Tendências para a gestão e preservação da informação digital

Organizadores

Miguel Ángel Márdero Arellano
Luiza Martins de Santana Araújo

ibict

ibict

Tendências para a gestão e preservação da informação digital

Tendencias para la gestión y preservación de la información digital

Organizadores

Miguel Ángel Márdero Arellano

Luiza Martins de Santana Araújo

Brasília, DF
2017

Diretoria
Cecília Leite Oliveira

Coordenação Geral de Pesquisa e Desenvolvimento de Novos Produtos (CGNP)
Arthur Fernando Costa

Coordenação Geral de Pesquisa e Manutenção de Produtos Consolidados (CGPC)
Lillian Maria Araújo de Rezende Alvares

Coordenação Geral de Tecnologias de Informação e Informática (CGTI)
Marcos Pereira Novais

Coordenação de Ensino e Pesquisa, Ciência e Tecnologia Da Informação (COEPPE)
Lena Vania Ribeiro Pinheiro

Coordenação de Planejamento, Acompanhamento e Avaliação (COPAV)
José Luis dos Santos Nascimento

Coordenação de Administração (COADM)
Reginaldo de Araújo Silva

Seção de Editoração
Ramón Martins Sodoma da Fonseca

ibict

Tendências para a gestão e preservação da informação digital

Tendencias para la gestión y preservación de la información digital

Organizadores

Miguel Ángel Márdero Arellano

Luiza Martins de Santana Araújo

Brasília, DF

2017

Equipe técnica

Organizador
Miguel Ángel Márdero Arellano

Editor executivo
Ramón Martins Sodoma da Fonseca

Editoras assistentes
Gislaine Russo de Moraes Brito

Normalização de referências
Priscilla Mara Bermudes

Revisão gramatical e visual
Margaret de Palermo Silva

Projeto Gráfico
Seção de Editoração - Sedit/Ibict

Capa
Rodrigo Azevedo

Tradução
Seção de Editoração - Sedit/Ibict

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)
Tendências para a gestão e preservação da informação digital [recurso eletrônico] / organizadores: Miguel Ángel Márdero Arellano, Luiza Martins de Santana Araújo. - Brasília : 2017.
228 p.

ISBN número 978-85-7013-136-2

1. Preservação digital. 2. Gestão da informação. I. Tendências para a gestão e preservação da informação digital.

CDU 025.85(0.034.1)

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Setor de Autarquias Sul (SAUS)
Quadra 05, Lote 06, Bloco H – 5º Andar
Cep: 70070-912 – Brasília, DF
Telefones: 55 (61) 3217-6360 / 55 (61) 3217-6350
www.ibict.br

Rua Lauro Muller, 455 - 4º Andar - Botafogo
Cep: 22290-160 – Rio de Janeiro, RJ
Telefones: 55 (21) 2275-0321
Fax: 55 (21) 2275-3590
http://www.ibict.br/capacitacao-e-ensino/pos-graduacao-em-ciencia-da-informacao
http://www.ppgci.ufrj.br

Comitê Editorial

Sumário

Dados de pesquisa:
o que são, impactos do grande volume produzido, como organizá-los e quais preservar

Maíra Murrieta Costa
Murilo Bastos da Cunha
Sonia de Assis Boeres

RESUMO

Discute aspectos sobre a ciência colaborativa do século XXI, a internacionalização e a virtualização da ciência que culminaram com a explosão de dados de pesquisa coletados on-line, dando origem ao fenômeno de big data e cyberinfrastructure, também denominado e-Science. Expõe os fatos que trouxeram à tona a ciência de uso intensivo de dados de pesquisa, dentre eles o desenvolvimento tecnológico dos instrumentos de coleta e análise de dados. Explica conceitualmente o que é e-Science e cyberinfraestructure. Ao mesmo tempo, apresenta os termos dados de pesquisa e dados científicos e argumenta que, por se tratar de um tema novo, ainda não há consenso na literatura sobre qual termo deve ser utilizado. Externa o conceito de big data, e como ele trabalha com a utilização de dados de redes sociais para o desenvolvimento de aplicativos. Evidencia a e-Science como uma parte do big data, que lida com dados em larga escala no âmbito científico. Procura definir o termo dado para então apresentar a expressão dados de pesquisa, suas peculiaridades, formas de coleta, tratamento e preservação. Discorre sobre os aspectos de armazenamento e preservação digital dos dados de pesquisa, abordando aspectos de infraestrutura tecnológica. Finaliza o capítulo com a apresentação de reflexões sobre a gestão de dados de pesquisa no Brasil.

Palavras-Chave: *Big data*. Ciência orientada a dados. Curadoria de dados. *Cyberinfrastructure*. Dado de pesquisa. *E-Science*. Gestão de dados de pesquisa. Preservação de dados de pesquisa.

INTRODUÇÃO

O que é a ciência orientada ao uso intensivo de dados? O que é *big data*? O que o *big data* e a denominada *e-Science* têm em comum? De maneira didática, este capítulo procura, por meio de alguns conceitos introdutórios, responder às questões acima mencionadas para o leitor. Também procura elucidar como se dá o uso das tecnologias da informação para a produção de grandes volumes de dados, seja no aspecto comercial (*big data*), ou no universo científico (*e-Science*).

Dentre os objetivos deste capítulo está o de apresentar o que é a *e-Science/ cyberinfraestructure*, bem como o que são os dados de pesquisa, por vezes denominados dados científicos ou dados de pesquisa. Além disso, argumenta-se que a terminologia na área, por estar em plena ebulição, ainda não está completamente consolidada. Por fim, ele traz reflexões sobre como o Brasil está se preparando para gerenciar os dados de pesquisa gerados nas universidades, nos institutos e nos seus centros de pesquisa, a fim de dar-lhes acesso e garantir-lhes a preservação de longo prazo. Os desafios são muitos, principalmente quando se compreende a variedade de dados produzidos pelas pesquisas, as nuances de cada campo, o comportamento das áreas mais internacionalizadas, como, por exemplo, a energia nuclear, a biodiversidade ou mesmo a área espacial.

O texto pretende trazer contribuições para pesquisadores que objetivem realizar a gestão de seus dados, assim como para profissionais da informação que já começam a se deparar com esse recente desafio. Além disso, traz reflexões para os profissionais de agências de fomento, associações de pesquisa e, ainda, para estudantes e pesquisadores que querem se preparar para essa nova realidade.

Por se tratar de tema extremamente novo, é relevante que o leitor não confunda conceitualmente o que é e-Science/cyberinfrastructure e o que são os dados de pesquisa, algumas vezes também denominados dados científicos.

E-Science e *cyberinfrastructure* são termos guarda-chuva que se referem à infraestrutura tecnológica necessária para apoiar a pesquisa científica do século XXI, como, por exemplo, a computação em grid1 e bancos de dados que suportem petabytes de dados não estruturados, com fluxo constante.

Os dados de pesquisa e/ou científicos, por sua vez, são aqueles coletados em grande volume, por sensores, telescópios, satélites, dentre outros instrumentos e que exigem a infraestrutura tecnológica já comentada para processamento e análise.

Isto posto, cabe ressaltar que ainda não há um consenso na literatura quanto ao uso da expressão *dados científicos* ou *dados de pesquisa*. Os autores Hey e Hey (2006), Bell (2011) e Rodrigues et al. (2010) utilizam dados científicos. Por outro lado, Borgman (2015), Sales (2014), Sayão e Sales (2014) utilizam o termo dados de pesquisa (*data scholarship*).

Também não há consenso quanto ao uso do termo *e-Science* ou *cyberinfrastructure*. Tal situação já havia sido observada na indexação das bases de dados Library and Information Science Abstracts (LISA) e Library and Information Science & Technology Abstracts (LISTA), em estudo bibliométrico sobre a literatura referente ao tema, realizado por Costa e Cunha (2015). Nesse estudo, os autores observaram que o termo *e-Science* foi mais utilizado como indexador, em detrimento do termo *cyberinfrastructure*.

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

A compreensão da ciência com o uso intensivo de dados de pesquisa produzidos em larga escala por sensores especializados, sem dúvida, perpassa o entendimento das origens da ciência moderna. Suas origens, por sua vez, se encontram na Inglaterra do século XVII (ALFONSO-GOLDFARB, 1994; SOLLA PRICE, 1976). Nesse período, a ciência não precisava de grandes justificativas. Quando sofria ataques, sua resposta estava sempre voltada para o futuro, e não para o passado.

A quantidade de publicações em cada campo do conhecimento, após a Segunda Guerra Mundial, cresceu exponencialmente, duplicando a cada dez ou quinze anos (SOLLA PRICE, 1976). Esse fenômeno deu origem à chamada *Big Science*[2]. Solla Price (1976) teorizou sobre a pequena ciência e a grande ciência (*little science* e *big science*), argumentando que a transição de uma para a outra foi gradual. O autor (1976, p. 3) defendeu a ideia de que "se um seguimento suficientemente amplo da ciência for medido de alguma forma razoável, o modo normal de crescimento é exponencial".

A literatura científica revela que a evolução da ciência está altamente relacionada com o aprimoramento do instrumental tecnológico, que permitiu a realização de observações de diversos fenômenos. Para Bell (2011), as teorias científicas do século XX foram baseadas em dados geralmente disponíveis em cadernos científicos pessoais. Já no início do século XXI, emergiu, de modo crescente, uma questão: os dados oriundos de pesquisas são coletados por meio de sensores especializados, telescópios, satélites e ensaios de laboratórios. Há autores, como Green (2011), Fox e Hendler (2011), que destacam a transformação pela qual passará a pesquisa científica em razão da criação e disponibilidade de grande volume de dados *on-line*.

O desenvolvimento tecnológico de instrumentos de coleta de dados em larga escala e a facilidade de troca de informações sobre determinada pesquisa, por meio da internet, permitiu que cientistas interagissem mais *on-line*. Nas palavras de Castells (2003) a internet tornou-se a espinha dorsal da sociedade contemporânea, "a base tecnológica para a forma organizacional da era da informação – a rede" (CASTELLS, 2003, p. 7).

[2] Representa um momento da ciência após as grandes guerras, marcado pelo alto grau de investimentos em C&T.

Esse conjunto de fatores permitiu o desenvolvimento de uma ciência colaborativa *on-line*, que produz dados em larga escala. A ciência colaborativa tem como um de seus marcos iniciais o Projeto Genoma Humano. Outros bons exemplos de colaboração são: a) as iniciativas do European Organization for Nuclear Research (CERN) para descobrir a partícula da vida – uma partícula subatômica que poderia ser o bóson de Higgs; b) o Projeto Netuno do Observatório Oceânico EUA-Canadá; c) o Projeto de Celeste Digital Sloan, dentre outros. Todos esses projetos têm em comum o enorme volume de dados coletados por sensores especializados.

Tapscott e Williams (2007) consideram que o Projeto Genoma Humano representou um divisor de águas. Afinal, as indústrias farmacêuticas pararam com as suas tentativas isoladas de mapear o genoma e passaram a apoiar colaborações abertas (*open-science*). A experiência desse projeto representa o resultado final de forte concentração de esforços públicos e privados em prol da informação genética do ser humano.

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

A pesquisa colaborativa, produzida por uma equipe multidisciplinar, que coleta grande quantidade de dados, em diversos lugares, fomentou o chamado dilúvio de dados. É nesse cenário que surgem os fenômenos de *big data* e da *e-Science*.

O dilúvio de dados, quando abordado na perspectiva de coleta de dados científicos por sensores, telescópios, radares, satélites, dentre outros, originou a *e-Science*, também denominada *cyberinfrastructure*.

A contemporaneidade do tema traz à tona questões conceituais que ainda não passaram pelo processo de reflexão necessário ao seu amadurecimento. Por exemplo, merece ser comentado que dentre as denominações utilizadas para *e-Science*, também se destacam na literatura os termos ciência orientada por dados (*data-driven science*), computação fortemente orientada a dados (*data-intensive computing*), ciberinfraestrutura (*cyberinfrastructure*), ciência com

uso intensivo em dados, quarto paradigma da ciência (*fourth paradigma of science*), dentre outros (ALVARO et al., 2011; CESAR JÚNIOR, 2011; MARCUM, GEORGE, 2010; GRAY, 2007; HEY, TREFETHEN, 2003).

A diferença básica entre o *big data* e a *e-Science* parece estar no fato de a *e-Science* tratar de grande volume de dados no âmbito científico. Poder-se-ia dizer que a *e-Science* é um aspecto particular do *big data* que, por sua vez, está voltado para o âmbito comercial do uso de grandes volumes de dados.

Ao retomar as questões sobre a evolução da ciência, tem-se que nos primórdios da ciência moderna a sociedade se preocupou com o armazenamento dos dados de pesquisa primários, registrados em cadernos pessoais, bem como com a preservação dos resultados das pesquisas, publicados em artigos de periódicos e livros. As atividades de armazenamento e preservação foram exercidas com primazia pelas bibliotecas. O momento atual é propício para a sociedade se preocupar com a gestão dos dados digitais e, consequentemente, com a preservação dos dados de pesquisa coletados *on-line*, a fim de garantir o acesso às futuras gerações de pesquisadores.

O DILÚVIO DE DADOS: *BIG DATA*, *E-SCIENCE*, DADOS DO GOVERNO E DADOS ABERTOS

A pesquisa colaborativa presente no século XXI é descrita como aquela que tem a "capacidade de gerar e armazenar dados em uma escala sem precedentes e muito além da capacidade humana de análise" (CESAR JÚNIOR, 2011). Suas características deram origem aos termos *big data* e *e-Science*.

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

Mayer-Schonbeerger e Cukier (2013) exemplificam o conceito de *big data* lembrando o surgimento do vírus H1N1 em 2009. Os autores relatam que pesquisadores da empresa Google analisaram os 50 milhões de termos de busca mais comuns entre os americanos e os compararam com a lista do Centers of Disease Control (CDC). A pesquisa nos termos de busca utilizados no *search engine* Google revelou o local onde o vírus estava se espalhando com mais velocidade que o sistema de informações do CDC.

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

São exemplos desses dados os *posts* das redes sociais, sejam elas Facebook, Twitter ou algum outro aplicativo social. Os dados postados e coletados podem ser pela tecnologia de RFID, dados de localização geográfica de um usuário de aplicativo de mapas da empresa Google – disponibilizados na rede por meio do seu telefone celular ou do aparelho GPS do automóvel, dados de compras *on-line* realizadas com cartão de crédito, dados dos programas de televisão e filmes assistidos na *smart* TV por meio do Netflix ou Youtube, dentre tantos outros exemplos. Esses dados podem ser utilizados em benefício de políticas públicas na área de saúde e educação. Também têm aplicação no conceito de *smart cities* e têm sido frequentemente utilizados por empresas de comércio eletrônico para aprimorar suas estratégias de vendas.

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

constante (*streaming data*[3]). Tecnicamente, o autor procura clarificar a diferença entre os conceitos do *big data* e os conceitos do *analytics tradicional*, conforme exposto no quadro 1.

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

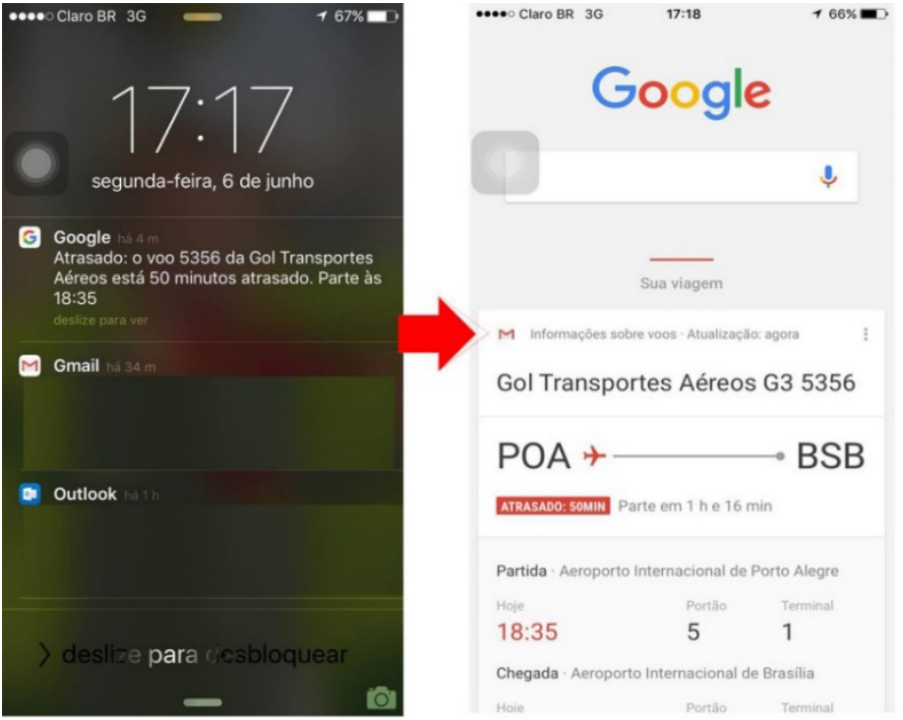
	Big data	**Analytics tradicional**
Tipos de dados	Formatos não estruturados	Dados formatados em linhas e colunas
Volume de dados	100 terabytes a petabytes	Dezenas de terabytes ou menos
Fluxo de dados	Fluxo constante de dados	Pool estático de dados
Métodos de análise	Aprendizado de máquina	Baseados em hipóteses
Objetivo principal	Produtos baseados em dados	Suporte ao processo decisório

Fonte: Davenport (2014, p. 4)

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

[3] São dados gerados por inúmeras fontes de dados de forma contínua, que geralmente enviam os registros de dados simultaneamente. Possuem ampla variedade de dados, como, por exemplo, arquivos de log gerados por clientes usando seus aplicativos móveis ou da web, compras de e-commerce, atividade de jogador durante o jogo, informações de redes sociais, pregões financeiros ou serviços geoespaciais, como também telemetria de serviços conectados ou instrumentação em datacenters.

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



Fonte: Costa (2016).

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

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

Outro ponto polêmico, onde o *big data* mostra seu lado perverso, refere-se à utilização de dados pessoais em larga escala em prol da segurança nacional de um país. O exemplo de maior repercussão, até 2016, foi o de Edward Snowden, que revelou o programa de vigilância da National Security Agency (NSA). Snowden revelou que o programa acessava vários tipos de dados[4] de usuários dos serviços de acesso à internet fornecidos pelas empresas AOL, Apple, Facebook, Google, Microsoft, Paltalk, Skype, Yahoo! e YouTube. As empresas negaram que tenham oferecido acesso aos dados para o governo americano.

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

A ironia é que em apenas 12 anos da data de lançamento, a ficção se tornou realidade, ou seja, muito antes do cenário de 2054 relatado no filme. A única diferença é que a divisão pré-crime do filme determinava suas ações por meio de *um possível futuro* visualizado pelos paranormais e clarividentes *precogs*. Em contrapartida, a realidade de 2014 é a possibilidade de prever comportamentos a partir do dilúvio de dados disponibilizados *on-line* pelo próprio usuário em seu *post* ou *tweet*.

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

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

Já no que diz respeito ao grande volume de dados produzidos no âmbito científico, conforme mencionado no primeiro tópico deste capítulo, há vários termos sendo utilizados para tratar as mudanças ocorridas na condução da ciência contemporânea. Dentre esses termos, destacam-se *e-Science* e *cyberinfrastructure*. Aparentemente surgem como termos sinônimos, mas em países com iniciativas diferentes no tratamento do grande volume de dados científicos *on-line*. Por esse motivo, faz-se necessário contextualizar o surgimento de ambos.

Para Jankowski (2007), o termo *cyberinfrastructure* está extremamente relacionado às iniciativas dos cientistas americanos, em 2003, de obter patrocínio da National Science Foundation. Essa iniciativa resultou na publicação do Atkins Report. Nas palavras do relatório: "(...) se a infraestrutura é necessária para a economia industrial, então a cyberinfrastructure é necessária para a economia do conhecimento" (ATKINS, 2003, p. 5).

Já o termo *e-Science* surge de iniciativas europeias, especialmente no Reino Unido, onde John Taylor – diretor geral do Escritório de Ciência e Tecnologia do Reino Unido cunha o termo em 1999[5], durante o lançamento de um programa de financiamento para pesquisas (JANKOWSKI, 2007). Na perspectiva de Hey e Trefethen (2003), Marcum e George (2010) e Gray (2007) a *e-Science* faz referência à coleção de instrumentos e tecnologias necessárias para apoiar a pesquisa científica do século XXI, e amparar o grande volume de dados produzidos que precisam estar em rede, com a característica intrínseca da colaboração e da multidisciplinaridade.

Para Gray[6] (2007), a *e-Science* é o ponto onde a tecnologia da informação encontra os cientistas. Ele explica que a coleta de dados de pesquisa é realizada por instrumentos (satélites, telescópios, sensores) ou é gerada por máquinas de simulação. Os dados capturados, ou obtidos por meio de simulação, são processados por um *software*, que providenciará o armazenamento da

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

[6] Jim Gray foi vencedor do Prêmio Turing de 1998. É considerado um dos pioneiros em aplicações e técnicas computacionais para o tratamento de grandes quantidades de dados gerados por cientistas de outras áreas. FONTE: CORDEIRO, D.; BRAGHETTO, K.R.; GOLDMAN, A.; KON, F. Da ciência à e-ciência: paradigmas da descoberta do conhecimento. Revista USP, n. 97, p. 71-80, março 2013.

informação em bancos de dados. Ao comentar que um telescópio é operado por 20 a 50 pessoas e que há milhares de pessoas escrevendo códigos para lidar com a informação coletada pelo instrumento, ele utiliza o campo da astronomia para defender a sua tese.

Na perspectiva de Jankowski (2007, p. 549), *e-Science* é um termo guarda-chuva, utilizado para as iniciativas de computação em *grid*, a colaboração global de pesquisadores e internet baseada em instrumentos. São esses dados, produzidos por esses instrumentos, que precisam passar por um processo de curadoria, armazenamento, divulgação, reutilização e preservação digital. Falaremos sobre as características desses dados no próximo tópico.

Em função do resultado do estudo bibliométrico sobre o tema, realizado por Costa e Cunha (2015), optou-se por utilizar, neste trabalho, o termo *e-Science* em detrimento dos demais. Cabe ressaltar que os autores não identificaram nas bases de dados LISA e LISTA o termo *cyberinfraestructure* indexado na ocasião em que os metadados[7] da base foram analisados. Por fim, os autores inferem que o termo criado no Reino Unido (*e-Science*) tenha ganhado mais adeptos, levando-o a ser um termo indexador.

No que diz respeito ao grande volume de dados produzidos por instituições, no caso do Brasil, não podemos nos esquecer das principais bases de dados do governo, como, por exemplo, as produzidas pelo IBGE, as coletadas pelo Datasus, as recolhidas e geradas pelo Ipea, ou mesmo os dados financeiros do governo federal disponíveis no Siafi[8]. Dentre esses dados, alguns classificam-se como dados abertos e serão abordados a seguir.

Ao analisar a questão dos dados produzidos pelo governo, Sayão e Sales (2015, p. 9) defendem que "embora estes dados não tenham sido originalmente coletados para fins de pesquisa, eles se tornam dados de pesquisa uma vez que tenham sido modificados ou expandidos". Os autores observam que a

[7] Os dados foram coletados nas bases de dados LISA e LISTA entre o período 19/03/2013 a 19/06/2013.

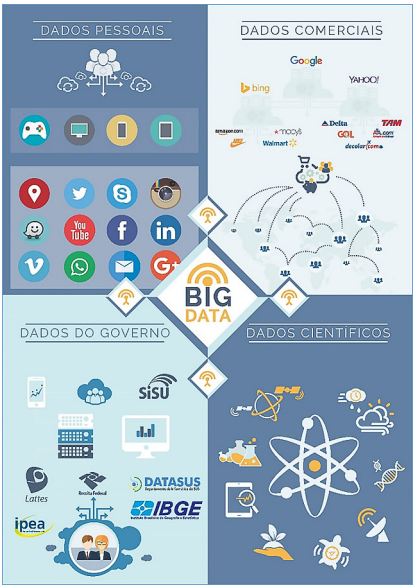
[8] O Sistema Integrado de Administração Financeira do Governo Federal consiste no principal instrumento utilizado para registro, acompanhamento e controle da execução orçamentária, financeira e patrimonial do governo federal. Disponível para acesso em < http://www.tesouro.fazenda.gov.br/siafi> .

partir do momento que os dados produzidos pelo governo são utilizados por alguma área de pesquisa e sofrem alguma modificação, eles passam a ser dado de pesquisa.

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

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

Figura 2 – Aspectos conceituais do *big data*.



Fonte: Costa (2016).

OS DADOS DE PESQUISA

De acordo com Borgman (2015, p. 4),

> a questão não declarada a fazer é: o que são dados" [grifo da autora]. Para a autora, "o único consenso sobre as diferentes definições é que nenhuma definição única será suficiente para definir o termo, uma vez que eles têm muitos tipos de valor, sendo que valor dos dados pode não ser aparente até muito tempo depois dos mesmos terem sido coletados, criados ou mesmo perdidos.

Em continuidade ao assunto, a autora defende que valor dos dados varia muito ao longo do tempo, lugar e contexto. Além disso, enfatiza que ter os dados corretos é geralmente melhor do que ter mais dados. Entretanto, é importante destacar que os dados não têm nenhum valor ou significado quando estão isolados.

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

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

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

Aproximadamente seis anos depois, com a explosão de dados produzidos e transmitidos por máquinas nos ambientes de pesquisa, Gray (2007, p. 35) propõe uma nova pirâmide informacional, trazendo mais uma vez o *dado*

na base da pirâmide e a literatura no topo. A questão é que Gray tem como projeto deixar todos os dados de pesquisa *on-line*, para assim contribuir com o desenvolvimento da ciência de jeito mais célere. A proposta de Gray (2007) é ilustrada na figura 3.

Figura 3 – Todos os dados científicos *on-line*.

Fonte: Gray (2007, p. 25).

Na perspectiva de Borgman (2015, p. 17), pela sua complexidade, o termo dado, por si só, é digno de um livro. A autora defende que "a questão o que é dado é melhor abordada como quando são dados " fundamentando-se em definições sobre o termo no *Dicionário de Oxford de* 1646**,** que traz o uso da palavra na teologia, bem como no estudo de Rosenberg, sobre o termo dado, no século XVIII. Além disso, a autora relembra que diversos[9] autores da ciência da informação já discutiram sobre o fato de o dado ser uma forma de informação.

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

[9] Blair (2010), Brown e Duguid (2009), Burke (2000, 2012), Day (2001), Ingwersen e Javelin (2005), Liu (2004), Meadows (2001), Buckland (1991) e outros autores.

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

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

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

Em continuidade ao assunto, Simberloff et al. (2005, p. 19): argumentam que

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

Já no que diz respeito aos dados experimentais, Simberloff et al. (2005, p. 19) defendem que "em princípio os dados de experimentos, que podem ser reproduzidos com precisão, não precisam ser armazenados por tempo indeterminado". Porém, os autores revelam que, na prática, pode não ser possível reproduzir com precisão todas as condições experimentais, particularmente quando algumas condições e variáveis não podem ser conhecidas. Além disso, há situações em que os custos de reprodução da experiência são proibitivos e nestes casos, em específico, a preservação de longo prazo deve ser garantida para essa categoria de dados. Em síntese, Simberloff et al. (2005) ponderam que as questões de custo e a capacidade de reprodutibilidade são a chave ao considerar-se políticas para a preservação de dados experimentais.

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

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

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

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

Em seu guia de pesquisa, Sayão e Sales (2015) propõem como formas de classificação de dados: **a) quanto à sua origem** (observacionais, computacionais e experimentais), **b) quanto à sua natureza** e **c) quanto à fase da sua pesquisa**. A proposição desses autores assemelha-se à proposta

de Simberloff (2005) nos itens **a** e **b**. E ainda trazem uma proposta nova ao proporem uma classificação quanto à fase da pesquisa (dados brutos, crus ou preliminares; dados derivados; dados canônicos ou referenciais).

Do ponto de vista prático, pode-se dizer que o sensor que está implantado nas tartarugas do Projeto Tamar[10], ou mesmo o sensor que estava implantado no Leão Cecil[11] geram dados de biodiversidade. E quando se trata de informação georreferenciada, pode-se ter a latitude e a longitude indicando a posição de uma espécie de bromélia rara na Floresta Amazônica. Em ambos os casos os dados são armazenados em grandes bancos de dados.

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

É preciso explorar semelhanças e diferenças na maneira como os dados são criados, utilizados e compreendidos nas comunidades acadêmicas (BORGMAN, 2015). A partir da observação da autora, é pertinente pensar no caso de um dado, como, por exemplo, da tartaruga do Projeto Tamar coletado pelo sensor. Quem é o autor desse dado? É o líder do projeto de pesquisa? É o pesquisador responsável pelo monitoramento daquela tartaruga específica? Como a comunidade acadêmica entende esses dados? Identificar o autor do dado traz à tona a resposta de como citar o dado. Identificar o modo como o dado foi gerado permite classificá-lo de acordo com as propostas já existentes, como, por exemplo, a de Simberloff et al. (2005). Ao classificar esse dado, por consequência, sabe-se o período de temporalidade dele no

[10] Um projeto de 35 anos que representa uma das mais bem-sucedidas experiências de conservação marinha desenvolvidas no Brasil e serve de modelo para outros países.

[11] Leão africano que vivia no Parque Nacional de Hwange localizado no Zimbábue. Era monitorado por cientistas da Universidade de Oxford, no Reino Unido, que estudavam a longevidade e a conservação de leões no Zimbábue. O leão foi morto, aos 13 anos de idade, no ano de 2015, por turista americano em caçada de lazer, abrindo a discussão sobre esta prática e a sobrevivência de animais selvagens.

repositório de dados. Essa cadeia de atividades alimenta o ciclo de gestão de dados de pesquisa que será abordado a seguir, no tópico sobre preservação dos dados. Outras reflexões que impactam na gestão de dados de pesquisa são ilustradas na figura 4.

Figura 4 – Reflexões sobre a gestão de dados de pesquisa.

Fonte: Costa (2016).

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

A PRESERVAÇÃO DOS DADOS

Tratar o objeto digital, dentre tantas atividades, implica viabilizar a sua preservação em longo prazo (*long-term-preservation*). Garantir essa preservação, por sua vez, envolve vários aspectos, pois o objeto digital pode sofrer, ao longo de sua vida, várias alterações. Essas alterações não devem impedir que *hardware* e *software*, no futuro, possam transformar os dados armazenados em informação legível para o usuário.

Um marco internacional no contexto da preservação digital (PD) foi a carta submetida a 32ª Sessão da Conferência Geral da Unesco, que ocorreu em Paris em 2003. Na ocasião, a Unesco apresentou, durante a conclusão da conferência**,** uma carta que ficou conhecida como a *Carta sobre a Preservação Digital*[12], que trouxe à tona conceitos sobre a importância de estabelecer princípios para a preservação e contínua acessibilidade ao patrimônio digital mundial.

Merece destaque o fato de que a Unesco justificou a importância da PD devido aos "recursos culturais, educacionais, científicos, públicos e administrativos e a informação técnica e médica estarem cada vez mais sendo produzidas, distribuídos e acedidos apenas em formato digital" (UNESCO, 2003). Levando-se em conta que a informação digital está sujeita a suscetibilidades de cair em desuso e decadência física, a carta teve como objetivo levantar um manifesto em prol de um compromisso de longo tempo**,** para assegurar o contínuo acesso aos conteúdos e à funcionalidade dos objetos digitais.

Segundo a Carta da Unesco (2003, p. 2-4), o desaparecimento do patrimônio, não importa em que forma esteja, é um empobrecimento das nações. Para a entidade, o patrimônio digital são recursos de informação e expressão criativa produzidos, distribuídos, acessados e mantidos em forma digital, e sua preservação é um benefício para a presente e para as futuras gerações.

[12] A Carta sobre a Preservação Digital foi publicada em 2003 pela Unesco.

A organização sem fins lucrativos Portico (2015), a maior comunidade mundial de arquivos digitais, que provê serviços de preservação digital, em seu *site*[13], apresenta a definição de PD como uma série de políticas e atividades de gestão necessárias para assegurar a duradoura usabilidade, autenticidade, descoberta e acessibilidade dos conteúdos em longo prazo.

Em consonância com a proposta, Rosenthal et al. (2005) argumentam que o objetivo do sistema de preservação digital é que a informação que ele contém permaneça acessível ao usuário por longo do tempo.

Márdero Arellano (2008) trouxe novas contribuições ao apresentar três pontos básicos para a preservação digital, sendo eles: autenticidade, confiabilidade e integridade. O autor resumiu como sendo a autenticidade dos dados a certeza de quem é seu criador (p. 135); já a confiabilidade é ligada à certificação e segurança dos dados digitais (p. 277). A integridade, por sua vez, representa a inteireza/confiabilidade do conteúdo, representada pela não alteração ou modificação para permitir o acesso continuado.

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

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

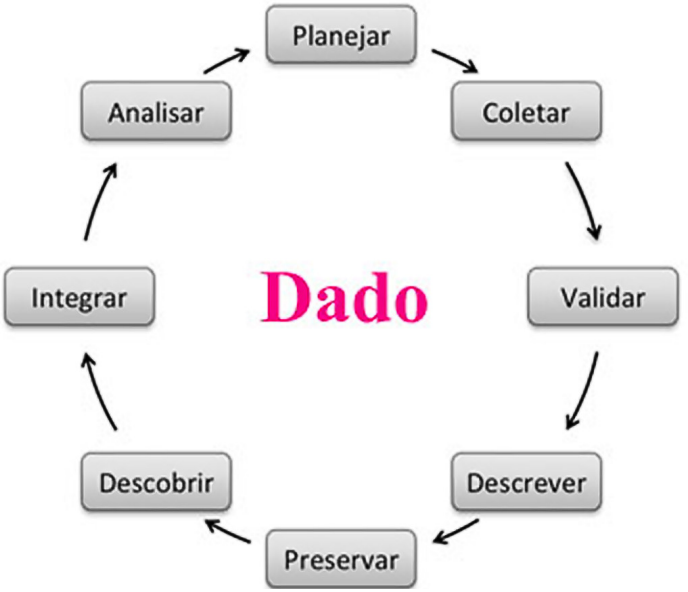
[13] Site - http://www.portico.org/.

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

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

Nos Estados Unidos, o Projeto Data Observation Network for Earth (DataONE), da National Science Foundation, com sede em Albuquerque, Novo México, tem envidado esforços para a preservação digital de dados de pesquisa. O projeto DataONE tem uma missão ambiciosa: "Fornecer o acesso universal aos dados sobre a vida na Terra e o ambiente que o sustenta [o acesso], bem como as ferramentas que os pesquisadores necessitam para tanto". Assim, o DataONE tem desenvolvido um framework distribuído e uma ciberinfraestrutura sustentável que atenda às necessidades da ciência aberta. A iniciativa vai ao encontro do movimento de ciência aberta e acata a diretriz do governo americano de aumentar o acesso aos resultados da investigação científica financiada pelo governo federal, conforme ilustra a figura 5.

Figura 5 – Ciclo de vida do dado na perspectiva do pesquisador.



Fonte: DataONE Project[14] (2016).

Na visão do DataONE, o dado tem vida própria. A figura 5 ilustra as etapas de sua criação e utilização. A gestão do dado começa quando o pesquisador ainda está planejando sua etapa de coleta. Os próximos três estágios (coletar, validar, descrever) são a base para o acesso do dado em longo prazo. Enquanto isso, os três últimos representam a descoberta e o uso dos dados.

A filosofia do modelo DataONE parte da pergunta – "se você compartilhar seus dados com um cientista ou colega que não está envolvido com seu projeto de pesquisa, eles estarão aptos a ver sentido nos dados? Será que eles vão ser capazes de usá-los de forma eficaz e adequadamente? "

Nesse sentido, o Projeto DataONE divulgou uma cartilha – *Primer on Data Management*: *What you Always wanted to Know*. A cartilha descreve algumas práticas de gestão de dados fundamentais, trazendo contribuições para

[14] Melhores Práticas - https://www.dataone.org/best-practices.

desenvolver-se um plano de gestão de dados, bem como sugestões para se criar o dado de modo eficaz, organizá-lo, gerenciá-lo, descrevê-lo, preservá-lo e compartilhá-lo, conforme retratado no quadro 2.

Quadro 2 –Visão geral do ciclo de vida DataONE

Atividade	Descrição
Plan	Fase de descrição dos dados que serão compilados, e como os dados serão administrados e tornados acessíveis ao longo da sua vida útil.
Collect	Observações são feitas à mão, ou com sensores, ou outros instrumentos, e os dados são colocados em um formato digital.
Assure	A qualidade dos dados é assegurada por meio de controles e inspeções.
Describe	Os dados são descritos com precisão e são usados os padrões de metadados apropriados.
Preserve	Os dados são submetidos a um arquivamento de longo prazo adequado.
Discover	Dados potencialmente úteis estão localizados e são obtidos junto com as informações relevantes sobre os dados.
Integrate	Dados de fontes diferentes são combinados para formar um conjunto homogêneo de dados que podem ser facilmente analisados.
Analyze	Os dados são analisados.

Fonte: Strasser et al. (2012).

Além do Projeto DataONE, merece ser comentado que na visão de Borgman (2015, p. 20), entre os princípios mais conhecidos para arquivamento de dados tem-se o documento Reference Model for a Open Archival Information System15 (OAIS). A autora comenta que esse documento apresenta um consenso sobre a prática originada na comunidade de Ciências Espaciais para tratamento e arquivamento de dados. A autora observa que essas orientações também têm sido amplamente adotadas nas ciências e ciências sociais como diretrizes para o arquivamento de dados.

[15] Consultative Committee for Space Data Systems.

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

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

Borgman (2015, p. 21) defende que "entre as categorias mais discretas dos dados estão os níveis de processamento definidos pelo Sistema de Informação de Dados sobre a Terra da Nasa". Nesse sistema, dados com uma origem comum se distinguem pela forma como eles são tratados, conforme demonstra o quadro 3. De acordo com a Nasa (2016)

> Produtos de dados da EOSDIS[16] são processados em diversos níveis, variando do **Nível 0** ao **Nível 4**. Os produtos de **Nível 0** são dados brutos na maior resolução do instrumento. Em níveis mais elevados, os dados são convertidos em parâmetros e formatos mais úteis. Todos os instrumentos da EOS devem gerar produtos de **Nível 1**. A maior parte gera produtos de **Nível 2** e **3**, e muitos geram produtos de **Nível 4**.

[16] Earth Observing System Data and Information System – em português - Sistema de Informação de Dados sobre a Terra.

Quadro 03 – Níveis de processamento de dados[17]

Nível do Dado	Descrição
Nível 0	Dados de instrumentos e de carga em resolução total, reconstruídos e não processados, com qualquer e todos os artefatos de comunicação removidos (por exemplo, quadros de sincronização, cabeçalhos de comunicação, dados duplicados). Na maioria dos casos, o Sistema de Operação de Dados EOS (EDOS) fornece esses dados para os Data Centers como conjuntos de dados de produção para processamento pelo Departamento de Ciência de Processamento de Dados ou por um SIPS (Science Investigator-led Processing Systems – Sistema de Processamento liderado por Investigador Científico) para produzir resultados de níveis superiores).
Nível 1A	Dados de instrumentos em resolução total, reconstruídos e não processados, com referência ao tempo e com informações auxiliares anotadas, incluindo coeficientes de calibração geométricos e radiométricos e parâmetros de georeferenciamento (por exemplo, Plataforma Ephemeris), computados e anexados, mas não aplicados ao Nível 0 de dado[18].
Nível 1B	Dados no Nível 1A que foram processados por unidade do sensor (nem todos os instrumentos possuem dados de origem para o Nível 1B[19].
Nível 2	Dados derivados de variáveis geofísicas na mesma resolução e posição que os dados de origem para o Nível 1.
Nível 3	Variáveis mapeadas em grades de escala uniforme do espaço-tempo, geralmente com alguma integridade e consistência.
Nível 4	Modelos derivados ou resultados da análise de dados de níveis inferiores (por exemplo, variáveis derivadas de múltiplas medições).

Fonte: Borgman (2015, p. 22); Feldman (2016), NASA (2016) – Livre tradução com fundamento nas fontes citadas.

[17] O quadro original pode ser visualizado no site da Nasa - http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/.
[18] De acordo com Feldman (2016) o Nível 1A de dados de arquivo é o preferido pelos cientistas da Nasa, pois se houver mudanças de calibração do sensor, os dados não precisam ser recoletados.
[19] De acordo com Feldman (2016), os dados de nível 1B são dados Nível 1A que tiveram calibrações de instrumentos / radiométricos aplicadas.

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

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

Strasser (2015) defende que enfrentar os desafios inerentes à pesquisa do século XX1 exige uma boa gestão de dados de pesquisa (GDP). Para o autor, ao se planejar com cuidado a documentação e preservação dos dados, os objetivos de ter-se dados de pesquisa reprodutíveis e transparentes são muito mais fáceis de alcançar. Além disso, dados bem geridos são mais fáceis de se utilizar e viabilizar sua reutilização, o que se traduz em maior colaboração para pesquisadores e o máximo de retorno do investimento para as agências de fomento.

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

Curadoria de dados pode ser entendida como a gestão e a preservação de dados em longo prazo, incluindo-se nesse contexto o fato de agregar valor aos dados digitais, bem como viabilizar a criação de novos dados, de maneira colaborativa, a partir dos já existentes.

Além disso, a atividade de curadoria também pode propiciar a redução dos riscos de obsolescência digital (DIGITAL CURATION CENTER, 2016; HEY, TANSLEY, TOLLE, 2011; ABOTT, 2008; GIARETTA, 2004).

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

Conway (1997), Sayão (2012) e Sales (2014) salientam que a teoria da curadoria digital traz, no contexto da preservação digital, o diferencial de que a informação não "apenas" deve ser preservada digitalmente, mas também de passar pelo processo de curadoria digital, o que envolveria o tratamento da informação desde a coleta dos dados de pesquisa até o reuso da informação por outros integrantes do fluxo informacional. [grifo nosso]

Em se considerando o já exposto a respeito da curadoria, deve ser comentado que essa atividade não aparece de maneira explícita no Modelo do DataONE, mas pode-se inferir que está implícita nas atividades de descrição e preservação.

ASPECTOS TECNOLÓGICOS

Os sistemas de gerenciamento de bancos de dados (SGBD) adequados para o processamento de grandes quantidades de dados não são os tradicionais (MySQL, PostgreSQL, Oracle, SQLServer etc.), até mesmo em função do custo de armazenamento, como será demonstrado no quadro 4. A respeito do assunto, Davenport (2014, p. 113) argumenta que "esses dados volumosos não podem ser bem manipulados por um software de banco de dados tradicional ou com servidores individuais (...) dessa forma uma nova geração de software de processamento de dados foi desenvolvida para resolver esse problema".

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

O Hadoop é um projeto de *software* livre desenvolvido pela Apache Software Foundation e por esse motivo às vezes é chamado de Apache Hadoop. A plataforma de computação distribuída do *software* Hadoop é em Java, voltada para *clusters* e processamento de grandes massas de dados. Possui alta escalabilidade, forte confiabilidade e tolerância a falhas. Para Davenport (p. 58), "O Hadoop é um ambiente de armazenamento e processamento de big data unificado em vários servidores". De acordo com o autor, "um *cluster* Hadoop com cinquenta nós e oitocentos núcleos de processamento é capaz de processar 1 *petabyte* de dados" (DAVENPORT, 2013, p. 59). A respeito da plataforma, Chechia (2013) comenta que os maiores colaboradores para o seu aprimoramento são o Facebook, a Google, o Yahoo e a IBM.

Quadro 4 - Custos de armazenamento de dados

Volume de dados	**Custo de armazenamento por 1 ano**		
	Banco de dados relacional tradicional	**Appliance de dados**	**Cluster Hadoop**
1 Terabyte	US$ 37 milhões	US$ 5 milhões	US$ 2 milhões

Fonte: Davenport (2014, p. 58).

Davenport (2014, p. 111) argumenta que o "big data é mais que apenas grande volume de dados não estruturados. Ele também inclui as tecnologias que possibilitam seu processamento e análise". No intuito de expor as tecnologias utilizadas no *big data*, o autor elaborou uma síntese, conforme demonstra o quadro 5.

Quadro 5 – Visão geral das tecnologias de *big data*

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

Fonte: Davenport (2014, p. 112).

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

Por todo o exposto, deve-se refletir sobre o sucateamento pelo qual muitas unidades de pesquisa têm passado no Brasil. Se o pesquisador tem dificuldades para obter financiamento para a execução da pesquisa, a realidade da organização à qual o pesquisador está vinculado não é diferente. O orçamento necessário para que as unidades de tecnologia da informação se preparem para tratar o volume de dados imposto pela *e-Science* requer alto investimento. Além disso, como se discutiu anteriormente, novas tecnologias estão sendo adotadas para facilitar o processamento desse grande volume de dados.

Logo, há que se investir tanto na compra de equipamentos e, por vezes, em licenças de *software*, como na capacitação do profissional de tecnologia da informação. Por fim, não se pode negligenciar o fato de os dados serem um ativo institucional que, portanto, precisam passar pelo ciclo de preservação de dados de longo prazo, seja o proposto pelo Modelo OAIS, ou pelo Modelo DataONE, ou mesmo um modelo customizado adequado à realidade da instituição, o que novamente requer investimento em capacitação profissional.

Sem dúvida nenhuma, o importante é dispor de um ambiente em que o dado seja preservado de forma a ser reutilizado em pesquisas futuras. Mas como oferecer esse ambiente sem a infraestrutura tecnológica adequada e sem a capacitação profissional necessária para viabilizar a gestão dos dados de pesquisa? Os próprios pesquisadores precisam entender a importância da preservação de longo prazo do dado produzido pela sua pesquisa, para a partir de aí sensibilizarem o alto nível estratégico das instituições de pesquisa e assim obterem apoio financeiro e institucional para os projetos em questão.

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

No presente, muitas das ações ligadas à biblioteca e/ou ao repositório digital envolvem a digitalização do material existente, como, por exemplo, livros e fotografias. Infelizmente, poucos projetos dessas bibliotecas digitais consideram a preservação além da digitalização inicial. A ação de copiar a informação sem alterá-la oferece uma solução de curto prazo para a preservação do acesso aos objetos digitais. Isto faz com que a informação seja armazenada em uma nova mídia antes que a mídia antiga se deteriore.

Porém, em longo prazo, essa simples migração nem sempre funciona. Aqui entra, portanto, a necessidade de implantar-se uma política de preservação digital de longo prazo que leve em consideração todos os outros aspectos relacionados com a informação digital, bem como aspectos relacionados aos dados de pesquisa produzidos em larga escala.

Merece ser ressaltado que as bibliotecas brasileiras ainda têm uma atuação tímida e isolada na preservação de documentos. Pode-se se dizer o mesmo sobre a realidade brasileira no que diz respeito à preservação de dados de pesquisa. Sob esse aspecto, os profissionais da informação precisam estar atentos às mudanças de necessidades de informação do usuário a fim de preencher esse espaço profissional, caso contrário, corre-se o risco de as atividades de curadoria, preservação e outros aspectos, inerentes ao tratamento desse grande volume de dados, serem realizadas pelos especialistas de tecnologia da informação, ou ainda por uma nova categoria de profissional que nasceu para atender às demandas do big data – o cientista de dados.

OS REPOSITÓRIOS DE DADOS DE PESQUISA NO BRASIL

No Brasil, a percepção de Sales (2014; p.49) é de que: "os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a entender que estes dados, se preservados e bem gerenciados, constituem uma excelente fonte de recursos informacionais que podem ser compartilhados e reutilizados como insumo para novas pesquisas".

Corroborando a percepção de Sales (2014), já é possível constatar que o Brasil possui os seguintes repositórios: a) Repositório de Dados do Programa de Pesquisa de Biodiversidade da Amazônia Ocidental (PPBIO), b) Repositório de Dados do Programa de Pesquisas Ecológicas de Longa Duração (PELD), Portal GEOINFO de infraestrutura de dados espaciais da Embrapa (*com 1.081 itens catalogados*) dentre outros que serão comentados.

Além desses, merece ser comentado o Portal da Biodiversidade[20] (SISBio) lançado pelo Ministério do Meio Ambiente e pelo Instituto Chico Mendes de Conservação da Biodversidade (ICMBio) em 26 de novembro de 2015. Seu desenvolvimento teve o auxílio de pesquisadores da Escola Politécnica da USP, que conseguiram reunir em uma única interface de busca as informações de bancos de dados mantidos pelo ICMBio e pelo Jardim Botânico do Rio de Janeiro. O Portal oferece buscas textuais e geoespaciais, visualização e *download* de registro de ocorrências de espécies. Além disso, "já conta com mais de um milhão de registros (coordenadas geográficas) de espécies, resultantes da integração de nove bases de dados mantidas pelo ICMBio" (BRASIL. ICMBio, 2015).

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

Outros órgãos **iniciaram** o desenvolvimento de seu repositório de dados, dentre eles merece ser citado como exemplo o Instituto de Energia Nuclear que já criou a plataforma CarpeDIEN[21] (Dados e Informações em Engenharia Nuclear) e aos poucos vem alimentando dados de pesquisas de energia nuclear do instituto. Em sua página inicial, já se observa que o sistema oferece a busca pelo autor do dado, por assunto e data de publicação.

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

[20] O Portal da Biodiversidade representa a interface web do Sistema de Autorização e Informação em Biodiversidade (SISBIO). Disponível para consulta em <https://portaldabiodiversidade.icmbio.gov.br/portal/>

[21] Disponível em < http://carpedien.ien.gov.br/>.

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

REFLEXÕES SOBRE A GESTÃO DE DADOS DE PESQUISA NO BRASIL

O cenário exposto evidencia a emergência do tema dados de pesquisa e sua complexidade. O Brasil carece de uma política explícita que norteie as ações do Estado em termos de gestão e preservação dos dados de pesquisa (GDP), bem como diretrizes para reutilização dos dados em questão. Também precisa posicionar-se quanto à necessidade de acesso aberto aos dados de pesquisas financiadas por agências de fomento brasileiras.

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

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

Existe uma diversidade de pensamentos de como as políticas de GDP são implantadas e monitoradas. O certo é que esta e uma área nova para todos. Também coexiste nesse contexto a criação de ações de GDP mesmo sem a existência de políticas públicas. Certamente, tal cenário tende a mudar rapidamente nos próximos anos, com a participação mais efetiva dos agentes públicos.

É fato que precisamos de políticas públicas, mas também já é fato que algumas instituições[22] de pesquisa, ainda que de forma embrionária, usam sua autonomia para desenvolver políticas locais que atendam aos editais de fomento internacionais, bem como às necessidades de diretrizes quanto ao armazenamento, à preservação e à reutilização de dados. Também já se constata que pesquisadores começam a procurar apoio para realizar a curadoria de seus dados, pois algumas revistas internacionais exigem acesso aos mesmos para publicar determinado artigo.

Em termos de movimentação nas instituições de pesquisa em prol de impulsionar um movimento de apoio à ciência aberta, pode-se dizer que o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) vem liderando no país o movimento de acesso aberto à informação científica desde o início dos anos 2000. Hoje a instituição é considerada referência em projetos voltados ao movimento do acesso livre à informação científica e tecnológica. Exemplo desse compromisso é a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), lançada em 2002, que utiliza a tecnologia de arquivos abertos e integra sistemas de informação de teses e dissertações de instituições de ensino e pesquisa brasileiros. Além da BDTD, o instituto apoia, desde 2009, em conjunto com a Finep, a criação de repositórios institucionais abertos em universidades públicas e centros de pesquisas financiados com recursos públicos. Outra iniciativa de destaque é a formação da Rede de Serviços de Preservação Digital – Cariniana, responsável pela preservação dos periódicos eletrônicos na plataforma OJS/SEER no Brasil, e que possui planos de ampliar o projeto, abrangendo documentos de outros tipos e em variadas mídias.

[22] A exemplo, cita-se o ICMBio, Museu Emilio Goeldi e Instituto de Energia Nuclear.

A respeito das iniciativas do Ibict de acesso aberto à informação científica, merece ser ressaltado que o instituto, apenas em novembro de 2015, se articulou e organizou o *workshop* **Desafios no Contexto Contemporâneo para promover a nova ciência baseada em dados de pesquisa**. Participaram com destaque no evento o Projeto DataONE, a Rede Nacional de Pesquisa (RNP), a Empresa Brasileira de Pesquisa Agropecuária, unidade de Satélites (Embrapa), o Instituto Nacional de Pesquisa da Amazônia (Inpa), o ICMBio, a Escola Politécnica da Universidade de São Paulo (USP) dentre outros. Durante o evento houve uma cobrança dos participantes para que o Ibict liderasse, perante o MCTI, a elaboração de um conjunto de diretrizes sobre a gestão de dados de pesquisa no Brasil. Desde então, o instituto tem, de maneira ainda tímida, tentado se articular perante os demais *stakeholders* para mapear as necessidades dos pesquisadores quando à gestão de dados, bem como os principais pontos de uma política que norteie a gestão dos dados de pesquisa.

Certamente a elaboração de uma política para a gestão de dados de pesquisa em nível nacional traz à tona a necessidade de se identificar quais tipos de dados de pesquisa o Brasil produz. Além disso, é preciso escolher qual tipo de dado e qual área de conhecimento se priorizará para armazenar e iniciar um programa de gestão de dados de pesquisa. Igualmente, mostra-se necessário refletir sobre a necessidade de produzir um repositório único de acesso aos dados científicos, ou produzir diversos repositórios temáticos, ou ainda, desenvolver, a exemplo da BDTD, uma interface única que busque em diferentes repositórios o dado pesquisado. Nesse caso, quem seria a instituição que apresentaria essa competência?

Uma política de gestão de dados ainda precisa abordar outros aspectos, tais como: a) definir as regras de compartilhamento e reuso dos dados; b) definir o prazo de carência para algumas categorias de dados; c) definir prazo de armazenamento para algumas classes de dados; d) definir padrões de metadados e interoperabilidade destes; e) exigir do pesquisador um plano de gestão de dados quando a pesquisa for fomentada pelo governo; f) definir os requisitos necessários para a implementação do DOI para dados.

A maior parte das indagações anteriores precisa ser refletida em diversos níveis da esfera governamental do um país. Parece coerente afirmar que um conjunto de respostas a essas indagações precisa ser elaborado por meio

de uma política pública, elaborada por um comitê interministerial, com a participação das agências de fomento e, na medida do possível, com a participação de membros da comunidade científica.

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

CONSIDERAÇÕES FINAIS

A literatura revela que o bom uso das informações é uma necessidade. Já a história revela que, pelo menos uma vez, o mau uso da informação teve um fim trágico. Tal fato se deu quando Albert Einstein, em 1939, preocupado com as pesquisas sobre fissão nuclear, informou ao então presidente dos Estados Unidos, Franklin Roosevelt, que já existia a possibilidade de criação de uma bomba de alto poder de destruição. Roosevelt, de posse de uma informação altamente estratégica, reuniu um grupo de cientistas, do qual Einstein não fez parte, e deu início ao Projeto Manhattan, que produziu com sucesso um artefato atômico – *a bomba atômica*, posteriormente lançada em Hiroshima e Nagasaki nos dias 6 e 9 de agosto de 1945.

A respeito do assunto, Einstein, que sempre foi um pacifista, um tempo depois afirmou que o maior erro de sua vida foi ter enviado a carta a Roosevelt. Também declarou que não foi sua descoberta a causadora da tragédia da Segunda Guerra Mundial, mas o uso que fizeram dela. Aqui fica a sugestão para que cientistas reflitam sobre a forma ética no uso de seus dados.

É nesse contexto que os profissionais da informação do presente precisam trabalhar para que as informações disponibilizadas em um volume cada vez maior na internet sejam utilizadas para o bem comum, como, por exemplo, para promover o desenvolvimento de pesquisas em saúde, visando promover a cura do câncer, da AIDS, dentre outras mazelas da humanidade.

Os dados são um componente importante da pesquisa econômica e social – eles são a base para a pesquisa e o produto final da pesquisa. A qualidade dos dados de pesquisa e a sua proveniência tornam-se fundamentais no compartilhamento e na subsequente utilização como documento secundário. A gestão eficaz dos dados é uma condição essencial para a geração de dados reutilizáveis de alta qualidade. Os investigadores precisam ter o conhecimento e as habilidades para garantir que os dados que criam e gerenciam podem ser explorados ao máximo, sendo, portanto, um promissor potencial para futuras pesquisas.

Outra questão que não pode deixar de ser abordada é: de que maneira os profissionais da informação podem tratar os dados de pesquisa? Há lugar para eles nas bibliotecas? Os centros de informação como estão hoje talvez não tenham condição física, de infraestrutura e de pessoal capacitado para trabalhar com o conteúdo dos dados de pesquisa. Luce (2010, p. 3) argumenta que, para as bibliotecas, a evolução gradual dos dados de pesquisa (que ele chama de *e-Science*) provoca desafios profundos, e ao mesmo tempo proporciona a elas uma oportunidade de redefinir seus papéis e agregar valor ao seu portfólio de serviços. Hoje os laboratórios são os locais mais "populares" para se desenvolver os dados de pesquisa.

Como um objeto acadêmico, os dados continuam a crescer em importância na comunidade de pesquisa e, paulatinamente, os bibliotecários e demais profissionais da informação têm, crescentemente, maiores responsabilidades na gestão e curadoria de dados. As novas iniciativas de bibliotecas e arquivos incluem as tarefas de ajudar os pesquisadores a encontrar conjuntos de dados para reutilização; localização e hospedagem em repositórios para o arquivamento necessário, consultas sobre o fluxo de trabalho, planos de gestão de dados e melhores práticas para a preservação.

Os profissionais de informação, ao olhar para as opções de fornecer serviços e produtos a essa nova e vibrante área, precisarão saber como tratar e difundir esse tipo de documento até então negligenciado pela área de ciência da informação

A preservação de dados em formato digital é imperativa. As transformações pelas quais a ciência contemporânea vem passando já não permitem que os dados coletados sejam preservados para serem recuperados apenas nesta década, mas sim também em um futuro distante, com a mesma qualidade e confiança em sua autenticidade. A informação digital é um recurso vital na economia do conhecimento, valiosa para a pesquisa, para a educação, para o desenvolvimento tecnológico, assim como para as atividades culturais e o aprimoramento de políticas públicas.

oOo

REFERÊNCIAS

ABBOT, D. *What is digital curation*. Edinburgh, UK: Digital Curation Center, 2008. Disponível em: <http://www.era.lib.ed.ac.uk/handle/1842/3362>. Acesso em: 30 de jan. 2013.

ALFONSO-GOLDFARB, A.M. *O que é história da ciência.* São Paulo: Brasiliense, 1994. (Coleção primeiros passos, 286).

ALVARO, E. et al. E-Science librarianship: field undefined. *Issues in Science & Technology Librarianship*, n. 66, p. 28-43, Summer 2011.

ATKINS, D.E. et al. *Revolutionizing science and engineering through cyberinfrastructure*: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington, DC, 2003. Disponível em: <http://www.nsf.gov/cise/sci/reports/atkins.pdf>. Acesso em: 10 abr. 2015.

BELL, G. Prefácio. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). *O quarto paradigma*: descobertas científicas na era da e-Science. São Paulo: Oficina de Textos, 2011. P. 11- 15.

BLUE RIBBON TASK FORCE ON SUSTAINABLE DIGITAL PRESERVATION AND ACCESS - BRTF. *Sustainable economics for a digital planet:* ensuring long-term access for digital preservation: final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. Disponível em: <http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf>. Acesso em: 21 de abr. 2015.

BOLLIER, D. *The promise and peril of big data*. Washington: Aspen Institute, 2010. Disponível em: <http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf>. Acesso em: 15 maio 2014.

BORGMAN, C.L. *Big data, little data, no data*: scholarship in the networked world. Cambridge, Massachusetts; London, England: MIT Press Books, 2015. 383 p.

CASTELLS, M. *A galáxia da internet*: reflexões sobre a internet, os negócios e a sociedade. Rio de Janeiro: Jorge Zahar, 2003.

CÉSAR JÚNIOR, R.M. Apresentação à edição brasileira. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). *O quarto paradigma*: descobertas científicas na era da e-Science. São Paulo: Oficina de Textos, 2011. p. 7- 8.

CHECHIA, D. *Big data:* do conceito a prática. (slides). Palestra apresentada na 7a. Edição da Conferência O Outro Lado - Security BSides São Paulo, 2013. Disponível em: <http://pt.slideshare.net/daniel.checchia/bigdata-da-teoria-pratica>. Acesso em: 13 de julho de 2016.

CONWAY, P. *Preservação no universo digital*. Rio de Janeiro: Arquivo Nacional, 1997. (Projeto Conservação Preventiva em Bibliotecas e Arquivos). Disponível em: <http://www.portalan.arquivonacional.gov.br/media/CPBA_52_ Preserva%C3%A7%C3%A3o_Universo_Digital.pdf>. Acesso: 27 abr. 2005.

COSTA, M.M. *Política de gestão de dados científicos*: panorama mundial e diretrizes para o Brasil. [Projeto de Tese]- Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília, Universidade de Brasília, Brasília, 2016.

DATAONE: best practices. 2016. Disponível em: <https://www.dataone.org/best-practices>. Acesso em: 10 abr. 2016.

DAVENPORT, T.H. *Ecologia da informação:* por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 2001.

_____. *Big data no trabalho:* derrubando mitos e descobrindo oportunidades. Rio de Janeiro: Elsevier, 2014.

DIGITAL CURATION CENTER. *What is digital curation?.* 2016. Disponível em: <http://www.dcc.ac.uk/digital-curation/what-digital-curation>. Acesso em: 01 jul. 2016.

ESTADOS UNIDOS. White House. Executive Office of the President. *Big data:* seizing opportunities, preserving values. Washington, 2014. Disponível em: <http://www.whitehouse.gov/sites/ default/files/docs/big_data_ privacy_report_may_1_2014.pdf >. Acesso em: 20 maio 2014.

FELDMAN, G.C. *Ocean color web:* products definitions. (2016). Disponível em: <https://oceancolor.gsfc.nasa.gov/cms/products>. Acesso em: 22 de setembro de 2017.

FOX, P.; HENDLER, J. E-science semântica: o significado codificado na próxima geração de ciência digitalmente aprimorada. In: HEY, T; TANSLEY, S.; TOLLE, K. (Org.). *O quarto paradigma*: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011.

_____; HARRIS, R. ICSU and the challenges of data information management for internacional science. *Data Science Journal*, v. 12, n. 10, Feb. 2013. Disponível em: <https:jstage.jst.go.jp/article/dsj/12/0/12_WDS-001/_article>. Acesso em: 01 jul. 2016.

GIARETTA, D. *DCC approach to digital curation*. [Draft]. [S. l.]: DCC, 2004. Disponível em: <http://www.dcc.ac.uk/sites/default/files/documents/DCCApproachtoDigitalCuration-20040827.pdf>. Acesso em: 02 abr. 2014.

GRAY, J. *EScience:* a transformed scientific method. Palestra apresentada no Conselho Nacional de Pesquisa dos Estados Unidos (NRC-CSTB), Mountain View, Califórnia, 11 janeiro de 2007. Disponível em: <http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt>. Acesso em: 30 ago. 2012.

GREEN, D. Infraestrutura-científica: introdução. In: HEY, T.; TANSLEY, S.; TOLLE, K. (Org.). *O quarto paradigma*: descobertas científicas na era da e-science. São Paulo: Oficina de Textos, 2011. P. 129-130.

HEY, T.; HEY, J. E-Science and its implications for the library community. *Library Hi Tech*, v. 24, n. 4, 2006. P. 515-528.

_____; TANSLEY, S.; TOLLE, K. (Org.). *O quarto paradigma*: descobertas científicas na era da e-Science. São Paulo: Oficina de Textos, 2011. 261 p.

_____; TREFETHEN, A. E-Science and its implications. *Philosophical Transactions of the Royal Society (A),* v. 361, p. 1809-1825, June 2003.

JANKOWSKI, N.W. Exploring e-science: an introduction. *Journal of Computer-Mediated Communication*, v. 12, p. 549-562, 2007.

INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE. *Portal da biodiversidade.* Brasília, 2015. Folder.

_____. *Instrução normativa nº 03, de 01 de setembro de 2014*. Fixa as normas para a utilização do Sistema de Autorização e Informação em Biodiversidade – SISBIO, na forma das diretrizes e condições previstas nesta instrução normativa e, regulamenta a disponibilização, o acesso e o uso de dados e informações recebidos pelo Instituto Chico Mendes de Conservação da Biodiversidade por meio do SISBIO. (Processo nº 02070.001067/2013-96). Brasília, 2014.

_____. *Instrução Normativa nº 2 de 25 de novembro de 2015.* Institui a política de dados e informações sobre biodiversidade do Instituto Chico Mendes de Conservação da Biodiversidade e dispõe sobre sua disponibilização, acesso e uso. Brasília, 2012.

MACHLUP, F.; MANSFIELD, U. (Org.). *The study of information*: interdisciplinary messages. New York: John Wiley & Sons, 1983.

MARCUM, D.B.; GEORGE, G. (Ed.). *The data deluge*: can libraries cope with e-science?. Santa Barbara, California: Libraries Unlimited, 2010.

MÁRDERO ARELLANO, M.A. *Critérios para a preservação digital da informação científica* 2008. 356 f. Tese (Doutorado em Ciência da Informação)- Universidade de Brasília, Brasília, 2008. Disponível em: <http://repositorio.unb.br/handle/10482/1518>. Acesso em: 01 jun. 2016.

MAYER-SCHÖNBERGER; V.; CUKIER, K. *Big data:* como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro, Elsevier, 2013. 256 p.

MEADOWS, J. *Understanding information.* München: KG Saur, 2001. 112 p.

NASA. *NASA Science Earth. Data Processing Level.* (2016). Disponível em: <https://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/>. Acesso em: 13 jun. 2016.

PORTICO. *About us.* 2015. Disponível em: <http://www.portico.org/digital-preservation/about-us>. Acesso em: 11 jun. 2015.

RODRIGUES, E. et al. *Os repositórios de dados científicos:* estado da arte. Portugal: Universidade do Minho; Universidade do Porto, 2010. (Projecto RCAAP D24 – Relatório).

ROSENTHAL, D.S. et al. Requirements for digital preservation systems: a bottom-up approach. *D-Lib Magazine,* v. 11, n. 11, Nov. 2005. Disponível em: <http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>. Acesso em: 28 jul. 2012.

SALES, L.F. *Integração semântica de publicações científicas e dados de pesquisa:* proposta de modelo de publicação ampliada para a área de ciências nucleares. 265 f. Tese (Doutorado em Ciência da Informação)- Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

SAYÃO, L.F.; SALES, L.F. Curadoria geral: um novo patamar para a preservação de dados digitais de pesquisa. *Informação & Sociedade,* v. 22, n. 3, p. 179-191, set./dez. 2012.

_____. Dados abertos de pesquisa: ampliando o conceito de acesso livre. *Revista Eletrônica de Comunicação Informação e Inovação em Saúde,* v. 8, n. 2, jun. 2014.

_____. *Gestão de dados de pesquisa para bibliotecários e pesquisadores.* Rio de Janeiro: CNEN: IEN, 2015. 90 p. Versão Preliminar.

SHEARER, K. *Comprehensive brief on research data management policies.* 2015. Disponível em: <www.science.gc.ca/1E116DB8-E7F3-4B6F-BB44-83342BAAA030/Comprehensive% 20Brief%20on%20Research%20 Data%20Management%20Policies.pdf>. Acesso em: 05 jul. 2016.

SIMBERLOFF, D. et al. *Long-lived digital data collections*: enabling research and education in the 21st century. [S.l.]: National Science Foundation, 2005.

SOLLA-PRICE, D.J. de. *O desenvolvimento da ciência:* análise histórica, filosófica, sociológica e econômica. Rio de Janeiro: Livros Técnicos e Científicos, 1976.

STRASSER, C. *Research data management.* Baltimore: National Information Standards Organization, 2015. (NISO Primer Series).

_____ et al. *Primer on data management*: what you always wanted to know. DataONE best practices primer. [Tennesse]: DataONE, 2012. Disponível em: <https://www.dataone.org/sites/all/documents/ DataONE_BP_Primer_020212.pdf>. Acesso em: 10 jun. 2016.

TAPSCOTT, D.; WILLIAMS, A.D. *Wikinomics:* como a colaboração em massa pode mudar o seu negócio. Rio de Janeiro: Nova Fronteira, 2007.

UNESCO. *Carta para la preservación del patrimonio digital.* [Paris?]: UNESCO, 2003. Disponível em: <http:// www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/charter_preservation_ digital_heritage_ es.pdf>. Acesso em: 23 jul. 2004.

Research data: what are the impacts of large volume produced, how to organize them and what to preserve

ABSTRACT

It discusses aspects of the collaborative science of the XXI Century, science internationalization and virtualization that led to the explosion of research data collected online, giving rise to the phenomenon of big data and cyberinfrastructure, also called e-Science. It exposes the facts which brought to light the science of intensive use of research data, including the technological development of instruments for data collection and analysis. It conceptually explains what e-Science and cyberinfrastructure are. At the same time, it presents the expressions "research data" and "scientific data" and argues that because it is a new subject, there is still no consensus in the literature about which term should be used. It outsides the concept of big data, and how it works with the use of social networking data to application development. It highlights the e-Science as a part of big data, which handles large-scale data in the scientific realm. It seeks to define the term "data" to then display the term "research data", its peculiarities, ways of collection, treatment and preservation. It discusses storage and digital preservation aspects of research data, addressing aspects of technology infrastructure. It concludes the chapter with the presentation of reflections on research data management in Brazil.

***Keywords**: Big data. Data-driven science. Data curation. Cyberinfrastructure. Research data. E-Science. Research data management. Digital preservation of research data.*

Datos de investigación: qué son, impactos del gran volumen producido, cómo organizarlos y cuáles preservar

RESUMEN

En el caso de la ciencia colaborativa del siglo XXI, la internacionalización y la virtualización de la ciencia que culminaron con la explosión de datos de investigación recogidos en línea, dando origen al fenómeno de big data y cyberinfrastructure, también denominado e-Science. Expone los hechos que trajeron a la luz la ciencia de uso intensivo de datos de investigación, entre ellos el desarrollo tecnológico de los instrumentos de recolección y análisis de datos. Explica conceptualmente qué es e-Science y cyberinfraestructure. Al mismo tiempo, presenta los términos datos de investigación y datos científicos y argumenta que, por tratarse de un tema nuevo, aún no hay consenso en la literatura sobre qué término debe ser utilizado. Externo el concepto de big data, y cómo trabaja con el uso de datos de redes sociales para el desarrollo de aplicaciones. Evidencia a e-Science como una parte de la gran fecha, que se ocupa de datos a gran escala en el ámbito científico. Se busca definir el término dado para entonces presentar la expresión datos de investigación, sus peculiaridades, formas de recolección, tratamiento y preservación. Discuta sobre los aspectos de almacenamiento y preservación digital de los datos de investigación, abordando aspectos de infraestructura tecnológica. Finaliza el capítulo con la presentación de reflexiones sobre la gestión de datos de investigación en Brasil.

***Palabras clave**: Big data. Ciencia orientada a datos. Curaduría de datos. Cyberinfrastructure. Datos de investigación. E-Science. Gestión de datos de investigación. Preservación de datos de investigación.*