tríade
comunicação, cultura e mídia

A era dos grandes e pequenos volumes de dados

The age of large and small data

La era de los datos grandes y pequeños

Adriana Alves Rodrigues - Universidade Estadual da Paraíba | Campina Grande | Paraíba | Brasil | profeadrianaalves@gmail.com |
https://orcid.org/0000-0003-2378-6934

Embora seu surgimento seja recente, iniciado nos anos 2000[1], o termo *Big Data* vem ganhando atenção em muitas esferas sociais, econômicas, políticas e acadêmicas e está diretamente vinculado à era dos grandes volumes de dados na *Web*. Mesmo sem uma definição consensual sobre o termo, a ideia predominante em sua emergência era de que o volume de informação já não cabia na memória de processamentos dos computadores, necessitando assim, de especialistas para expandir a capacidade de armazenamento. Esse momento deu a origem de tecnologias como o *MapReduce*, da *Google* e do *Hadoop*, do *Yahoo,* por exemplo, que possibilitam a gerência de uma grande quantidade de dados, mais que outros poderiam processar. No entanto, com o olhar mais crítico, Christine Borgman, no seu livro *Big Data, little data, no data: Scholarship in the networked world* (2015), já demonstra novos olhares para o dilúvio emergente dos dados numa perspectiva holística.

Dividido em três partes com dez capítulos, a obra inicia com alguns questionamentos que vão dando a tônica do livro. Na primeira parte do livro, há os questionamentos relativos à terminologia e seu uso. Dados são ubíquos, estão em todos os lugares, mas é preciso questionar o que é um dado ou quais valores estão neles. De acordo com a definição do *Oxford English Dictionary*, dado pode ser um item de informação, um conjunto de dados, quantidades, símbolos, características que são operacionalizadas por computador considerado coletivamente. A autora foca na valorização dos

RESENHA

[1] Para Schonberger e Cukier (2013), a grande mudança de paradigma para que o *Big Data* explodisse em diversas esferas ocorreu em meados de 2000, mais especificamente nos campos da Astronomia e Genômica.

dados e afirma que as características deles, combinadas às tendências sociais e técnicas maiores, estão contribuindo para a reconfiguração crescente de que os dados estão se tornando mais úteis, mais valiosos e mais problemáticos para a área das pesquisas.

Para a autora, *Big data* está adquirindo grande atenção, assim como aconteceu com o *Big Science* há 15 anos. A autora explica que o dado apresenta muitos tipos de valores e, que estes, podem não ser apresentados depois que esses dados são coletados, curados ou perdidos. O valor dos dados varia de acordo com o lugar, tempo e contexto. Ela ressalta que "Os dados não têm nenhum valor ou significado em isolamento. Eles podem ser ativos ou passivos ou ambos. Eles existem dentro de uma infraestrutura conhecimento - uma ecologia de pessoas, práticas, tecnologias, instituições, material objetos anda relações"[2] (BORGMAN, 2015, Kindle Version, Location 345).

Em relação à grande disponibilidade desses dados na sociedade, para ela, a falta de dados (*no data*) torna-se mais nítida com os dados em abundância em que qualquer cidadão pode analisar, compartilhar, etc. É aqui que reside a grande inspiração do livro. Para esse contexto dos "dados pobres" (*data-poor*), Borgman diz que são uma espécie de "possessões valorizadas" (*prized-possessions)* que podem ser direcionadas para a seleção de métodos e teorias. Relevantes dados podem existir, mas são mantidos por entidades que estão sob nenhuma obrigação de liberá-los ou que podem ser proibidos por lei de liberá-los. Dados podem ser anônimos ao grau razoável, tais como pesquisas sociais em geral, são os mais propensos a se tornar disponíveis para reutilização.

Na segunda parte, o foco recai sobre os estudos dos dados no universo acadêmico. Uma vez lançados para o público, perde-se o controle de quem, como, quando e por que esses dados podem ser usados. Mas a

[2] No original: "Data have no value or meaning in isolation. They can be assets or liabilities or both. They exist within a knowledge infraestrutucture - an ecology of people, practices, technologies, institutions, material objects and a relationships".

autora faz um alerta: "nas ciências físicas e da vida, os pesquisadores podem proteger acesso a sites de pesquisa, espécies, observações e experiências[3]" (Kindle Version, Location 514). Borgman lança seis provocações sobre a natureza, função e uso dos dados no âmbito acadêmico.

1) Reprodutibilidade, compartilhamento e reuso dos dados: direcionados para quem possui, controla, acessa e sustenta os dados de investigação, vai determinar como seu valor pode ser explorado e por quem;
2) Transferência de conhecimento através de contextos ao longo do tempo é difícil: algumas formas de representação dos dados podem ser compartilhadas prontamente por meio de disciplinas, contextos;
3) As funções de publicações acadêmicas podem permanecer estáveis apesar da proliferação de formas e gêneros: Dados servem de diferentes propósitos na comunicação acadêmica. A função dos dados pode ser examinada sob perspectivas variadas no campo de investigação;
4) Os trabalhos acadêmicos estão sendo mais amplamente divulgados através de movimentos como a publicação de acesso aberto, dados abertos e software de fonte aberta: as diferentes finalidades de dados e publicações em bolsa influenciam os incentivos, meios e práticas de divulgação;
5) Infraestrutura de conhecimento envolvendo para acomodar acesso aberto, a pesquisa de dados intensivos, as novas tecnologias, mídias sociais, e as mudanças na prática e política: conhecimentos são necessários, mas as suas aplicações serão variáveis em diferentes contextos e domínios de investigação;

[3] No original: "In the physical and life sciences, researchers can protect access to research sites, species, observations and experiments".

6) Infraestrutura de conhecimento desenvolve adaptar-se ao longo de gerações de acadêmicos (Kindle Version, Location 627).

A questão dos dados está se tornando ubíqua, pervasiva na sociedade contemporânea, mas seu entendimento depende do contexto no qual os dados estão inseridos e do olhar do pesquisador. "Os dados não são objetos puros ou naturais com uma essência própria. Eles existem em um contexto, tendo um significado no contexto e forma perspectiva do observador" (p. 607)[4]. A definição de dados está numa fronteira não muito clara do que é e o que não é dado. Borgman enxerga a problemática de definição sobre o que se entende por dados e os define como formas de informação, um amplo conceito que está cada vez mais difícil de definir. Problemas epistemológicos e ontológicos são abundantes, resultando em muitos livros dedicados à explicação de informação e conhecimento.

O termo *data* é explorado como uma possibilidade de definição ao longo do livro. Para Borgman, quando ela refere data como entidades, usa-a na forma plural, seguindo o padrão literatura e pesquisas em comunicação. *Data* é usado no singular quando refere ao conceito. Os níveis de processamento desses dados têm implicações significantes de como eles são curados e mantidos para um uso futuro. A autora afirma que as origens dos dados podem influenciar nas decisões operacionais em que o dado está inserido. De acordo com a *US National Science Board* (NSB[5]), [6]são estabelecidas três categorias de dados, a saber: 1) dados observacionais (*observational data*): são aqueles que se caracterizam por reconhecer, notificar e registrar fatos ou ocorrências do fenômeno, geralmente com instrumentos para observação (*notebook*, satélite). São considerados os mais importantes para preservar, porque são largamente replicáveis; 2) dados computacionais *(computacional data*): são produtos de modelos

[4] No original: "The data are not pure natural objects or with their own essence. They exist in a context, having a meaning that of context and forming the perspective of the observer".

[6] Disponível em: https://www.nsf.gov/nsb/. Acesso em: 22 jun. 2020.

executados por computador, simulações ou fluxo de trabalho. Encontrados nas ciências sociais e nas humanidades. Podem ser reusados em um futuro num modo extensivo de documentação de *hardware, software* e *input data.* 3) dados experimentais (experimental data): são resultados de procedimentos em condições controladas para testar ou estabelecer hipóteses ou para descobrir ou para testar novas leis e experimentos.

A longa vida dos dados é enfatizada na terceira parte do livro e traz implicações políticas destas três categorias de dados, cada um com distintas características de curadoria. Vários tipos de registros são associados com os dados observacionais, experimentais, e computacionais em distintos registros históricos, campos registrados e notas escritas à mão. Registro, para ela, é um termo essencial para o entendimento dos dados e que é raramente definido academicamente. Borgmam posiciona registro como uma quarta categoria da origem dos dados, porque esta engloba formas de dados que não se encaixam facilmente em categorias de observação, experimentação, computação ou que resultem de qualquer uma destas categorias. Registro de qualquer fenômeno ou atividade humana pode ser tratado como dados para pesquisa em sua visão.

Os esforços para categorias dos conjuntos de dados digitais (observacional, experimental, computacional) significam afirmar que o mesmo dado pode ser incorporado dentro de múltiplas coleções, mas diferentemente representados em cada situação. O exemplo apontado é a pesquisa em coleção de dados (*Research data collections*) da *National Science Board* (NSB)[7] que, segundo a autora, é fruto do resultado de muitos projetos feitos em vários contextos e finalidades diferenciadas. Esses dados têm tido processamento e curadoria diferenciados e podem não estar em conformidade com os padrões dos estudos acadêmicos em relação aos formatos e estruturas, já que "essas coleções de dados podem estabelecer normas para esta comunidade, por adoção ou através do desenvolvimento

RESENHA

[7] Fundação Nacional de Pesquisas Científicas com sede nos EUA.

de novos padrões[8]" (BORGMAN, 2015, Kindle Version, Location 763, tradução nossa).

A ampliação do uso do conceito de dados na mídia é outra esfera abordada na obra, e nas pesquisas acadêmicas reflete na ubiquidade das pesquisas em dados que estão disponíveis no formato digital. Neste sentido, Borgman afirma que em todos os setores, os dados digitais tornam-se mais fáceis para gerar, minerar e distribuir em vários formatos e em vários espaços de ação. Uma das áreas é o *Open Acess* para publicações está acelerando o fluxo de conteúdo acadêmico em muitas áreas, por um lado, e por outro, acarreta tensões e conflitos dialógicos e conceituais, tendo em vista que o fluxo de informação depende cada vez mais fortemente da infraestrutura tecnológica.

Ainda no universo da pesquisa, foco principal da autora, Borgman explora a definição de infraestrutura do conhecimento e define como "Rede robusta de pessoas, artefatos e instituições que geram, compartilhar e mantem um conhecimento específico sobre as palavras humanos e naturais"[9] (Kindle version, Location 456, tradução nossa). Estas redes incluem tecnologia, atividades intelectuais, aprendizagem, colaboração e acesso distribuído para a experiência humana e para a informação documentada. Borgman diz que dado é uma forma de informação que parece estar sempre em movimento, difícil de corrigir de forma estática, mas produz novas formas de conhecimento como o *data mining* e *crowndsourcing,* que auxiliam no remapeamento e na reformatação de território intelectual.

Para ela "dados muitas vezes são 'objetos de fronteira', que existem tenuamente nas fronteiras entre as áreas de domínio, tais como coletar, criar, analisar, interpretar e gerenciamento de dados exige conhecimentos

RESENHA

[8] No original: "These data collections may establish standards for these community, wether by adoption or by developing new standards".
[9] No original: "robust network of people, artifacts and instituitions that generate, share and mantain specific knowledge about the human and natural words".

no domínio da investigação" (Kindle Version, Location 965, tradução nossa)[10]. Métodos comuns de representação de dados como metadados, linguagens de marcação, formatos, rotulação, tesauros, ontologias facilitam o intercâmbio dos dados dentro do campo de conhecimento.

Assim, a infraestrutura de conhecimento foi formatada e permanece num processo contínuo de reconfiguração, onde se desenvolve, entre outros aspectos, o *open acess, open data*, ambos com o propósito de melhorar o fluxo informacional, minimizar possíveis restrições de uso, e aumentar a transparência na prática de pesquisa. A autora destaca que o *open acess* é um conceito simples, mas que é mal-entendido em muitas áreas de abrangência, e que pode facilitar a criação de dados e de como os dados são tratados, tendo as bases de dados são exemplos desse movimento. Assim, o *Open data* depende de tecnologias abertas e que são desenvolvidas para compartilhar, expandir as possibilidades de disseminação da informação que têm impacto na gerência dos dados abertos. O grau de abertura dos dados, padrões e tecnologias influencia a capacidade de trocar dados entre ferramentas, laboratórios, parceiros, e ao longo do tempo.

As normas podem melhorar o fluxo de informações dentro das comunidades, mas também podem criar limites entre eles. Novas Tecnologias facilitam meios de notícias de comunicação, mas eles também desestabilizam os modelos existentes[11] (Location, 1195, tradução nossa). Por fim, é proposto o *Data Metaphor* (metáfora dos dados) baseado em três propósitos: 1) legitimação (geralmente realizada por revisão por pares; 2)

RESENHA

[10] No original: "data often are 'boundary objects' that exist tenuously on the borders between the domain areas such as colleting, creating, analyzing, interpreting and managing data requires expertise in the research domain".

[11] No original: "The degree of openness of data, standards, and technologies influences the ability to exchange data between tools, labs, partners, and over time. Standards can improve the flow of information within communities but also can create boundaries between them. New Technologies facilitate news means of communication, but they also destabilize existing models".

disseminação e 3) acesso, preservação e curadoria. Dessa maneira, os estudos dos dados ganhariam mais profundidade acadêmica e de pesquisa.

Ao longo do livro fica visível a tentativa em estabelecer parâmetros conceituais para o termo dado, cujos resultados apresentados sobre o modo como o termo vem sendo tratado em distintos âmbitos demonstram ser uma problemática recorrente que parece estar aumentando. Esta é uma das contribuições de Borgman, quando enfatiza a terminologia de modo contextual, ao invés de refletir apenas por um ponto de vista tecnicista. O descompasso entre o dilúvio de dados e a sistematização acadêmica ainda é um caminho arenoso. Trata-se, portanto, de uma questão em aberto e com muitas possibilidades de reflexão para os pesquisadores e demais especialistas que lidam com o *Big Data* na contemporaneidade. A crítica direcionada destina-se à emergência era dos grandes volumes de dados, ou seja, na contramão da firula do fenômeno, é preciso levar em consideração que nem sempre os grandes são os melhores ou os mais adequados a serem processados. Na verdade, o valor, muitas vezes reside nos pequenos dados, independentemente do contexto em que estão inseridos. O *Big Data* é um processo dinâmico e complexo e, ao mesmo tempo, requer uma reflexão mais profunda sobre questões éticas, processamento e análise dos dados. A mensagem que se apreende é que se faz necessário ir além de todo esse conjunto de dados, contextualizando e tensionando o fenômeno.

Referências

BORGMAN, C. L. **Big data, little data, no data**. Scholarship in the networked world. Cambridge: The MIT Press, 2015. (Kindle version).

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

http://dx.doi.org/10.22484/2318-5694.2020v8n17p181-188

Recebido em fevereiro 2020 – Aprovado em junho 2020.

RESENHA