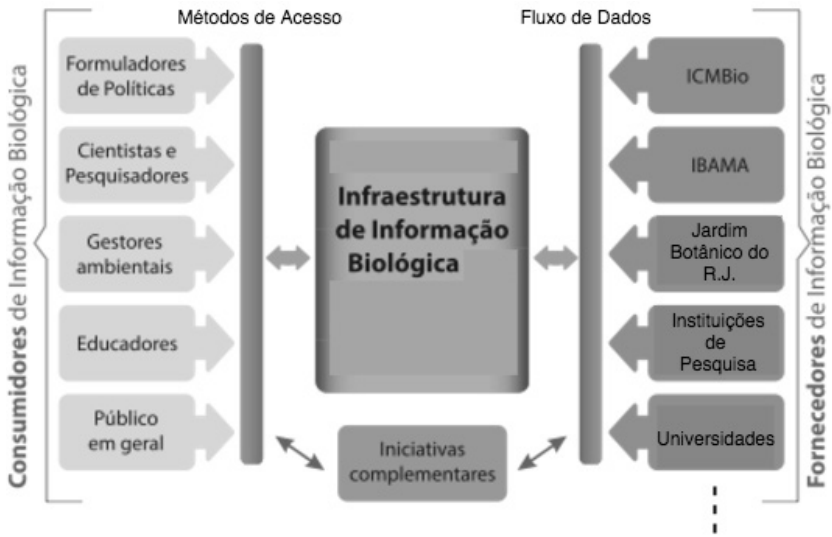
Modelo organizacional adotado

O Modelo Organizacional foi definido para a gestão de dados integrados de biodiversidade baseado nas infraestruturas anteriores. Esse modelo organizacional define a infraestrutura de informação e as diretrizes para apoiar a política de dados institucional voltados para a integração de dados de biodiversidade.

A infraestrutura de informação biológica (CAMPBELL, 2003) é o elemento-chave de integração e articulação de fornecedores e consumidores

de informação biológica, como demonstra a Figura 1. Essa infraestrutura é responsável por integrar, compartilhar, publicar e sintetizar dados biológicos gerados e manipulados no âmbito dessa organização. Os fornecedores de informação são responsáveis por organizar os dados biológicos primários das Instituições de pesquisa, de órgãos governamentais, de centros de pesquisas e de universidades participantes, interessados em compartilhar dados biológicos. Na Figura 1, citam-se, como exemplos, as instituições governamentais e de pesquisa vinculadas ao Ministério do Meio Ambiente: o Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio[4]), o Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA[5]) e o Jardim Botânico do Rio de Janeiro (JBRJ[6]).

Figura 1: Representação conceitual da infraestrutura de informação biológica



Fonte: Baseado em Corrêa (2013-a)

4 Portal do ICMBio: http://www.icmbio.gov.br
5 Portal do IBAMA: http://www.ibama.gov.br
6 Portal do JBRJ: http://www.jbrj.gov.br

Os consumidores de informação biológica são os que gerenciam, estudam e utilizam dados biológicos, recursos e ferramentas computacionais. Podem ser do setor público ou do privado - cientistas, formuladores de políticas públicas, analistas ambientais do Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) e do governo federal, governos estaduais e governos locais, indústria, professores, estudantes e cidadãos.

Essa infraestrutura de informação integra consumidores e fornecedores de informação biológica, articulados também com sistemas de informação e iniciativas complementares externas, como, *Global Biodiversity Information Facility*[7] (GBIF), Sistema de Informação sobre a Biodiversidade Brasileira[8] (SIBBr), Infraestrutura Nacional de Dados Espaciais[9] (INDE), *Specieslink*[10], *Data Observation Network for Earth*[11] *(*DataONE) e Scielo[12].

Características da infraestrutura biológica

A Infraestrutura de Informação Biológica integra dados de biodiversidade disponíveis em sistemas de informação e bancos de dados. Os dados de biodiversidade disponíveis são os já digitalizados e que são gerenciados por algum sistema de informação já disponível atualmente ou por bases de dados em Excel, Access ou CSV disponíveis nos centros de pesquisa. A infraestrutura computacional é baseada em padrões de metadados e em protocolos internacionais para integrar dados de

7 Portal de Dados do GBIF: <www.gbif.org>
8 Portal de Dados de Biodiversidade Brasileiro: <www.sibbr.gov.br>
9 Portal da INDE: < www.inde.gov.br>
10 Portal de Dados de Biodiversidade: <www.splink.cria.gov.br>
11 Portal DataONE: <www.dataone.org>
12 Portal de Referências Bibliográficas de Biodiversidade: <www.scielo.br>

biodiversidade. Atualmente, os principais padrões e protocolos utilizados para o compartilhamento de dados de biodiversidade são definidos por meio de padrões abertos, mantidos pelo GBIF e pelo *Biodiversity Information Standards*[13] (TDWG).

Um dos componentes fundamentais para a gestão integrada de dados de biodiversidade é o estabelecimento de um modelo organizacional para o sistema. Assim, considerou-se o modelo utilizado por diferentes organizações internacionais, que é baseado em 'nós', como GBIF, NBII[14] (EUA), CONABIO[15] (México), dentre outras.

A gestão de dados de biodiversidade de cada 'nó' de informação deve ser definida a partir de Política de Dados de Biodiversidade Institucional, que estabelece as diretrizes da instituição para o compartilhamento e a preservação dos dados por ela mantidos.

Apoio à tomada de decisão

A infraestrutura de Informação Biológica é formada por sistemas de apoio à tomada de decisão para conservação da biodiversidade em diferentes níveis de informação, de acordo com a Figura 2. O **nível operacional** trata de decisões para monitorar a biodiversidade relativa à aquisição, à análise e à fiscalização do processo de sua conservação. O **nível de integração** agrupa e sintetiza dados do nível operacional e possibilita relacionar e compartilhar dados de diferentes sistemas desse nível. O **nível de integração** apoia a decisão gerencial sobre a biodiversidade de médio e de longo prazos. As ferramentas de síntese possibilitam ao nível gerencial

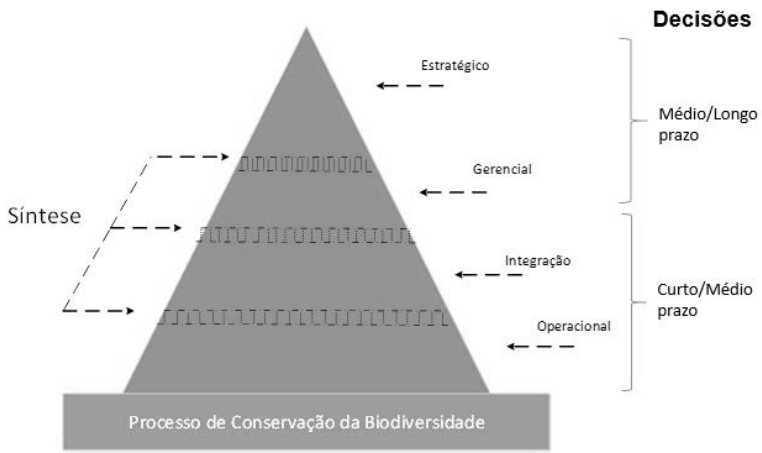
13 Portal de padrões de metadados e protocolos de interoperabilidade de biodiversidade: <www.tdwg.org>

14 National Biodiversity Information Infrastructure of USA: <http://www.usgs.gov/core_science_systems/Access/p1111-1.html>

15 Portal de Biodiversidade do México: <http://www.conabio.gob.mx>

análises mais complexas de dados. O **nível estratégico** usa as informações integradas para tomar decisões de longo prazo, como por exemplo, a criação de uma nova unidade de conservação.

Figura 2: Pirâmide dos níveis de tomada de decisão para conservar a biodiversidade.



Fonte: Baseado em Corrêa (2013)

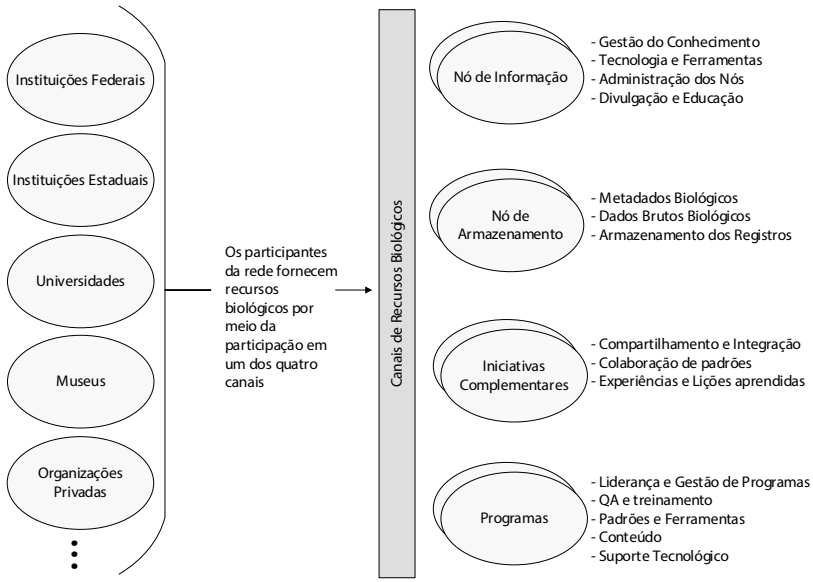
Deve-se deixar claro que esse sistema viabiliza a integração de dados entre os sistemas operacionais e os disponibiliza para sistemas gerenciais. O nível de integração não responde diretamente a todas as questões de nível gerencial e estratégico. Entretanto, viabiliza o desenvolvimento de sistemas mais sofisticados, que darão suporte à tomada de decisão em outros níveis.

Modelo baseado em nós

Uma infraestrutura de informação biológica (CONABIO, 1993) é organizada para garantir a gestão e o compartilhamento de informações de

biodiversidade por meio da participação dos parceiros (nós) (CAMPBEL, 2003) em um ou mais canais disponíveis na infraestrutura, que têm atribuições e responsabilidade bem definidas. Na Figura 3, apresenta-se essa organização.

Figura 3: Representação da relação dos parceiros (nós) na infraestrutura



Fonte: Baseado em Corrêa (2013-b)

Os **nós de informação** são os parceiros tecnológicos da rede que suportam o compartilhamento das informações e o bom funcionamento dos recursos disponíveis. Eles são interconectados e formam uma rede de informações, que também faz parte da rede global da infraestrutura. Os nós de informação podem disponibilizar serviços como:

- análise dos dados;
- desenvolvimento de ferramentas e suporte tecnológico;

- mineração dos dados;
- armazenamento dos dados;
- colaboração;
- treinamento.

Os **nós de armazenamento** são responsáveis pelo armazenamento de dados e metadados e auxiliam os consumidores de dados de biodiversidade a localizar, analisar e acessar esses dados e recursos de forma eficiente. Os 'nós' são responsáveis pelo compartilhamento dos dados e sua disponibilização para consumo. Questões sobre a agregação de dados dos parceiros e a qualidade dos dados compartilhados também são atividades desses 'nós'.

Os '**nós' das iniciativas complementares** são, em geral, parceiros dos nós de informação e de armazenamento, e sua participação na infraestrutura possibilita a criação de redes maiores e até globais de informações de biodiversidade. Eles também são importantes para garantir a troca de dados, de recursos e de experiências entre as diversas redes. Em geral, os nós de iniciativas complementares são focados no:

- compartilhamento de dados;
- desenvolvimento de padrões, protocolos e boas práticas;
- fornecimento de experiências e boas práticas conquistadas na gestão e no compartilhamento de dados científicos.
- Já os **nós de programas** objetivam dar suporte técnico, liderança, gestão e direcionamento estratégico para toda a infraestrutura. Além dessas atribuições, esses 'nós' também podem contribuir com informações de biodiversidade e disponibilização de recursos computacionais na rede.

Política de dados institucional

Política é um conjunto de orientações em conformidade com os objetivos de uma instituição. Essas orientações tomam como base os

valores fundamentais e a vocação da instituição e servem para definir estratégias, táticas e planos operacionais. A política de dados define objetivos estratégicos de longo prazo para gerenciar os dados em todos os aspectos de um projeto, agência ou organização. Pode-se definir a política de dados como um conjunto de princípios de alto nível, que estabelecem uma orientação para a gestão de dados (BURLEY; PEINE, 2009).

Em uma política de dados, são definidos os limites de responsabilidade, autoridade e possibilidade de acesso e utilização a um conjunto de dados. A política também pode ser utilizada para tratar de questões estratégicas, como aspectos jurídicos, aquisição, administração e custódia dos dados, dentre outras ações.

Alguns fatores devem ser considerados quando se vai definir uma política de dados, a saber:

- **Dinamismo e flexibilidade**: uma política de dados pode ser modificada diante de desafios não vislumbrados em sua concepção, diferentes tipos de projetos e parcerias potencialmente vantajosas, mantendo sua orientação no foco estratégico da organização;
- **Custo**: deve-se considerar o custo da obtenção dos dados *versus* o custo de acesso aos dados. O custo pode ser uma barreira tanto para os utilizadores adquirirem certos conjuntos de dados, quanto para os provedores fornecerem dados no formato ou na extensão requeridos;
- **Propriedade dos dados**: a propriedade dos dados deve ser claramente definida na política, que deve respeitar os direitos de propriedade intelectual que possam existir em diferentes níveis. Os proprietários normalmente têm direitos legais sobre esses dados, juntamente com os direitos de autor e de propriedade intelectual.

Normalmente, uma política deve contemplar a definição dos proprietários ou autores dos dados, os direitos de propriedade intelectual

e do autor, as obrigações legais e não legais relevantes para a organização, para que as atividades sejam compatíveis com as políticas de segurança de dados, o controle de divulgação, o respeito a contratos de licença, os acordos externos etc. Outras questões relacionadas aos autores dos dados são:

- **Custódia e carência**: depósito físico do dado e tempo em que um dado deve ser mantido em algum nível de sigilo por questões de segurança, por proteção para a publicação original ou por proteção à população ou indivíduo em que foi coletado, por questões de interesse econômico ou de outra natureza;
- **Privacidade**: uma política deve esclarecer quais dados são privados ou têm alguma restrição, assim como explicitar quais devem ser disponibilizados para um domínio público;
- **Responsabilidade legal**: uma política pode definir funções de gerenciamento de dados e responsabilidades (NPS, 2008) e deve determinar claramente os papéis associados às funções, estabelecer a propriedade dos dados em todas as fases de um projeto, incutir responsabilidades e assegurar o uso adequado, visando à qualidade dos dados.

Uma política deve informar como uma organização está protegida legalmente quanto ao uso dos dados, especialmente quando é possível causar algum dano a um indivíduo, espécie ou organização como resultado de abuso ou de má utilização dos dados. Normalmente é tratada como um "acordo de usuário final". O usuário, ao utilizar, concorda com o exposto na política. Uma política pode recomendar que uma "declaração de renúncia de responsabilidade" seja incluída no sistema de recuperação de metadados e dados de modo a isentar o fornecedor, o coletor de dados ou qualquer instituição associada ao conjunto de dados de qualquer responsabilidade legal pelo uso indevido ou imprecisões nos dados (BURLEY; PEINE, 2009).

Quanto à sensibilidade, alguns dados são ditos sensíveis. Por exemplo, a localização de espécies ameaçadas de extinção. É preciso que uma política deixe claro como será o acesso à base de dados, principalmente a informações sensíveis, que apresentam algum tipo de restrição. Uma política precisa abordar como se classifica a sensibilidade dos dados e como eles serão tratados em função dela. Um dado é dito sensível quando sua publicação pode resultar em um "efeito adverso" no táxon ou atributo em questão ou a um ser vivo (CHAPMAN; GRAFTON, 2008), (CHAPMAN, 2005).

Para classificar a sensibilidade dos dados, normalmente são considerados fatores como: o tipo das ameaças, a vulnerabilidade do táxon ou atributo, a constatação de que esses dados foram divulgados em outras bases, dentre outros.

Três classes de fatores são ponderados na determinação da sensibilidade dos dados: risco, impacto e sensibilidade da publicação (ver Figura 4).

Figura 4: Classes de fatores que determinam a sensibilidade dos dados

Fonte: Autor

O risco avalia se o táxon é prejudicado por alguma atividade humana (caça, pesca, extração, etc.). O impacto avalia as consequências da atividade humana com a publicação dos dados; a sensibilidade da publicação avalia se a publicação ou o uso dos dados de uma maneira incorreta aumenta a probabilidade de algum dano relacionado à conservação das espécies envolvidas.

A política de dados institucional deve se orientar pelos acordos internacionais assinados pelo Brasil e a legislação vigente. Em relação aos acordos internacionais, podemos citar como exemplo a Convenção de Diversidade Biológica (ECO-92), que estabelece o compartilhamento de dados de biodiversidade. Em 2012, o Brasil assinou um acordo com a *Global Biodiversity Facility* (GBIF) para compartilhar dados de biodiversidade brasileiros (http://www.gbif.org/country/BR/summary). Quanto à legislação nacional vigente e às regulamentações específicas, podemos citar a Lei nº 12.527, de 18 de novembro de 2011- lei de acesso à informação; a Lei nº 9.610, de 19 de fevereiro de 1998 – que dispõe sobre direitos autorais; e a Lei nº 10.650, de 16 de abril de 2003 – que dispõe sobre o acesso público aos dados e às informações existentes nos órgãos e nas entidades integrantes do Sistema Nacional de Meio Ambiente[16] (SISNAMA).

A seguir, são discutidos marcos regulatórios internacionais e nacionais que estabelecem diretrizes para a política de dados institucional.

Convenção sobre diversidade biológica

A Convenção sobre Diversidade Biológica[17] (CDB) é um tratado da Organização das Nações Unidas (ONU) que foi estabelecido durante

16 Portal do SISNAMA: http://www.mma.gov.br/governanca-ambiental/sistema-nacional-do-meio-ambiente

17 Portal sobre a CDB: http://quipronat.wordpress.com/2009/07/06/eco-92-cdb-93-e-ai/

a ECO-92: Conferência das Nações Unidas sobre Meio Ambiente e Desenvolvimento (CNUMAD), realizada no Rio de Janeiro em junho de 1992. Mais de 160 países assinaram o tratado, que entrou em vigor em dezembro de 1993.

O conceito de diversidade biológica, e inicialmente, associava-se ao número de espécies que habitavam determinado espaço geográfico e era sinônimo de riqueza específica. Com a CDB, novos aspectos foram incorporados a essa definição. O conceito atual de diversidade biológica procura referir e integrar toda a variedade e variabilidade que se encontra nos organismos vivos, em seus diferentes níveis, e os ambientes onde estão inseridos. O art. 2º, III, da Lei Brasileira n.º 9.985/2000, define diversidade biológica como a variabilidade de organismos vivos de todas as origens, que compreende, entre outros, os ecossistemas terrestres, marinhos e outros ecossistemas aquáticos e os complexos ecológicos de que fazem parte, além da diversidade de espécies e de ecossistemas.

Os objetivos da CDB são de conservar a biodiversidade, utilizar, de forma sustentável, seus componentes e repartir, justa e equitativamente, os benefícios derivados da utilização dos recursos genéticos, mediante, inclusive, o acesso adequado a esses recursos e a transferência de tecnologias pertinentes.

Apesar de a CDB ser um acordo internacional com força de lei, dá aos países- membros liberdade para estabelecerem as normas e os mecanismos que possibilitarão o alcance dos objetivos nela previstos, instituindo mais um compromisso do que uma obrigação específica. Cada nação deve, mais do que integrar a questão da política ambiental em sua política nacional, criar programas específicos de proteção de sua biodiversidade, além de identificar elementos importantes dela e lhes assegurar tratamento especial, gestão e proteção.

Para a CDB, a diversidade biológica é uma preocupação comum à humanidade e reconhece explicitamente a soberania dos países em sua gestão. Essa mudança de paradigma na titularidade da biodiversidade tem importantes implicações no contexto do poder e das relações internacionais, pois, a partir de sua entrada em vigor, os componentes da biodiversidade, com incalculável valor econômico potencial, não mais podem ser acessados livremente.

Global biodiversity information facility

A rede GBIF foi estabelecida em 2001 por governos para estimular o acesso livre e aberto a dados de biodiversidade por meio da Internet. É resultante do Fórum da Ciência Global da OCDE, que considera que o esforço necessário para integrar recursos de informática e seus usuários em uma unidade sinérgica e interoperável é que torna a informática biológica um esforço de megaciência.

A rede GBIF tem como visão um mundo onde dados sobre biodiversidade estão disponíveis de forma livre e universal para a ciência e para a sociedade, rumo a um futuro sustentável. Sua missão é de ser o recurso mais importante para a biodiversidade e gerar soluções inteligentes para questões ambientais e bem-estar humano (GBIF, 2016).

Trata-se de uma iniciativa multilateral estabelecida por acordos intergovernamentais, baseados em um memorando não vinculativo (*Memorandum of Understanding* - MoU). São participantes do GBIF: países (governos), economias, organizações intergovernamentais ou internacionais, organizações com uma missão internacional ou instituições designadas por essas organizações. O Brasil aderiu ao GBIF em outubro de 2012.

O órgão máximo do GIBF é o Conselho de Administração (*Governing Board*, GB), ao qual é subordinado o Comitê Executivo, responsável por acompanhar as ações do Secretariado e dos demais participantes da implementação das decisões do GBIF. O Secretariado é a instância executiva do GBIF, juntamente com os 'nós' dos países participantes.

Para que uma instituição possa publicar dados na rede GBIF, é necessário ter o aval de um dos participantes da rede, de um país ou de uma organização internacional. Portanto, um provedor de dados integrará sua base informacional à rede GBIF, através de um 'nó' ou com o seu aval. O principal usuário-alvo da rede GBIF é a comunidade científica.

A rede GBIF tem como missão disponibilizar dados sobre biodiversidade na Internet de maneira livre e aberta. Procura estruturar um ambiente positivo para o compartilhamento de dados e seu uso para beneficiar a ciência e a sociedade e contribuir para um futuro sustentável. O MOU estabelecido com os provedores de dados tem um parágrafo específico sobre propriedade intelectual no qual se indica claramente que as leis e os acordos estabelecidos serão respeitados. O GBIF não detém qualquer direito sobre os dados compartilhados na rede, mas pode proteger as ferramentas que irá desenvolver. Todavia, o MoU indica que o GBIF deverá transferir os aplicativos desenvolvidos para as instituições de pesquisa, principalmente as de países em desenvolvimento. Indica, também, que a autoria dos dados será explicitamente atribuída a cada provedor, a quem cabe determinar que dado será de acesso público e explicitar as restrições (se houver) em relação ao uso de seus dados.

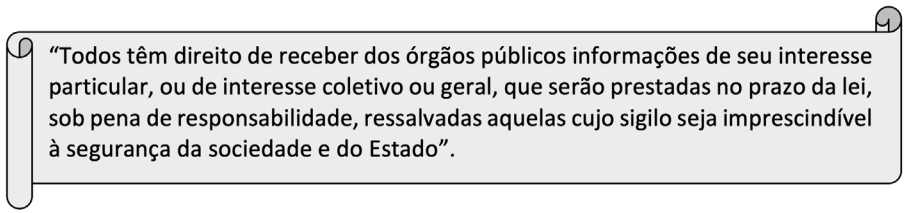
Diferentemente da Convenção da Diversidade Biológica, em que a maioria dos pontos focais é governamental, na rede GBIF, tem-se também

a participação de instituições de direito privado, sem fins lucrativos, como representantes técnicos oficiais. São exemplos de tais representantes a CONABIO, uma instituição governamental interministerial do México; o INBio[18] (INBIO, 2016), uma organização civil de direito privado sem fins lucrativos da Costa Rica; e o Instituto Humboldt[19] (HUMBOLDT, 2016), da Colômbia, uma organização civil de direito privado sem fins lucrativos.

Regulamentações federais brasileiras

A publicidade dos atos de governo é um princípio democrático, que, no Brasil, aparece expresso no artigo 5º da Constituição, principalmente em seu inciso XXXIII (ver Figura 5).

Figura 5: Artigo da Constituição que define o direito a acesso a dados governamentais



Fonte: Baseado no artigo 5º da Constituição Brasileira, inciso XXXIII

Do ponto de vista dos dados digitais, o projeto *Open Definition*[20] (OPEN_DEFINITION, 2016) define o dado aberto como o que pode

18 Portal INBio: http://www.inbio.ac.cr/
19 Portal do Instituto Humboldt da Colômbia: http://www.humboldt.org.co/
20 Portal do Open Definition: http://opendefinition.org/

ser livremente utilizado, reutilizado e redistribuído por qualquer pessoa. A definição completa dá detalhes específicos do significado do termo, que se resumem em alguns pontos importantes:

- **Disponibilidade e acesso**: o dado precisa estar disponível por inteiro e a um custo razoável de reprodução, preferencialmente por meio de *download* na Internet, e em um formato conveniente e configurável;
- **Reuso e redistribuição**: o dado precisa ser fornecido em condições de serem reutilizados e redistribuídos, incluindo o cruzamento com outros conjuntos de dados;
- **Participação universal**: todos podem usar, reutilizar e redistribuir, pois não se discriminam áreas de atuação, pessoas ou grupos.

Modelo de política de dados institucional

A seguir, apresenta-se um modelo de tópicos e conteúdos de um documento que descreve a Política de Dados Institucional (Quadro 1). Para fins de ilustração, são colocados trechos de exemplos baseados em uma organização científica que lida com dados marinhos. Esse exemplo foi baseado no modelo proposto pelo *National Biodiversity Network*[21] (NBN). (NBN, 2016).

21 Portal da NBN: http://www.nbn.org.uk/Share-Data/Managing-Permissions/Model-Agreements/Model-Data-Sharing-Use-Policy.aspx

Quadro 1: Modelo de Política de Dados Institucional

Política de Dados de Biodiversidade

1. Introdução

Inicia-se a política com uma breve descrição do propósito da política. É uma oportunidade para identificar públicos-alvo e demais entidades relacionadas.

Exemplo:

Esse documento define a política de utilização e compartilhamento dos dados de biodiversidade do âmbito da organização XYZ. Por meio da política, busca-se alcançar uma abordagem racional e consistente para a gestão de dados e o uso. Essa política se destina a auxiliar os provedores de dados e os usuários a entenderem os objetivos e as intenções da organização.

2. Objetivos

Nesta seção, definem-se os objetivos estratégicos a serem alcançados por meio dessa política e qual será o papel da gestão dos dados nesse processo. Nesse ponto, normalmente deixa-se claro o escopo de quais dados serão tratados para o alcance dos objetivos.

Exemplo:

Os dados de espécies marinhas mantidos pela organização XYZ fazem parte de uma rede de dados que atua em colaboração com outras organizações, como o Departamento ZWY do governo brasileiro. Visa-se salvaguardar dados de pesquisa sobre espécies marinhas, dentro do território nacional, disponibilizando-os como recurso nacional de informação e apoio à ciência marinha, para melhorar a gestão do ambiente marinho.

3. Fonte de dados

Nessa seção, explicitam-se quais tipos de dados são geridos e suas fontes. Isso possibilita que potenciais provedores de dados identifiquem interesse e que potenciais usuários saibam quais os dados que são armazenados e seus fornecedores.

Exemplo:

A empresa XYZ coleta, gerencia e armazena dados e imagens coletados por uma variedade de agências públicas e privadas que realizam pesquisas no ecossistema marinho e que estão credenciadas segundo a Resolução 1234/2013. Os dados armazenados correspondem a uma série de espaço-temporal de longo prazo, com registros de fauna e de flora marinha, além de dados de *habitat*.

A empresa XYZ não armazena dados marinhos físicos e químicos. Os dados foram limitados a espécies marinhas mamíferas, répteis, peixes e plâncton.

4. Uso de dados

Nessa seção, apresenta-se o que pode ser feito com os dados recebidos. Se há um padrão de armazenamento e formato de dados, é aqui que eles são explicados. Questões como propriedade, curadoria e demais responsabilidades sobre os dados também podem ser incluídas nessa seção.

Exemplo:

Os dados são mantidos em um banco de dados relacional e, mensalmente, são feitas cópias de segurança. Os dados estão sob a responsabilidade do gestor de dados institucional, que responderá por sua preservação e disponibilização. O técnico é responsável pela execução das atividades que operacionalizam o acesso e o armazenamento dos dados.

5. Compartilhamento de dados

Nessa seção, apresenta-se o conjunto de diretrizes relacionado ao compartilhamento de dados. Normalmente são apresentadas diretrizes comuns a todos os casos (premissas), e se houver exceções, são apresentadas. Recomenda-se que sejam apresentadas, de forma clara, as razões que possam restringir o acesso a algum dado. É importante considerar questões legais e demais diretrizes da organização durante a escrita desse item.

Exemplo:

A empresa XYZ tem como princípio permitir o acesso aos dados contidos em seu repositório. Esses acessos são permitidos quando para usados sem fins lucrativos, educacionais e de pesquisa, e outros fins de utilidade pública. No entanto, podem existir situações em que a empresa XYZ, como um repositório de dados, pode restringir o acesso a toda ou parte de alguns recursos de dados. Sempre que restrições forem aplicadas, sua devida justificativa e a decisão lógica devem ser documentadas e disponibilizadas. São focos de restrições liberações, cujo efeito pode aumentar o risco de dados ao meio ambiente ou colocar espécies particularmente sensíveis em risco.

6. Demais termos e condições

Nessa seção, podem ser apresentadas as demais permissões para quem acessa os dados, que podem ser diferentes para diferentes usuários, com suas devidas justificativas. É possível, por exemplo, afirmar que o conjunto de dados são públicos para fins de pesquisa, sem fins lucrativos, mas é possível também garantir que uma porção dos dados seja licenciada para algum uso comercial específico.

Exemplo:

O repositório de dados da XYZ é formado de dados oriundos de diversas entidades públicas e privadas além de pesquisadores individuais. O titular do direito de cada dado deve ser informado nos metadados associados. Cada dado que possa ser copiado para fins comerciais deve ser explicitamente identificado pelos provedores de dados. As regras para uso comercial para fins específicos são definidas na norma ABC dessa organização.

8. Mais informações

Nessa seção, são apresentadas referências para diretrizes, legislações, regulamentações etc., que possam ter servido de base para formular a política. Recomenda-se que seja explicitado um canal direto de comunicação, para auxiliar a implementar e a cumprir essa política, assim como esclarecer eventuais dúvidas.

Fonte: Adaptado de National Biodiversity Network[22] (NBN)

22 Exemplo de Política de Dados da NBN: http://www.nbn.org.uk/Share-Data/Managing-Permissions/Model-Agreements/Model-Data-Sharing-Use-Policy.aspx

Considerações finais

As possibilidades de utilizar e reutilizar os dados científicos para gerar novas hipóteses e investigações é o cerne da Nova Ciência, que apresenta uma inovação no contexto científico e uma mudança na forma de pensar, de fazer e de reproduzir a pesquisa.

A definição e a aplicação de modelos para a Gestão de Dados Científicos Abertos em universidades, institutos de pesquisa e agências de fomento de pesquisas é fundamental para resolver alguns dos desafios no contexto contemporâneo de Ciência Aberta no Brasil. Assim, é fundamental avaliar experiências e soluções aplicadas em escala global e aberta, que possam ser reproduzidas nas instituições brasileiras, em coordenação integrada e de maneira multidisciplinar, em seus aspectos políticos, científicos e tecnológicos.

O Modelo de Gestão Integrada de Dados Científicos de Biodiversidade apresentado neste capítulo contribui com esse desafio, pois define e estabelece um modelo conceitual para a integração de dados brasileiros de biodiversidade, baseado em um modelo organizacional de federação de 'nós' de informação e diretrizes de política de dados institucional. Esse modelo foi validado por meio do estudo de caso aplicado no Ministério de Meio Ambiente.

Estabelecer um modelo inicial de gestão de dados científicos e apoiar a definição de mecanismos de compartilhamento de dados de biodiversidade contribui para que os atores responsáveis por fornecer e consumir dados de biodiversidade se entendam.

Referências

BOHLE, S. What is e-science and how should it be managed? Nature.com, Spektrum der Wissenschaft. 2013.

BURLEY, T E.; PEINE, J. D. USGS - Science for a changing world: NBII-SAIN Data Management Toolkit. 2009. Disponível em: <http://pubs.usgs.gov/of/2009/1170/>. Acesso em: 20 maio 2014.

CAMPBELL, J. NBII Enterprise Architecture - Section 2 - Business Architecture. NBII Report Program. 2003.

CHAPMAN, A.D.; GRAFTON G. *Guide to best practices for generalizing primary species-occurrence data*, version 1.0. Copenhagen: Global Biodiversity Information Facility, 2008. 27 p. (87-92020-06-2).

CHAPMAN, A.D. Uses of primary species-occurrence data. *Report for the Global Biodiversity Information Facility*, Copenhagen, v. 1, 2005.

CONABIO. *Dados nacionais de biodiversidade do México*. Disponível em: www.conabio.gob.mx. Acesso em: 10 jan. 2016.

CORRÊA, P. L. P. *Arquitetura para integração de sistemas de informação de biodiversidade*. Brasília – DF. Diretoria de Conservação da Biodiversidade - Secretaria de Biodiversidade e Florestas - Ministério do Meio Ambiente. Março de 2013. **Relatório** 4-a

CORRÊA, P. L. P. *Infraestrutura organizacional para gestão de dados de Biodiversidade do Ministério do Meio Ambiente*. Brasília – DF. Diretoria de Conservação da Biodiversidade - Secretaria de Biodiversidade e Florestas - Ministério do Meio Ambiente. Março de 2013. **Relatório** 5-b

CORRÊA, P. L. P.; STANZANI S. L., DA SILVA D. L. *Arquitetura para a integração de dados de biodiversidade do Instituto Chico Mendes de Conservação da Biodiversidade*. Brasília – DF. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. Junho de 2013. Relatório 3-c

DA SILVA D. L.; CORRÊA, P. L. P.; JUAREZ, K.M.; FONSECA; R.L. *Diretrizes para a integração de dados de biodiversidade***.** Ministério

do Meio Ambiente. Apoio da Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. ISBN: 9788577381951. 2014. 100p.

GBIF. *Global Biodiversity Information Facility*. Fonte: http://www.gbif.org/Acesso em: 05 fev. 2016

HUMBOLDT. Instituto Humboldt da Colômbia. Disponível em: http://www.humboldt.org.co/. Acesso em: 05 fev. 2016.

IABIN_PTN. Inter American Biodiversity Information Network – Pollinators Thematic Network. Disponível em: http://www.oas.org/en/sedi/dsd/iabin/. Acesso em: 05 fev. 2016.

INBIO. Instituto de Biodiversidade da Costa Rica. Fonte: www.inbio.ac.cr. Acesso em: 05 fev. 2016.

NBN. National Biodiversity Network (NBN). Rede de dados de biodiversidade nacional da Inglaterra – Política de uso e acesso aos dados. Fonte: http://www.nbn.org.uk/Share-Data/Managing-Permissions/Model-Agreements/Model-Data-Sharing-Use-Policy.aspx. Acesso em: 05 fev. 2016.

OPEN_DEFINITION. Projeto Open Definition que estabelece os princípios de dados abertos. Fonte: http://opendefinition.org/. Acesso em: 05 fev. 2016.

8

OS PRINCÍPIOS FAIR: viabilizando o reuso de dados científicos

Guilherme Ataíde Dias
Renata Lemos dos Anjos
Adriana Alves Rodrigues

Introdução

A questão do uso e do reuso dos dados científicos é um tema presente e considerado muito importante por toda a comunidade de pesquisa internacional. As nações associadas à Comunidade Europeia, os Estados Unidos e o Canadá estão em um estágio de desenvolvimento avançado em ciência e tecnologia que endossam e promovem iniciativas que possibilitam o uso e o reuso de dados científicos. Essa situação é abordada por Mons *et al* (2017), que asseveram que o Conselho Europeu, assim como o G7, em sua quadragésima segunda reunião de cúpula, realizada no Japão em 2017, estabeleceram que a *Open Science* (Ciência aberta) assim como a reusabilidade de dados de pesquisa são assuntos primordiais em suas agendas.

A adoção de práticas que possibilitam que pesquisadores brasileiros reusem dados científicos tem começado a ser difundida de forma mais sistemática recentemente, embora de forma ainda tímida. A partir

de pesquisa em andamento realizada por pesquisadores da Universidade Federal da Paraíba (UFPB) (Processo CNPq 310825/2015-6) acerca das práticas de gestão de dados por pesquisadores brasileiros na área da Ciência da Informação (CI), foi possível verificar alguns fatos pertinentes a esse domínio específico, a saber: a maioria dos pesquisadores da área da CI alega conhecer o processo de gestão de dados de pesquisa, mas têm dúvidas sobre a devida operacionalização do processo e mencionam a falta de tempo e de recursos para prosseguir com a prática. Em relação a essas constatações, Costa (2017, p.77) acrescenta:

> 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 Ciência da Informação para a *e-science* no contexto brasileiro. (COSTA, 2017, p.77).

No âmbito institucional brasileiro, existem iniciativas relacionadas ao processo de gestão de dados científicos, mas as consideramos ainda em sua gênese. Os esforços ainda não são realizados de forma coordenada entre os diversos atores interessados e não existe uma política nacional sobre o tema. Indicamos como iniciativas pioneiras no processo de gestão de dados científicos o *Programa FAPESP de Pesquisa em eScience*[1] e as ações associadas ao desenvolvimento do *Portal da Biodiversidade*[2] (COSTA, 2017). No momento, sabemos de iniciativas do IBICT/RNP e de algumas universidades federais para o desenvolvimento de ações que possibilitem soluções para a gestão de dados científicos em seus respectivos domínios de atuação. Quanto à existência de repositórios específicos para dados, no

1 Disponível na URL http://www.fapesp.br/8436
2 Disponível na URL http://portaldabiodiversidade.icmbio.gov.br

Brasil, indicamos os listados na pesquisa de Costa (2017) e mais alguns que encontramos ao longo de nossa investigação, de forma não exaustiva: o Repositório de Dados de Estudos Ecológicos do Programa de Pesquisa de Biodiversidade da Amazônia Ocidental (PPBIO); o Repositório de Dados Ecológicas de Longa Duração (PELD[3]); o Repositório de Dados de Pesquisa UNIFESP[4]; e a Plataforma GeoInfo da Embrapa[5].

Contrastando com os passos iniciais dados por poucas instituições brasileiras, com foco no desenvolvimento de políticas e de infraestruturas de suporte para a gestão de dados científicos, várias ações já foram desenvolvidas no exterior. Essa iniciativa, que merece destaque e está se se tornando bastante conhecida, é identificada pela sigla FAIR (princípios FAIR) e seu objetivo é de garantir que os dados sejam encontrados com mais facilidade (***F**indable*), acessíveis (***A**ccessible*)**,** interoperáveis **(*I**nteroperable*) e** reutilizáveis **(*R**eusable*).**

Doorn e Dillo (2016) esclarecem que os princípios FAIR foram criados em janeiro do ano de 2014, em um seminário sobre ciências da vida, com o objetivo de contribuir com a gestão de grandes volumes de dados que estão sendo produzidos pela ciência.

Para deixar mais claro o que são os princípios FAIR, trazemos a explicação de Rodriguez-Iglesias *et al*:

> FAIR é um acrônimo de Encontrável, Acessível, Interoperável e Reutilizável. Resumidamente, os Princípios FAIR sugerem que cada elemento de dados deve ter um identificador único global, que deve ser associado a metadados contextuais e pesquisáveis ('Encontrável'). Esses identificadores devem resolver dados ou metadados usando um protocolo aberto, padrão ('Acessível'); os dados e os metada-

3 Disponível na URL http://memoria.cnpq.br/repositorio-peld
4 Disponível na URL https://repositoriodedados.unifesp.br/
5 Disponível na URL http://www.embrapa.br/geoinfo

> dos devem usar uma linguagem de representação formal, amplamente aplicável, e vocabulários e ontologias abertas e amplamente aceitas para o domínio relevante ('Interoperável'); e, finalmente, os dados devem ser ricamente descritos com referências cruzadas abundantes e com um mecanismo claramente definido para acessar informações de proveniência e licença ('Reutilizável'). (RODRIGUEZ-IGLESIAS *et al*, 2016, p.1, tradução nossa[6]).

A ideia de um conjunto de princípios que pode contribuir para o uso e o reuso de dados científicos foi muito bem recepcionada pela comunidade interessada no compartilhamento de dados e causou impacto em outros domínios tão distintos como a "arqueologia e monitores ambientais para 'cidades inteligentes'" (MONS *et al*, 2017, p.49-50). Ressaltamos que os princípios FAIR não abordam a questão da qualidade dos dados. Esse tópico pode ser encontrado nas etapas dos diversos ciclos de vida de dados existentes.

Os princípios FAIR não estão associados exclusivamente aos processos de interação de dados com seres humanos (MONS et al, 2017), mas também à facilidade dos processos de interação dos dados com agentes artificiais. É cada vez mais comum o uso de agentes computacionais inteligentes para a coleta automática de dados científicos. A adesão aos princípios FAIR pode contribuir para melhorar o trabalho desses agentes.

A seguir, apresentamos cada um dos princípios FAIR de forma mais detalhada e discutimos sobre eles.

---

6 Texto original: FAIR is an acronym of Findable, Accessible, Interoperable, and Reusable. Briefly, the FAIR Principles suggest that every data element should have a globally-unique identifier, and that this identifier should be associated with contextual, searchable metadata ("Findable"); these identifiers should all resolve to data or metadata using an open, standard protocol ("Accessible"); the data and metadata should use a formal, broadly applicable representation language, and utilize open and widely-accepted domain-relevant vocabularies and ontologies ("Interoperable"); and finally, the data should be richly described with an abundance of cross-references, and with a clearly-defined mechanism for accessing provenance and license information ("Reusable").

Discutindo sobre os Princípios FAIR

O Quadro 1 apresenta os quatro princípios orientadores FAIR e os conceitos associados a cada um dos seus princípios.

Quadro 1: Os princípios orientadores FAIR

FAIR: Princípios orientadores
F - Ser encontrável (*Findable*)
E1. Os (meta)dados são atribuídos a um identificador persistente, único e global.
E2. Os dados são descritos com metadados ricos (definidos por R1 a seguir).
E3. Os metadados incluem, de forma clara e explícita, o identificador dos dados que descrevem.
E4. Os (meta)dados são registrados ou indexados em um recurso pesquisável.
A - Ser acessível (*Accessible*)
A1. Os (meta)dados são recuperáveis por seu identificador, usando-se um protocolo de comunicação padronizado.
A1.1 O protocolo é aberto, gratuito e universalmente implementável.
A1.2 O protocolo possibilita um procedimento de autenticação e autorização, quando necessário.
A2. Os metadados são acessíveis, mesmo quando os dados não estão mais disponíveis.
I - Ser interoperável (*Interoperable*)
I1. Os (meta)dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para representar o conhecimento.
I2. Os (meta)dados usam vocabulários que seguem os Princípios FAIR.
I3. Os (meta)dados incluem referências qualificadas para outros (meta)dados.
R - Ser reutilizável (*Reusable*):
R1. Os meta(dados) são ricamente descritos com uma pluralidade de atributos precisos e relevantes.
R1.1. Os (meta)dados são disponibilizados com uma licença de uso de dados clara e acessível.
R1.2. Os (meta)dados estão associados a uma proveniência detalhada.
R1.3. Os (meta)dados estão de acordo como padrões comunitários relevantes para o domínio.

Fonte: Adaptado de Wilkinson et al (2016, *online)*

A primeira premissa para que os dados científicos possam ser usados e reutilizados por pesquisadores está fundada no fato de que esses dados e os respectivos metadados associados devem ser fáceis de ser encontrados por quem deseje usá-los. De que serve disponibilizar dados através de um serviço se eles não podem ser encontrados?

Para maximizar a encontrabilidade dos dados, é muito importante que sejam atribuídos aos conjuntos de dados identificadores persistentes, únicos e globais. Atualmente, o identificador persistente mais utilizado para identificar objetos digitais é o *Digital Object Identifier* (DOI), que pode ser atribuído por diversas organizações. Mas, no caso específico dos dados científicos, uma organização que se destaca é o *DataCite*[7], cujo foco são dados. Dessa forma, essa organização está apta a prover serviços customizados para um domínio específico de usuários. O objetivo do *DataCite*, conforme explicitado no seu *website*, é de "[...] ajudar a comunidade de pesquisa a localizar, a identificar e a citar dados de pesquisa com confiança" (DATACITE, 2018, *online*, tradução nossa[8]).

A ideia de descrever dados com metadados ricos (*rich metadata*) está relacionada ao fato de que o pesquisador deveria ser capaz de encontrar os dados desejados, independentemente de ter acesso ao seu identificador (Princípio ***F****indable*). Para isso, baseia-se nas informações que podem ser obtidas por meio dos metadados associados aos dados. Uma pesquisa feita com uma ferramenta de busca deveria ser suficiente para encontrar o conjunto de dados desejado. Então, sugere-se que o pesquisador use extensivamente metadados descritivos e indique detalhadamente todos os atributos que caracterizem os respectivos conjuntos de dados. É uma

7 https://www.datacite.org

8 Texto original: [...] help the research community locate, identify, and cite research data with confidence.

boa prática prover o maior detalhamento possível dos dados usando seus metadados (GOFAIR, 2018).

Quando o conjunto de dados é encontrado por uma pessoa ou por algum mecanismo automatizado, eles devem ser accessíveis (Princípio ***A****ccessible)*, levando-se em consideração também questões relacionadas à autenticação e à autorização (GOFAIR, 2018).

O Princípio *Accessible* está relacionado ao fato de que os dados deveriam poder ser acessados por protocolos-padrões. Tecnológicas "esotéricas", fechadas, com poucas implementações e mal documentadas devem ser evitadas. Gofair (2018) esclarece que a possibilidade de não prover acesso seguro para dados sensíveis através de protocolos automatizados está em conformidade com a ideia dos princípios FAIR. Uma forma de contornar essa situação seria disponibilizar nos metadados informações de contato, como *e-mail*, telefone e outras que possibilitem que a pessoa interessada nos dados contate seu detentor.

Associada a necessidade de os dados serem acessíveis está a diretiva de que os metadados devem ser acessíveis, mesmo quando os dados não estão mais disponíveis. Essa ideia diz respeito ao fato de que, com o passar do tempo, os dados poderão não estar mais disponíveis *online* devido a uma série de possibilidades, dentre elas, as questões associadas ao custo de disponibilizar os conjuntos de dados nas redes. Nessa situação, é preciso disponibilizar metadados com informações que possibilitem identificar os indivíduos ou as instituições detentoras dos dados.

A questão da interoperabilidade entre conjuntos de dados e metadados é abordada pelo Princípio FAIR ***I****nteroperable*, que está relacionado à necessidade de integrar dados a outros conjuntos de dados (considerando-se também sempre os metadados) e com as mais variadas aplicações ao longo do seu ciclo de vida. Para que seja possível uma efetiva interoperabilidade entre conjuntos de dados, é importante que existam

instrumentos para padronizar semanticamente os sistemas envolvidos no processo. Nesse cenário, os instrumentos que podem contribuir são os vocabulários controlados, os tesauros e as ontologias (GOFAIR, 2018), que também devem seguir os princípios FAIR.

A inclusão de referências qualificadas em conjuntos dados para outros conjuntos de dados envolve um conjunto de dados que pode ter sido construído a partir de outros conjuntos de dados ou em casos em que as informações complementares acerca de um conjunto de dados estão armazenadas em diferentes conjuntos de dados (GOFAIR, 2018).

Entendemos que o elemento precípuo para preservar dados é a possibilidade de reusá-los. Isso reduz os custos nos esforços de pesquisas e possibilita que os resultados de outras pesquisas sejam validados por qualquer pesquisador que tenha acesso aos conjuntos de dados. É nesse contexto em que o princípio FAIR ***R****eusable* se insere. Entendemos que o princípio *Reusable* representa as ideias associadas à iniciativa FAIR, pois a possibilidade de usar e de reusar os dados é o norte da iniciativa. Para possibilitar um aumento no re(uso) dos dados, o profissional responsável por sua publicação, além de incluir metadados relacionados ao processo de descoberta dos conjuntos de dados, deve prover metadados que descrevam, de forma detalhada, o contexto associado à criação dos dados. Dentre as diversas possibilidades, listamos algumas: a versão usada na criação/coleta dos dados, a maneira como determinado sensor foi calibrado, a data e as coordenadas geográficas associadas ao momento da coleta etc. (GOFAIR, 2018).

No contexto do uso dos dados, as questões sobre os direitos de usar os conjuntos de dados são importantes, complexas e devem estar disponíveis tanto para humanos quanto para agentes artificiais. O que pode ser feito com os dados e quem pode usá-los deve ser especificado de forma bastante clara em um termo de uso. Questões relativas à propriedade dos conjuntos de dados e o que pode e não pode ser feito com

eles, na perspectiva de determinado ordenamento jurídico, devem estar explicitadas.

As licenças do tipo *Creative Commons* são um exemplo de modelo de licenciamento que pode ser aplicado no contexto dos conjuntos de dados, contudo, essas licenças, embora úteis, devem estar sintonizadas com as respectivas legislações nacionais. Rodrigues *et al* (2016) discutem sobre o ordenamento jurídico brasileiro a respeito dessa temática.

Para que seja viável re(usar) os conjuntos de dados, é importante conhecer de onde eles vieram (proveniência). Gofair (2018) explica que, para outros usarem seus dados, devem saber sua origem, como fazer as devidas citações/referências e apresentar um fluxo de trabalho que possibilite a resposta para as seguintes questões: Quem gerou ou coletou os dados? Como eles foram processados? Esses dados já foram publicados antes? Os dados incluem conteúdos de outras fontes?

Finalmente, indicamos que, para estar de acordo com o princípio *Reusable,* é importante aderir a padrões e a formatos que sejam compreendidos pela comunidade de usuários. Exemplos disso são os metadados *Dublin Core*, os arquivos em formato de texto, vocabulários controlados padronizados etc.

Considerações finais

Os princípios FAIR são muito recentes no âmbito da comunidade associada ao uso e ao reuso de dados científicos, mas já causam impacto. Discussões sobre sua aplicação estão acontecendo em eventos relacionados à *e-Science*. Entendemos que os princípios FAIR podem e devem ser usados para ampliar o acesso aos dados científicos.

Poucos são os trabalhos que avaliam efetivamente a aplicação do FAIR em casos concretos. No caso específico da área da Ciência da

Informação, no Brasil, não conhecemos nenhum trabalho em periódico especializado que aborde o assunto. Entendemos que o número de pesquisas sobre esse tema deverá aumentar à proporção que aumentarem os dados científicos disponibilizados em repositórios especializados.

Referências

COSTA, M. M. *Diretrizes para uma política de gestão de dados científicos no Brasil*. 2017. 288f. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade de Brasília, Brasília, 2017.

DATACITE. *Our mission*. Disponível em: <https://www.datacite.org/mission.html>. Acesso em: 14 jun. 2018.

DOORN, P; DILLO, I. *Fair data in trustworthy datarRepositories Webinar*. EUDAT. 2016. Disponível em: <https://eudat.eu/events/webinar/fair-data-in-trustworthy- data-repositories-webinar > Acesso em: 14/06/2018.

GOFAIR. *FAIR principles*. Disponível em: <https://www.go-fair.org/fair-principles/>. Acesso em: 14 jun. 2018.

MONS, B. *et al*. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European open science cloud. *Information Services & Use*. v.37, n.1, p.49-56. 2017. Disponível em: <http://doi.org/10.3233/ISU-170824>. Acesso em: 07 nov. 2017.

RODRIGUEZ-IGLESIAS, A. *et al*. Publishing FAIR Data: an exemplar methodology utilizing PHI-Base. *Frontiers in plant science*. 7, 641. 2016. Disponível em: <http://doi.org/10.3389/fpls.2016.00641>. Acesso em: 07 nov. 2017.

RODRIGUES, A. A.; DIAS, G. A.; VIEIRA, A. A. N. CREATIVE COMMONS E PRODUÇÃO COLABORATIVA NO CONTEXTO DO ORDENAMENTO JURÍDICO BRASILEIRO. In: Luísa Neto; Fernanda Ribeiro. (Org.). *Direito e informação na Sociedade em Rede*: atas. Porto: Faculdade de Direito da Universidade do Porto, 2016.

WILKINSON, M.D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016. Disponível em: <http://doi.org/10.1038/sdata.2016.18>. Acesso em: 07 nov. 2017.

9

O COMPARTILHAMENTO DE DADOS CIENTÍFICOS NA ERA DO E-SCIENCE

Flavio Ribeiro Córdula
Wagner Junqueira de Araújo

Introdução

Ao longo dos Séculos XX e XXI, principalmente a partir do pós-Segunda Guerra Mundial, iniciou-se uma ruptura paradigmática na comunicação científica, com o advento e a popularização da Internet e das tecnologias – digitais – de informação e comunicação. A comunicação científica é parte inerente do desenvolvimento da ciência e está fundamentada na informação científica, gerando conhecimentos, divulgado, sobretudo, por meio de periódicos científicos (GUEDES, 1998). O periódico científico, por sua vez, impulsiona a disseminação da produção científica e desempenha um papel fundamental no meio acadêmico, promovendo avanços e destacando autores e editores. Na década de 1990, com a propagação da informação digital e o uso da Internet, houve a grande ruptura no modo de editorar e de disseminar informações, principalmente da produção científica (FACHIN, 2002).

Uma consequência inconteste e, possivelmente, natural dessa revolução científica (KUHN, 1997) pressupõe que os dados gerados, obtidos, coletados e utilizados na produção de artigos científicos também devam ser disponibilizados em repositórios digitais. Para Sayão e Sales (2014, p. 85), "de uma forma definitiva, a ciência orientada por dados e pelas tecnologias digitais criam um ponto de inflexão no ciclo tradicional da comunicação científica", ou seja, que impacta diretamente o modo de conduzir as pesquisas, e por consequência, o registro científico.

A utilização de infraestrutura computacional e de *software* científico possibilitou o surgimento de novas técnicas de organização e troca de informações em prol da ciência. Os métodos de obtenção de resultados científicos, por intermédio de computação intensiva, e o grande volume de dados são conhecidos como e-Ciência ou *e-Science*. Mesmo no Brasil, o termo *e-Science* é o mais reconhecido e, por isso, utilizado neste artigo. Na década de 1990, John Taylor, diretor-geral do Escritório de Ciência e Tecnologia do Reino Unido, criou esse termo para se referir ao uso da tecnologia para realizar investigações científicas (YANG; WANG; VON LASZEWSKI, 2009).

Segundo Costa (2017, p. 52), a contemporaneidade do tema *e-Science* "traz à tona questões conceituais que ainda não passaram pelo processo de reflexão necessário ao seu amadurecimento". Segundo a autora, encontram-se na literatura alguns termos relacionados ao *e-Science*, tais como: ciência orientada por dados (*data-driven science*), computação fortemente orientada para dados (*data-intensive computing*), ciberinfraestrutura (*cyberinfrastructure*), ciência com uso intensivo de dados, quarto paradigma da ciência (*fourth paradigm of science*), dilúvio de dados (*data deluge*), E-infraestrutura (*E-infrastructure*), entre outros.

Os dados são criados ou produzidos de várias maneiras: por meio de observação, de visualização, de monitoramento e de sensores, na criação de metadados, por análise de comportamento, por cálculos matemáticos e estatísticos etc. Praticamente tudo o que é produzido e coletado no ambiente de pesquisa pode ser considerado dado (SAYÃO, SALES, 2014; BORGMAN, 2015; DATAONE, 2017). Os dados se proliferam de duas maneiras diferentes: em formato digital e material. Diferentemente do formato material – papel, pinturas, desenhos – os dados, quando em formato digital, não podem ser interpretados sem o devido suporte tecnológico (BORGMAN, 2015).

O compartilhamento de dados é um dos vários aspectos relacionados à gestão de dados e é um conceito distribuído ao longo de todo o seu ciclo de vida (DATAONE, 2017). O acesso a dados de pesquisas relacionados – dados que se relacionam a sua pesquisa – possibilita que membros da comunidade científica reproduzam, comparem e avaliem melhor métodos e resultados.

Este texto foi produzido com alicerce metodológico no campo das Ciências Sociais e visa investigar os principais aspectos relacionados à gestão de dados na era do *e-Science*, com foco no compartilhamento. Como norte, foi empregado o ciclo de vida dos dados da *Data Observation Network for Earth* (*DataONE*). Para isso, recorreu-se aos materiais oficiais disponibilizados pelo *site* da *DataONE,* a publicações em livros, dissertações, teses, anais de congressos e ao Portal de Periódicos Capes, no qual foram encontrados artigos científicos em que foram usados termos relacionados à temática 'compartilhamento de dados científicos', conforme demonstrado no Quadro 1. Esses termos foram pesquisados em sua forma singular e plural, além de suas expressões equivalentes na língua inglesa. Foram baixados 43 arquivos, armazenados, lidos e estudados, embora nem todos tenham sido citados.

Quadro 1: Termos utilizados no Portal de Periódicos Capes

TERMOS OU EXPRESSÕES DE BUSCA
e-Ciência, dados científicos, dados de pesquisa, dados abertos, dados abertos de pesquisa, compartilhamento de dados, compartilhamento de dados científicos, curadoria de dados, ciclo de vida dos dados, gestão de dados, gestão de dados científicos.
e-Science, scientific data, research data, open data, open research data, data sharing, scientific data sharing, data curation, data life cycle, data management, scientific data management.

Fonte: Elaborado pelos autores

Nos tópicos seguintes, este trabalho traz, primeiramente, alguns esclarecimentos sobre dados e, em seguida, aprofunda-se em seus desdobramentos, com uma discussão sobre questões relativas à ciência orientada para dados, para o ciclo de vida dos dados, segundo a *DataONE* e para o compartilhamento de dados. Por fim, apresenta as considerações finais sobre a pesquisa.

Dados

Dados é um termo bastante explorado pela comunidade acadêmica de Ciência da Informação (CI) e bem mais complexo do que geralmente sugerido por pesquisadores ou agências financiadoras de pesquisa (BORGMAN, 2015). Mesmo depois de mais de cinco séculos de uso (BORGMAN, 2015), ainda não há uma definição consensual para esse termo s. Geralmente, e assim como está sendo feito neste artigo, os dados são discutidos e exemplificados, mas não definidos de fato. Até mesmo a *Data Documentation Initiative* (*DDI*) não define o termo 'dados' de forma

concreta, apenas mostra exemplos do que é ou não é um dado. A *DDI* é um padrão internacional que descreve os dados produzidos por pesquisas e outros métodos de observação nas ciências sociais, comportamentais, econômicas e de saúde.

Muito se discute sobre dados, principalmente no meio acadêmico e em políticas governamentais, mas pouco esforço é empregado para definir esse termo. O tripé dados - informação - conhecimento, segundo Borgman (2015), tende a simplificar demais a relação e os conceitos desses complexos termos. Ainda segundo essa autora, a melhor sumarização do termo dados indica que eles são "representações de observações, objetos e outras entidades utilizadas como evidências de um fenômeno para propósitos de pesquisa" (BORGMAN, 2015, p. 42, tradução nossa).

Para Sayão e Sales (2016), o termo 'dado de pesquisa' tem um significado amplo e se transforma de acordo com os domínios científicos, com os objetos de pesquisa com as metodologias de geração e a coleta de dados, entre outras variáveis. Os autores afirmam, ainda, que esse termo pode ser o resultado de um experimento realizado em um ambiente controlado de laboratório, um estudo empírico na área de ciências sociais ou, até mesmo, a observação de um fenômeno cultural. Entende-se, porém, que o termo que corresponderia mais adequadamente a essa afirmação seria 'dados científicos'.

Os dados dão garantias e evidências da veracidade dos resultados das pesquisas e dos artigos científicos publicados e servem de alicerce para o progresso da ciência (MOLLOY, 2011; BORGMAN, 2015). Sayão e Sales (2014, p. 77-78) referem que "uma sequência genômica, a velocidade de partículas subatômicas, as respostas de levantamento social [...], as imagens de satélites de outros planetas, todos esses recursos informacionais são como dados de pesquisa". Embora os dados sejam um meio para se atingir um fim – geralmente um artigo científico ou livro – raramente

as pesquisas são elaboradas considerando-se uma possível reutilização dos dados (BORGMAN, 2015), mesmo que essa reutilização seja para uso próprio, em uma possível continuação ou expansão de pesquisa científica.

A *long tail* – ou calda longa – é um termo estatístico utilizado para identificar a distribuições de dados. Segundo Borgman (2015, p. 25, tradução nossa), é "uma maneira popular de caracterizar a disponibilidade e o uso de dados nas mais diversas áreas de pesquisa". Esse termo foi cunhado por Chris Anderson, em um artigo na *Revista Wired*, em 2004, em que ele afirmou que grandes empresas usam essa estratégia, como a Amazon, da Apple e da Netflix (ANDERSON, 2004), por exemplo.

Quando aplicada a pesquisas científicas, a *long tail* revela, segundo Borgman (2015), que 15% da distribuição dos dados estão na cabeça da curva, e os 85% restantes se encontram distribuídos ao longo da calda. Isso significa dizer que apenas alguns campos de pesquisa – como a Astronomia, a Física, a Geologia, a Macroeconomia e outras áreas, por exemplo – realmente trabalham com grandes quantidades de dados (BORGMAN, 2015). Em poucas palavras, o volume de dados é distribuído de forma desigual pelos vários campos da ciência.

O *bigness*, por sua vez, é uma importante característica a ser considerada quando se estuda a relação entre os dados e a *e-Science*. *Big Science* e *little Science* são termos diferenciados por Price (1963), não pelo tamanho de seus projetos, mas por sua maturidade. Enquanto o *big Science* pode ser caracterizado como internacional e colaborativo, o *little Science* remete a características de pesquisas independentes. Já a distinção entre *big data* e *little data* pode ser vista por alguns autores como uma escala relativa ao invés do tamanho absoluto (BORGMAN, 2015).

Assim, uma dezena de planilhas eletrônicas, cada uma com algumas centenas de linhas e colunas de dados, pode ser considerada *big data* para algumas pesquisas, em alguns campos. Porém, o mais comum

é se referir a um grande volume de dados – sempre – acompanhado de um conjunto de soluções tecnológicas para tratá-los. Segundo Borgman (2015), provavelmente, a característica mais importante do *big data* é a viabilização de perguntas usando-se conjuntos de dados infinitos: C = {1, 2, 3, 4, 5, ..., *n*}, em que *n* é infinito. Além disso, pode-se afirmar que os dados que compõem o *big data* são – quase sempre - digitais.

Ciência orientada para dados

A disponibilização, a disseminação e o compartilhamento aberto de dados científicos são alicerces do progresso da ciência. O reuso de dados de pesquisas possibilita a contínua correção, a atualização e a validação de estudos que já haviam sido feitas. A utilização da tecnologia, da nanotecnologia, de técnicas avançadas de engenharia etc. em sensores e instrumentos de análise, simulação e coleta tornou possível a produção de dados em escala exponencial. Esse fato caracteriza o *e-Science*, que proporciona um "conjunto de ferramentas tecnológicas para a coleta e a análise de dados de pesquisa e possibilita que novos enfoques, aplicações, inovações e serviços sejam oferecidos pela ciência moderna" (SAYÃO; SALES, 2014, p. 78). O fenômeno do *e-Science* é responsável por gerar e usar dados em grandes quantidades.

Rotulado de "*data deluge*" – dilúvio de dados - em uma tradução literal, por Hey et al. (2009), o universo dos dados e as tecnologias relacionadas tornaram-se indispensáveis para os procedimentos científicos atuais e criaram um universo em que os dados podem ser coletados, curados, analisados e exibidos em todos os lugares. A ciência aberta (*open science*) faz referência a um modelo de prática científica que, em consonância com o desenvolvimento da cultura digital, visa disponibilizar e compartilhar as informações em redes, de forma contrária à pesquisa fechada dos laboratórios de pesquisas. Essa tendência a proporcionar

livre acesso aos periódicos científicos (digitais) agora se estende aos dados gerados pelas pesquisas científicas.

Reconhecer a importância dos dados digitais para o modelo de ciência atual é uma forma de aperfeiçoar a visão que caracterizava dados de pesquisa, geralmente registrados em mídia impressa ou em formatos digitais, porém esquecidos em armazenamentos locais, como meros subprodutos dos processos de pesquisa. Várias terminologias, como *big Science*, *little Science*, *big data*, *little data* e *no data* (BORGMAN, 2015), entre outras, surgiram de forma concomitante à era do *e-Science* e "possibilitaram a emergência de novos campos de estudo, como a Astroinformática e a Bioinformática. Existem, hoje, disciplinas científicas que são totalmente – em todos os seus ciclos – orientadas por dados" (SAYÃO; SALES, 2014, p. 79). Assim, a colaboração entre cientistas de áreas distintas tornou-se indispensável aos novos domínios do conhecimento.

O ciclo de vida dos dados segundo o *DataONE*

A *Data Observation Network for Earth* (*DataONE*) é um projeto comunitário que disponibiliza o acesso a dados de diversos repositórios parceiros e dá suporte a pesquisas avançadas e à descoberta de dados ambientais e da terra. É um projeto inovador, de plataforma distribuída e com ciberinfraestrutura sustentável que atende às necessidades da ciência e da sociedade para o acesso aberto, persistente, robusto e seguro a dados. A *DataONE* promove um guia de melhores práticas em gerenciamento de dados e é um projeto apoiado pela *National Science Foundation* (*NSF*). Ajuda a preservar, a acessar, a usar e a reutilizar dados científicos multidisciplinares, por intermédio da construção de uma infraestrutura cibernética primária e de um programa de educação e divulgação. Além disso, fornece armazenamento científico para dados ecológicos e

ambientais. O objetivo da *DataONE* é de preservar e possibilitar acesso a dados – compartilhamento – multiescalar, multidisciplinar e multinacional. Os usuários são cientistas, gerentes de ecossistemas, formuladores de políticas, estudantes, educadores, bibliotecários e o público em geral. Vincula a infraestrutura cibernética existente para fornecer um *framework* distribuído, gerenciamento de dados e tecnologias que possibilitem que esses dados sejam preservados em longo prazo.

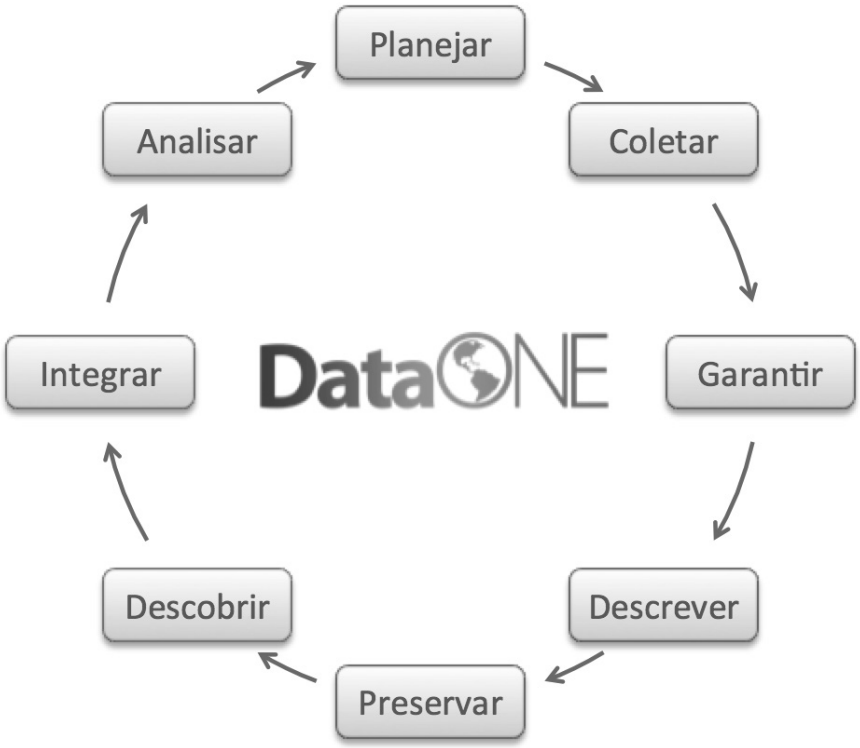
A função essencial do gerenciamento de dados é de apoiar e facilitar o compartilhamento e o reuso dos dados. Feito de forma correta, na visão do pesquisador, pode melhorar a capacidade de recuperar dados, facilitar o acompanhamento dos processos científicos, visando à reprodutibilidade, melhorar as versões de seu controle, controlar sua qualidade de forma mais eficiente, evitar sua perda e aumentar a credibilidade e o reconhecimento das pesquisas.

Para a ciência como um todo, um bom gerenciamento de dados, alinhado a políticas de acesso aberto, é importante porque:

- Aumenta o impacto e a visibilidade da pesquisa;
- Promove a inovação e novos usos potenciais de dados;
- Conduz a novas colaborações entre usuários de dados e criadores;
- Maximiza a transparência e a responsabilidade;
- Possibilita o escrutínio dos resultados da pesquisa;
- Incentiva a melhoria e a validação dos métodos de pesquisa;
- Reduz o custo para duplicar sua coleta;
- Fornece recursos importantes para educação e treinamento.

Com o intuito de alcançar um bom gerenciamento dos dados, a *DataONE* um ciclo de vida que ilustra as etapas pelas quais os dados bem gerenciados passam desde o início de um projeto de pesquisa até sua conclusão. Entretanto, na realidade da pesquisa científica, as etapas, nem sempre, seguem um círculo contínuo. Observe a Figura 1 a seguir:

Figura 1: O ciclo de vida dos dados segundo a *DataONE*



Fonte: *DataONE* (2017, *online*, tradução nossa)

O ciclo de vida dos dados é um processo contínuo de desenvolvimento, manipulação, gerenciamento e estágios de armazenamento de dados. O compartilhamento de dados, por exemplo, deve ser abordado ao longo de todo esse ciclo de vida.

O planejamento do gerenciamento de dados (PGD) é o ponto de partida no ciclo de vida dos dados e se encaixa na etapa "planejar" desse ciclo. No entanto, o plano deve ser revisado frequentemente, ao longo do desenvolvimento do projeto, para garantir que a documentação e o gerenciamento dos dados sejam adequados. O PGD é um documento

formal que descreve o que deve ser feito com os dados durante o desenvolvimento de um projeto de pesquisa e em seu final. Os planos de gerenciamento de dados destinam-se a garantir que os dados sejam preservados e úteis tanto agora quanto no futuro. Já a criação, a inserção e a manipulação dos dados em um formato digital relacionam-se às fases 'planejar' e 'coletar' do ciclo de vida dos dados.

O controle e a garantia da qualidade (CQ/GQ) dos dados são assegurados pelas fases 'coletar' e 'garantir'. O CQ/GQ é uma estratégia que visa prevenir que erros entrem em um conjunto de dados. Essas atividades acontecem em três momentos distintos: antes da coleta dos dados, durante sua inserção e depois que são inseridos. Proteger os dados significa preservá-los e proporcionar sua segurança física, os *backups* e as tecnologias de criptografia. Esse processo envolve as etapas 'garantir' e 'preservar'.

A gestão dos metadados é essencial para os processos de recuperação, compartilhamento e reuso de dados e acontece na etapa 'descrever' do ciclo de vida. A multiplicidade de definições para o termo metadados se deve à diversidade de usos para os quais pode ser empregado. Segundo Borgman (2015), os metadados são informações estruturadas que descrevem, explicam, localizam ou, de alguma forma, facilitam a recuperação, o uso ou o gerenciamento de um recurso.

A citação de dados é parte das fases 'descrever', 'preservar' e 'descobrir' do ciclo de vida e consiste em fornecer uma referência aos dados, da mesma maneira que os pesquisadores rotineiramente fornecem referências bibliográficas. Por fim, a fase 'analisar' contempla toda a gestão de análise de dados e fluxos de trabalho, geralmente conduzida por computadores, modelos matemáticos e estatísticos, entre outras tecnologias.

Tendo em vista que muito do que é estudado sobre dados está aberto à interpretação pessoal (BORGMAN, 2015), as informações a

seu respeito podem ser difíceis de descrever, representar e gerenciar. Na próxima sessão, discorre-se sobre o compartilhamento de dados científicos, um processo abordado ao longo de todo o ciclo de vida dos dados.

Compartilhamento de dados

O compartilhamento de dados não é um tópico novo. Relatos sobre razões e motivações para melhorar o compartilhamento e a curadoria de dados datam, pelo menos, da década de 1980 (BORGMAN, 2012). Tendo em vista sua multiplicidade de significados, o compartilhamento de dados, para o propósito deste artigo, deve ser entendido como a divulgação e a disseminação de dados de pesquisa para uso de outros.

Ao longo da última década, pesquisadores enfrentaram transformações significativas nos ambientes – tecnologia – e nas políticas que regem o compartilhamento de dados (PHAM-KANTER; ZINNER; CAMPBELL, 2014). Uma das mudanças mais significativas foi a criação de uma política de dados pelo *National Institutes of Health* (*NIH*), uma instituição norte-americana de pesquisa e agência de fomento vinculada ao Departamento de Saúde dos EUA em fevereiro de 2003.

A partir do estabelecimento dessa política, nomeada de "*NIH Data Sharing Policy (Final NIH Statement on Sharing Research Data)*" – Política de compartilhamento de dados do NIH (Declaração final da NIH sobre o compartilhamento de dados de pesquisa), em uma tradução literal – o *NIH* passou a exigir de todos os pedidos de financiamento com custos anuais superiores a $500.000,00 a inclusão de planos de compartilhamento de dados (PHAM-KANTER; ZINNER; CAMPBELL, 2014). Vale ressaltar que, desde a criação dessa política de dados, ocorrida, conforme mencionado, em fevereiro de 2003, outras dezesseis já foram

elaboradas por essa mesma instituição, a mais recente datada de agosto de 2016 (NIH, 2017).

Em 2010, a *National Science Foundation* (*NSF)* anunciou que todas as futuras propostas de subsídio exigiriam um plano de gerenciamento de dados, com o intuito de encorajar e facilitar o compartilhamento de dados. Esse PGD, assim como as demais publicações, estaria sujeito à revisão por pares (NSF, 2017). A popularização do *e-Science* e o crescimento do movimento *open data* fizeram com que muitos dos principais *publishers* de conteúdo científico do mundo avançassem na elaboração de políticas de compartilhamento e gerenciamento de dados (NASSI-CALÒ, 2014), como a *PLoS*, a Elsevier, a Springer e a SciELO.

Como dito, o compartilhamento de dados deve ser abordado em todas as fases de seu ciclo de vida, e isso não ocorre ao acaso. Pesquisadores, colaboradores, estudantes, funcionários etc. geralmente dedicam enormes quantidades de trabalho físico e intelectual para coletar, gerenciar e analisar seus dados e publicar seus resultados. Alguns desses dados podem estar em formas compartilháveis, outros, não. Alguns dados são de valor reconhecido para a comunidade, outros não. Alguns pesquisadores desejam compartilhar todos os seus dados, o tempo todo, alguns desejam nunca compartilhar nenhum dos seus dados, e a maioria está disposta a compartilhar apenas parte de seus dados, e não, o tempo todo. Essas perspectivas concorrentes, a variedade de tipos de dados e as origens e a variedade de circunstâncias locais contribuem para dificultar o compartilhamento dos dados (BORGMAN, 2012).

Borgman (2012) enfatiza que, quando as políticas beneficiam os autores ou produtores dos dados, a possibilidade de esses dados serem compartilhados com outros pesquisadores aumenta significativamente. A autora menciona quatro razões para que o compartilhamento dos dados

de pesquisa exista e seja eficiente, a saber: porque reproduz ou verifica a pesquisa, disponibiliza para o público os resultados de pesquisas financiadas com verbas públicas, proporciona a outros pesquisadores a possibilidade de fazer novos questionamentos sobre os mesmos dados e avança nas pesquisas e em inovações.

A reprodutibilidade da pesquisa, vista como "o padrão-ouro" para a ciência, é a razão mais problemática para o compartilhamento de dados de pesquisa. Isso porque, apesar de fundamentalmente orientada, a pesquisa também é vista como um serviço prestado ao bem público. Reproduzir um estudo comprova os resultados e confirma a ciência e, ao fazê-lo, ratifica que os investimentos públicos foram bem empregados. No entanto, o argumento pode ser aplicado apenas a certos tipos de dados e de pesquisa (BORGMAN, 2012).

A segunda razão destacada por Borgman (2012) tem relação com o sentimento público de que dados científicos oriundos ou produzidos por pesquisas financiadas com fundos públicos, por intermédio de governos ou agências de fomento, devem estar disponíveis para o uso geral e não ser reservados a pesquisadores. O período de embargo, nesse caso, não deveria existir.

As duas últimas razões para o compartilhamento de dados científicos são mais focadas na ciência. A terceira discorre sobre o imperativo para que o compartilhamento de dados possibilite outros pesquisadores a fazerem novos e melhores questionamentos. Isso está intimamente ligado à quarta razão mencionada por Borgman (2012), que afirma, categoricamente, que compartilhar dados científicos favorece o avanço da ciência e o surgimento de inovações para a sociedade.

Conforme pesquisa desenvolvida por Costa (2017, p. 159), 47% dos pesquisadores entrevistados, doutores envolvidos com a gestão de dados

científicos no Brasil, não preservam os dados produzidos por suas pesquisas, e 70% demonstraram interesse em compartilhar seus dados de pesquisa. No entanto, apenas dois afirmaram que compartilhariam seus dados com qualquer outro pesquisador. Percebe-se, assim, que existem restrições quanto à adoção do conceito de dados abertos na comunidade científica brasileira.

Por fim, vale salientar que o compartilhamento de dados de pesquisa normalmente – e atualmente – acontece por intermédio do fornecimento direto dos dados de um pesquisador para outro, ou de um grupo de pesquisa para outro. Essa é uma maneira pequena e analógica de se compartilharem dados científicos na era do *e-Science*.

Considerações finais

As razões que fundamentam a promoção do compartilhamento de dados refletem preocupações legítimas das partes interessadas, como a capacidade de reproduzir estudos, disponibilizar recursos públicos, alavancar investimentos em pesquisa e avançar nas pesquisas e em inovações. As discussões acerca do compartilhamento de dados científicos, na era do *e-Science*, sugerem que essas razões, geralmente, levam a políticas genéricas que não refletem a vasta diversidade de dados científicos dentro dos domínios e entre eles, desconsiderando, assim, as particularidades que esses dados carregam e que deveriam ser observadas na criação de políticas de compartilhamento a eles relacionadas.

Pouca atenção é voltada para as motivações científicas – e sociais – que envolvem o compartilhamento, o reuso, ou, até mesmo, o conhecimento sobre a infraestrutura e os valores de investimento necessários para que a gestão de dados ocorra de forma efetiva. Assim, as políticas de dados, principalmente dos *publishers*, parecem estar mais preocupadas com o acúmulo dos dados e com

o incentivo ao *open data* do que com as reais políticas de compartilhamento. Segundo Borgman (2015), a disponibilização, o compartilhamento e o reuso dos dados são mais bem compreendidos quando vistos como problemas de infraestrutura de conhecimento.

Existem mais perguntas do que respostas sobre essa temática:

- Como devem ser tratados os direitos dos autores dos dados científicos?
- Deveria existir alguma compensação financeira caso os dados gerados por uma pesquisa produzissem um produto ou serviço que fosse comercializado como resultado de outra pesquisa?
- Como conseguir financiamento para projetar e desenvolver a infraestrutura necessária para depositar e guardar esses dados?
- Como citar dados utilizados em uma pesquisa quando foram produzidos por outra?
- Onde armazená-los?

Atualmente, no Brasil, as iniciativas de repositórios de dados científicos ainda estão em fase embrionária.

Observa-se que, na era do *e-Science*, o compartilhamento de dados científicos é uma temática ainda pouco explorada, carente de pesquisas que se proponham a estudar, a discutir e a compreender as questões que a envolvem. A tese produzida por Costa (2017) ainda é um dos poucos trabalhos que apresentam resultados de um estudo sobre esse tema no Brasil. Esse é um ponto importante e que deve ser levado em consideração para a realização de trabalhos futuros.

Por fim, considerando que o compartilhamento de dados pode ajudar a diminuir a desigualdade na distribuição dos dados pelos vários campos da ciência, como evidenciado por Borgman (2015, p. 25-26), ao analisar a aplicação da *long tail* nas pesquisas científicas, é preciso entender

quem compartilha e reusa os dados, como, quando, por que e para quais fins.

Referências

ANDERSON, C.. The long tail. *Wired*, jan. 2004. Disponível em: <https://www.wired.com/2004/10/tail/>. Acesso em: 24 jun. 2017.

BORGMAN, C. L.. The conundrum of sharing research data**.** *Journal of the American***.** Society for Information Science and Technology. Maryland, v. 63, n. 6, p. 1059-1078, junho de 2012. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.v63.6/issuetoc>. Acesso em: 24 jun. 2017.

BORGMAN, C. L.. *Big data, little data, no data***:** scholarship in the networked world. Londres, Inglaterra: The MIT Press Cambridge, Massachusetts, 2015.

COSTA, M. M. *Diretrizes para uma política de gestão de dados científicos no Brasil*. 2017. 288 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2017.

DATAONE. *Data observation network for earth*. 2017. Disponível em: <https://www.dataone.org>. Acesso em: 22 jun. 2017.

FACHIN, G. R. B. *Modelo de avaliação para periódicos científicos on-line*: proposta de indicadores bibliográficos e telemáticos. 2002. 210 f. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina, Santa Catarina, 2002. Disponível em:<https://repositorio.ufsc.br/bitstream/handle/123456789/83088/185438.pdf?sequence=>. Acesso em: 11 jul. 2017.

GERHARDT, T. E.; SILVEIRA, D. T. (Orgs.). *Métodos de pesquisa*. Porto Alegre: Editora da UFRGS, 2009.

GIL, A. C.. *Métodos e técnicas de pesquisa social*. 6. ed. São Paulo: Atlas, 2009.

GUEDES, M. das G. T. M.. *Comunicação científica*: o artigo de periódico nas atividades de ensino e pesquisa do docente universitário brasileiro na pós-graduação. 1998. 387 f. Tese (Doutorado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 1998.

KUHN, T. S. *A estrutura das revoluções científicas*. 5. ed. São Paulo: Editora Perspectiva S.A, 1997.

MARCONI, M. A. de.; LAKATOS, E. M. *Técnicas de pesquisa*. 5. ed. São Paulo: Atlas, 2002.

MOLLOY, J. C.. The Open Knowledge Foundation: open data means better science. *PLoS Biology*, Reino Unido, v. 21, n. 12, p. 1-4, dez. 2011. Disponível em: <http://journals.plos.org/plosbiology/article/file?id=10.1371/journal.pbio.1001195&type=printable>. Acesso em: 23 jun. 2017.

NASSI-CALÒ, L. Movimento Open Data se consolida internacionalmente. *SciELO em Perspectiva*, 2014. Disponível em: <http://blog.scielo.org/blog/2014/07/14/movimento-open-data-se-consolida-internacionalmente/>

NIH. *NIH Sharing policies and related guidance on NIH-Funded Research Resources*. 2017. Disponível em: <https://grants.nih.gov/policy/sharing.htm>. Acesso em: 21 jun. 2017.

NSF. NSF National Science Foundation**.** *Where discoveries begin***.** Disponível em: < https://www.nsf.gov/bfa/dias/policy/>. Acesso em: 25 jun. 2017.

PHAM-KANTER, G.; ZINNER, Darren E.; CAMPBELL, E. G.. Codifying collegiality: recent developments in data sharing policy in the life sciences. *PLoS One,* v. 9, n. 9, p. 108451, 2014.

PACKER, A. L. et al. (Org.). SciELO – 15 years of open access: an analytic study of open access and scholarly communication. Paris: UNESCO, 2014.

PRICE, D. J. de S.. *Little science, big science***.** New York: Columbia University Press, 1963.

SAYÃO, L. F.; SALES, L. F.. Dados abertos de pesquisa: ampliando os conceitos de acesso livre. *RECIIS – Rev. Eletron. de Comun. Inf. Inov. Saúde*, v. 8, n. 2, p. 76-92, 2014. Disponível em: <https://www.reciis.icict.fiocruz.br/index.php/reciis/article/viewFile/611/1252>. Acesso em: 22 jun. 2017.

SAYÃO, L. F.; SALES, L. F.. Algumas considerações sobre os repositórios digitais de dados de pesquisa. *Informação & Informação*, Londrina, v. 21, n. 2, p. 90-115, mai./ago., 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939/20122>. Acesso em: 22 jun. 2017.

YANG, X.; WANG, L.; VON LASZEWSKI, G.. Recent research advances in e-science. *Cluster Computing*, v. 12, n. 4, p. 353-356, 2009.

SOBRE OS AUTORES[1]

Adriana Alves Rodrigues

Doutoranda em Ciência da Informação na Universidade Federal da Paraíba (PPGCI/UFPB). Mestre em Comunicação e Cultura Contemporâneas (linha Cibercultura) pela Universidade Federal da Bahia - UFBA e Especialista em Jornalismo Contemporâneo pelo Centro Universitário Jorge Amado - UNIJORGE – Salvador/BA. Graduada em Comunicação Social - Jornalismo pela Universidade Estadual da Paraíba - UEPB. Foi professora substituta no curso de Comunicação Social - Jornalismo, na UEPB e professora dos cursos de Pós-Graduação (Especialização) em: Jornalismo e Convergência Midiática (FSBA/Salvador); Mídias Digitais e Convergência (Fesp/João Pessoa) e Comunicação Digital (Cesrei). Integra o Grupo de Pesquisa em Jornalismo e Mobilidade - MOBJOR. Pesquisa sobre Cibercultura com ênfase em *Big Data*, Ciência da Informação, jornalismo digital, jornalismo de dados, narrativas multimídia, visualização de dados e Ciência de Dados.

Bernardina Maria Juvenal Freire de Oliveira

Doutora em Letras e Mestre em Ciência da Informação, ambos pela UFPB. Professora do Departamento de Ciência da Informação atuando junto aos cursos de graduação em Arquivologia e Biblioteconomia e nas Pós-graduações *Stricto Sensu* em Ciência da Informação, nível de Mestrado e Doutorado e no Programa de Pós-Graduação em Gestão das Organizações Aprendentes. Líder do Grupo de Pesquisa em Cultura, Informação, Memória e Patrimônio (GECIMP) cadastrado junto ao CNPq. A professora é também Membro do Instituto Histórico e Geográfico Paraibano da Cidade de Areia e Presidente da Academia Feminina de Letras e Artes da Paraíba.

1 Informações obtidas diretamente dos autores ou através dos seus respectivos Currículos Lattes.

Flávio Ribeiro Córdula

Doutorando em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba (PPGCI/UFPB). Mestre em Ciência da Informação pela Universidade Federal da Paraíba (2015). Graduado em Ciências da Computação pelo Centro Universitário de João Pessoa (2008). Analista de tecnologia da informação na Superintendência de Tecnologia da Informação da Universidade Federal da Paraíba – STI/UFPB.

Guilherme Ataíde Dias

Graduado em Ciência da Computação (UFPB) e Direito (UNIPE), Mestre em Administração (CCSU/USA), Doutor em Ciência da Informação (USP) e Pós-Doutor (UNESP). Professor Associado IV na Universidade Federal da Paraíba, lotado no Departamento de Ciência da Informação. Membro do Programa de Pós-Graduação em Ciência da Informação, do Programa de Pós-Graduação em Gestão das Organizações Aprendentes (ambos da UFPB) e do Programa de Pós-Graduação – Gestão e Organização do Conhecimento (UFMG). Líder do Grupo de Pesquisa *Web*, Representação do Conhecimento e Ontologias (WRCO) cadastrado junto ao CNPq. Tem interesse de pesquisa nas áreas de Ciência dos Dados, *e-Science*, curadoria digital e propriedade intelectual. Atualmente é bolsista de produtividade em pesquisa (PQ-1D) do CNPq. Audiófilo e ex-ciclista.

Laerte Pereira da Silva Júnior

Doutor em Informação e Comunicação em Plataformas Digitais pela Universidade do Porto e Universidade de Aveiro - Portugal (2017). Mestre em Ciência da Informação pela Universidade Federal da Paraíba (UFPB, 2012), graduado em Tecnologia em Telemática pelo Centro Federal de Educação Tecnológica da Paraíba (2004) e em Licenciatura Plena em Letras pela Universidade Federal da Paraíba (1989). Trabalha como Analista de Tecnologia

da Informação no Centro de Ciências Humanas, Letras e Artes da UFPB. Participa da comissão de trabalho do Repositório Institucional da UFPB. Atua como pesquisador da Rede de Serviços de Preservação Digital do Instituto Brasileiro de Informação em Ciência e Tecnologia. Tem interesse de pesquisa nas áreas de curadoria digital, repositórios institucionais, repositórios de dados científicos e usabilidade na *Web*.

Luana Farias Sales

Doutora em Ciência da Informação pelo Programa de Pós-Graduação do IBICT/UFRJ (2011-2014). Mestre em Ciência da Informação pelo convênio UFF/IBICT (2004-2006), Graduação em Biblioteconomia e Documentação pela Universidade Federal Fluminense (2003). Atuou como Analista em C & T da CNEN, no Instituto de Engenharia Nuclear, participando da criação da linha de pesquisa de Gestão do Conhecimento Nuclear. Atuou ainda como docente do Curso de Graduação em Biblioteconomia da Universidade Federal do Estado do Rio de Janeiro - UNIRIO e da Universidade Federal Fluminense, ministrando disciplinas relacionadas à organização do conhecimento. Atualmente é Analista em C & T do MCTIC/IBICT, atuando como docente do Programa de Pós-Graduação em Ciência da Informação do convênio IBICT/UFRJ e Coordenadora da Rede de Implementação do GOFAIR Brasil. Tem experiência na área de Ciência da Informação, com ênfase em organização e representação do conhecimento e recuperação de informações, atuando principalmente nos seguintes temas: taxonomias, ontologias, vocabulários controlados, tesauros, terminologia e *software* de tesauro. Possui interesse em tópicos ligados à comunicação científica, tecnologia de informação e gestão do conhecimento. Desenvolve pesquisas especificamente nas temáticas de *e-Science*, curadoria digital de dados de pesquisa, biblioteca digital, metadados, repositórios institucionais, repositórios de dados, sistemas CRIS e objetos digitais.

Luís Fernando Sayão

Graduado em Física (UFRJ/IF), Mestre e Doutor em Ciência da Informação (UFRJ/IBICT). Trabalha na Comissão Nacional de Energia Nuclear (CNEN) onde já exerceu os cargos de: chefe do Centro de Informações Nucleares (CIN); chefe da Divisão de Tecnologia da Informação (DITEC); coordenador-geral de Informática; representante do Brasil no INIS - *International Nuclear Information System* (AIEA/ONU); coordenador-geral da RRIAN - *Red Regional de Información en el Área Nuclear*. No Conselho Nacional de Arquivos (CONARQ) foi conselheiro e é membro da Câmara Técnica de Documentos Eletrônicos. Docente do Programa de Pós-Graduação em Biblioteconomia da UNIRIO - Universidade Federal do Estado do Rio de Janeiro e e do Programa de Pós-Graduação em Memória e Acervos da Fundação Casa de Rui Barbosa. Foi membro do Comitê Técnico-Científico do IBICT e da Comissão de Ensino da CNEN. Guia Escalador e velejador.

Pedro Luiz Pizzigatti Corrêa

Possui graduação em Ciência da Computação pela Universidade de São Paulo (1987), Mestrado em Ciência da Computação e Matemática Computacional pela Universidade de São Paulo (1992), Doutorado em Engenharia Elétrica pela Escola Politécnica da Universidade de São Paulo (2002), pós-doutoramento em Data Science na *University of Tennessee* (2015) e Livre Docência pela Universidade de São Paulo (2017). Atualmente é Livre Docente do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo. Tem experiência na área de Ciência da Computação, com ênfase em banco de dados distribuídos, atuando principalmente nos seguintes temas: banco de dados, Ciência dos Dados, modelagem de sistemas computacionais, arquitetura de sistemas distribuídos, computação e biodiversidade, automação agrícola e governo eletrônico.

Plácida Leopoldina Ventura da Costa Santos

Livre-docente em Catalogação pela UNESP (2010), Doutora em Letras - Semiótica e Linguística Geral pela FFLCH/USP (1994), Mestre em Ciência da Informação pela PUC de Campinas (1983) e Bacharel em Biblioteconomia pela UNESP (1980). Docente permanente do Programa de Pós-Graduação em Ciência da Informação da FFC/UNESP, na linha de pesquisa Informação e Tecnologia. Vice-líder do Grupo de Pesquisa Novas Tecnologias em Informação (GP-NTI). Desenvolve suas pesquisas nas temáticas: metadados, catalogação e tecnologias, intersemiose digital, redes de informação, mapa do conhecimento humano. Pesquisadora do CNPq, coordenadora do GT8 - Informação e Tecnologia, da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação - Ancib (2013-2016). Editora da revista Informação & Tecnologia (Itec), membro do corpo editorial das revistas *Brazilian Journal of Information Science*: *research trends* e Revista Eletrônica Informação e Cognição. Parecerista *ad hoc* de agências de fomento e de periódicos científicos, participa como revisora e como membro de Comitês Científicos de periódicos científicos em Ciência da Informação no Brasil e no exterior. Membro da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação - ANCIB e membro da Diretoria da Sociedade Brasileira de Ciência Cognitiva - SBCC.

Renata Lemos dos Anjos

Mestra em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação, na linha de pesquisa Organização, Acesso e Uso da Informação, pela Universidade Federal da Paraíba. Possui Graduação em Biblioteconomia pela Universidade Federal da Paraíba. Tem experiência na área de gestão de dados de pesquisa, ciclo de vida dos dados, bibliotecário de dados.

Ricardo César Gonçalves Sant'Ana

Graduado em Matemática e Pedagogia, Mestrado e Doutorado em Ciência da Informação e Livre-Docente em Sistemas de Informações Gerenciais. Possui especializações em Orientação à Objetos (1996) e Gestão de Sistemas de Informação (1998). Professor Associado no Curso de Administração (FCE) e do Programa de Pós-Graduação em Ciência da Informação (FFC), ambos da UNESP. Presidente da Comissão de Acompanhamento e Avaliação dos cursos de Graduação - CAACG, Coordenador Local do Centro de Estudos e Práticas Pedagógicas – CENEPP. Líder do Grupo de Pesquisa Tecnologias de Acesso a Dados - GPTAD e membro do Grupo de Pesquisa - Novas Tecnologias em Informação GPNTI. Desenvolveu e coordena o projeto Competências Digitais para Agricultura Familiar – CoDAF e seus desdobramentos como a Revista Eletrônica Competências Digitais para Agricultura Familiar - RECoDAF. Atuou no setor privado como consultor, integrador e pesquisador de novas tecnologias informacionais de 1988 a 2004.

Sanderli José da Silva Segundo

Mestre em Ciência da Informação pela Universidade Federal da Paraíba (2018). Bacharel em Biblioteconomia pela Universidade Federal da Paraíba (2013). Especialista em Comunicação e Marketing em Mídias Digitais pela Universidade Estácio, Rio de Janeiro (2017).

Tassyara Onofre de Oliveira

Doutoranda em Ciência da Informação pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Paraíba. Mestra pelo Programa de Pós-Graduação em Gestão nas Organizações Aprendentes, na linha de Pesquisa Gestão de Projetos e Tecnologias Emergentes, pela Universidade Federal da Paraíba. Especialista em Direito Constitucional e Direito Eletrônico, pela Universidade Cândido Mendes com sede no Rio de Janeiro. Possui Graduação

em Ciências Jurídicas, pelo Centro Universitário de João Pessoa. Tem experiência na área de Direito, gestão e tecnologias com ênfase em proteção de dados pessoais no ordenamento jurídico brasileiro, gestão de dados, tecnologias voltadas à ética, responsabilidade social, à gestão da informação e do conhecimento, a democratização, usos e impactos da informação.

Thais Helen do Nascimento Santos

Doutora em Informação e Comunicação em Plataformas Digitais pela Universidade do Porto e Universidade de Aveiro - Portugal (2017). Mestrado em Ciência da Informação pela Universidade Federal da Paraíba (2013). Graduação em Arquivologia pela Universidade Estadual da Paraíba (2010). Professora Adjunta do Departamento de Ciência da Informação da Universidade Federal de Pernambuco. Tem interesse e desenvolve pesquisas junto as seguintes temáticas: Ciência da Informação; Arquivologia; representação e recuperação da informação; acesso e uso da Informação; serviços de informação e fontes de informação.

Wagner Junqueira de Araújo

Doutor em Ciência da Informação pela Universidade de Brasília – UNB (2009); Mestre em Ciência da Informação - UNB (2001) e Bacharel em Ciência da Computação pela Universidade do Oeste Paulista (1993). Professor do Programa de Pós-Graduação em Ciência da Informação – PPGCI/UFPB. Professor do Programa de Pós-Graduação em Gestão nas Organizações Aprendentes - PPGOA/UFPB. Professor Associado - I no Departamento de Ciência da Informação da Universidade Federal da Paraíba - UFPB.

Este livro foi diagramado pela Editora da UFPB em 2019.
Impresso em papel Offset 75 g/m²
e capa em papel Supremo 250 g/m².

Dados científicos são a matéria-prima pela qual os pesquisadores geram novos conhecimentos. Mais recentemente, a contar das últimas décadas do século XX, com o uso ubíquo das tecnologias digitais da informação e comunicação em praticamente todos os domínios da ciência, os dados científicos passaram a ser gerados de forma massiva e contínua. A partir desta avalanche de dados, tornou-se necessária a criação e o aperfeiçoamento de técnicas que possibilitassem a sua correta gestão ao longo das variadas etapas associadas com os diversos ciclos de vida de dados existentes. Este contexto apresenta excelentes possibilidades de trabalho para uma vasta gama de profissionais (bibliotecários, arquivistas, estatísticos, engenheiros de *software* e outros).

A emergência dos dados científicos também possibilita a condução de ações que transcendem o simples fazer técnico e apontam para o desenvolvimento de pesquisas que podem enveredar para os aspectos metodológicos e epistemológicos que orbitam em torno do conceito que define dados. Esta é uma oportunidade de avanço intelectual que deve ser observada por todos os cientistas que possuem interesse na Ciência da Informação e áreas correlatas.

ISBN 978-85-237-1411-6
9 788523 714116