https://doi.org/10.14195/2182-7974_36_1_9

Edmond, J., Horsley, N., Lehmann, J., & Priddy, M. (2022). *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. Bloomsbury Academic.

ANABELA PIRES DUARTE
Doutoranda em Ciência da Informação
Faculdade de Letras, Universidade de Coimbra
duarte.anabela@gmail.com
ORCID: https://orcid.org/0000-0002-0597-5777

'In God we trust. All others, bring data.[1]' (Walton, 1989)

Jennifer Edmond, Nicola Horsley, Jörg Lehmann e Mike Priddy, investigadores e académicos de diferentes áreas da tecnologia e cultura, em 2017-2018, integraram um projeto financiado pelo programa de pesquisa Horizon 2020 da Comissão Europeia – Knowledge complexity (KPLEX).

As linhas de investigação do projecto KPLEX, direcionadas para as *Digital Humanities* (DH), foram o terreno fértil para o conjunto de ensaios que constituiu o livro *The Trouble With Big Data*. O fenómeno de dataficação foi analisado como um conjunto potencial de riscos que incluíam, entre outros, perdas no processo de criação de conhecimento, na medida em que as "abordagens são altamente seletivas, excluindo qualquer input que não possa ser efetivamente estruturada, representada ou, de facto, digitalizada" («KPLEX – KNOWLEDGE COMPLEXITY», 2017).

O KPLEX Project surgiu com o propósito de observar as lacunas, tendências e consequências observadas num pequeno grupo de investigadores de diferentes áreas académicas, na sua relação com a partilha de dados. Estas falhas de comunicação, estes *gap's*, foram descritas por um dos autores como "a game being played of mutual misunderstanding, according to

[1] Citação atribuída a W. Edwards Deming.

which all sides may feel they have won, but only because they are using different rule books".

Selecionaram três fontes potenciais da falta de comunicação dentro das equipas, que realizavam pesquisas relacionadas com dados: i) a questão linguística, do discurso e da narrativa - a heterogeneidade na definição de dados, as implicações dos seus diferentes (des)entendimentos, as consequências da *datafication* na diversidade dos elementos culturais, reduzindo e/ou eliminando a polissemia na proveniência dos dados; ii) a ameaça da diluição das narrativas de histórias e identidades nacionais, estando acessíveis apenas as visíveis digitalmente (uma baixa percentagem), por impedimentos de ordem tecnológica, mas também por alguma desconfiança nos intermediários que disponibilizam informação (como demonstram os autores ao analisarem as práticas das *Cultural Heritage Institutions* (CHI's); iii) a tentativa de compreensão da variedade de estratégias que os investigadores profissionais aplicavam na integração dos dados nos seus processos de criação de conhecimento, numa abordagem comparativa de vários projetos de investigação (incluindo os das áreas de humanidades e de estudos culturais).

Foi neste contexto que surgiu o livro, *The Trouble With Big Data*. Trata-se de uma coletânea de ensaios que explora as implicações culturais e sociais da crescente dependência de dados, alerta para a potencial falta de preservação de práticas culturais importantes, num mundo cada vez mais orientado por algoritmos e pretende dar um contributo para que os investigadores não tenham de cruzar regularmente as fronteiras entre a investigação baseada em dados e as tradições qualitativas dos estudos culturais, para prosseguirem os seus objetos de estudo.

Os autores alargaram o escopo de estudo a outras áreas para além das Humanidades, analisando de forma mais abrangente o fenómeno, estabelecendo relações de causa-efeito do fenómeno B*ig Data* com as diversas formas de expressão cultural e de conhecimento na sua dimensão humana: "What we advocate is the recognition of humans and their biases in every step of knowledge production and that the design of systems that change cultural heritage practitioner's practice be driven by them".

Datafication é o processo de transformação de práticas culturais, sociais e outras formas de atividade humana em dados quantificados, para que possam ser analisados e processados por computadores (Mayer-Schönberger & Cukier, 2013; Van Dijck, 2014; Dourish & Gómez Cruz, 2018).

Esta obra estimula a discussão sobre as consequências do deslumbramento com o fenómeno *Big Data*, apresenta argumentos fundamentados dos perigos reais das práticas de dataficação e demonstra como constituem

uma perda efetiva da singularidade e densidade das práticas culturais e um empobrecimento, pela homogeneização forçada, para que "encaixem" num padrão adequado às estruturas de análise de dados, ou seja, reduzir a valores numéricos as manifestações da complexidade e subtileza humana.

A digitalização em larga escala da cultura material, o acesso a livros on-line, simulações de fenómenos de dimensões extremas, rápidos ou lentos ou muito complexos para serem explorados num laboratório de pesquisa, tornaram-se possíveis pela existência de uma ciência intensiva de dados, que Gray cunhou como o 4º paradigma da ciência, denominada e-Science ou Ciberinfraestrutura: uma síntese da tecnologia da informação e ciência, que possibilitava enfrentar desafios a escalas anteriormente inimagináveis, uma infraestrutura que permitia formas de conhecimento intensivas em informações e dados, partilhadas, colaborativas e multidisciplinares (Gray, 2007; Borgman, 2007). Uma realidade ambivalente no que concerne ao seu impacto.

Dentre os benefícios da e-Science foi apontado a convergência semântica de ferramentas de dados, cruzando fronteiras disciplinares e epistemológicas (Gray et al., 2005). No entanto, esta moldura, ainda enleada nas perspetivas mais promissoras de avanço, ciência colaborativa e democratização do conhecimento, já suscitava algumas vozes que faziam soar o alarme para a interação com o elemento humano, com todas as suas idiossincrasias (Saracevic, 1999; Borgman et al., 2015; Regan, 2012).

O processo de conhecimento está agora intrinsecamente ligado aos dados. E há que manter uma visão crítica face ao seu uso e impacto. Christine Borgman afirmou, que deve estar presente o elemento humano e considerou o conhecimento como "redes robustas de pessoas, artefactos e instituições" sublinhando que "apesar das divergências, as fronteiras entre as ciências e as humanidades estão a desvanecer-se no que toca às práticas na eScience" (Borgman et al., 2015; 2007).

Considerámos importante esta brevíssima alusão ao contexto do fenómeno *Big Data*, como incremento para a discussão levantada pelos autores desta obra.

A perspetiva dos autores perante *data/big data* pende para um lado mais sombrio.

> […] fantasy of data seems to have established itself as a contemporary fetish object, touted as 'the new oil, or even […] the secret to living happily ever after; you may feel that you own it, but […] you are a data subject, not a controller, or indeed an owner. (p. 15)

Estas são algumas das expressões utilizadas na caracterização, não dos dados *per si*, mas da forma como estão a ser tratados e tornados (in)acessíveis e omnipotentes, com consequências diretas na vida das pessoas. Clama a urgência de linhas de raciocínio empiricamente fundamentadas, de pensamento crítico que "complemente as estratégias para orientar as respostas aos desenvolvimentos na sociedade digital que crescem a partir das ciências sociais" e crie condições para que se estabeleça um "paradigma que permita que os métodos e o conhecimento das humanidades assumam um papel de liderança no estudo do digital nas culturas humanas e vidas".

Numa nota com algum humor, aligeiram o discurso face ao fenómeno *big data*, no que toca às expectativas defraudadas de uma acessibilidade democrática: "The above is not to say that we are advocates of calling up Bill Gates to switch the internet off".

A esta visão subjazem profundas preocupações no choque entre a lógica do *Big Data* e "as normas e valores culturais estabelecidos, valorizados e essenciais" ou na falta de acesso à tecnologia e às suas inovações, o que representa, do ponto de vista dos autores, "uma rutura que cria classes epistémicas que mudam rapidamente daqueles que 'possuem' para os que 'não possuem'".

Note-se que o foco não é apenas no *deluge* de dados ou na forma de lidar com o crescimento exponencial de dados. A questão reveste-se do conceito de dados enquanto representação *versus* a realidade humana, diversificada e pouco passível de ser reduzida a algoritmos. A preocupação que perpassa na obra, não se foca apenas no choque entre *big data* e práticas humanas. Está em causa a própria perda do fator humano no registo e tratamento da memória coletiva.

Apesar destes receios, podemos constatar que os valores e normas da sociedade continuarão a adaptar-se, aliadas a uma maior exposição ao fenómeno *Big Data*. Esse choque, esperemos, manterá uma tensão contínua de forças que, eventualmente, formará um caminho seguro, rumo ao conhecimento global e democrático. É da maior importância a existência de publicações e vozes divergentes, para uma visão equilibrada dos passos a seguir. Estimular o debate de ideias e perspetivas é crucial em terrenos ainda tão pouco definidos. Esta obra reveste-se de um carácter assertivo e provocatório, permitindo levantar questões necessárias e urgentes do impacto real do tratamento dos dados, tanto na evolução do conhecimento como na vida quotidiana de todos nós.

Estrutura-se de forma articulada, sendo o seu primeiro capítulo um guião claro do fio condutor da narrativa apresentada. Numa abordagem manifestamente holística, os autores refletem, sobre as falhas nas estratégias

existentes no trabalho com dados, ilustrando com exemplos práticos elucidativos e um conjunto de 38 entrevistas realizadas a um grupo multidisciplinar que, pela sua heterogeneidade, forneceram uma "perspetiva valiosa e única sobre alguns dos lados mais sombrios da pesquisa orientada por dados, da sua propensão a incorporar tendências das equipas e dos processos que as estruturam e usam, e do risco que representa para indivíduos e códigos culturais".

No capítulo 2 os autores observam, com uma incursão nas questões semânticas e linguísticas, como a falta de consenso em torno da definição do termo e do conceito de *data* que espelham a falha de comunicação entre as diferentes culturas epistémicas. Uma replicação do que constataram no projeto KPLEX. À falta de consenso na utilização do termo e das suas definições chamam de "dysfunctional relationship with the concept of data".

O contexto dos dados apresenta-se como fulcral e os autores revisitam algumas conceções de académicos como Sabina Leonelli que considera os dados como "relational category applied to research outputs that are taken, at specific moments of inquiry, to provide evidence for knowledge claims of interest to the researchers involved", ou Christine Borgman que considera que "the relationship, '[data]' exist in a context, taking on meaning from that context and from the perspective of the beholder".

O 3º capítulo mostra como a questão da objetividade *versus* subjetividade, nas diferentes áreas do conhecimento, está presente e como pode contribuir para um distanciamento quase perverso, a vários níveis, entre as diferentes áreas.

As práticas das diferentes culturas epistémicas têm perfis diferentes que devem ser considerados: os cientistas das ciências exatas coletam dados para responder a questões imediatas e específicas, enquanto os cientistas das humanidades perspetivam a coleta de dados a longo termo, sendo criados e preservados ao longo do tempo, respondendo a questões de natureza mais criativa e passíveis de interpretação humana.

As humanidades realizam uma avaliação crítica das fontes, com uma abertura a duplas verificações, diferentes interpretações e abordagens alternativas, com narrativas enriquecidas com descrições e contexto, muito distantes das interpretações de *big data*, permitindo orientação e criação de sentido socialmente pertinente e reflexões críticas do passado para o presente e o futuro.

A digitalização nas humanidades, em que as informações devem ser trazidas para novos formatos de acordo com os padrões de metadados, garantindo a troca de dados entre instituições, interoperabilidade, agregação

e dimensionamento demonstram as desigualdades que se estabelecem entre projetos provenientes de Instituições com maior ou menor capacidade de investimento. Quando a digitalização não é possível por impedimentos de várias ordens (financeiras, tecnológicas) vota certos projetos a uma invisibilidade e quase inexistência. Neste sentido os autores encaram a digitalização como uma força disruptiva.

Sublinham que os dados são uma construção social e um processo cultural, que moldam e são moldados pelos contextos culturais e pelas práticas de linguagem onde se operam.

Os capítulos 4 a 6 focam o silenciamento ou empobrecimento de vários cenários da cultura, os perigos de uma exposição não controlada, a ausência de conhecimento tácito essencial em certas áreas, o determinismo a que todos nós podemos estar sujeitos ao sermos encaminhados por algoritmos para o que devemos ouvir, ler, pensar. São notórios os receios de perda da História da Humanidade e diria até receios da perda de Humanidade em toda a sua complexidade, diversidade, pensamento crítico e liberdade e acessibilidade nas escolhas individuais.

No capítulo seguinte mantêm, talvez de forma mais explícita e incisiva, a crítica à dataficação como potencial silenciador de vozes de todo o espectro humano e cultural e como certas tendências reforçam as assimetrias e desigualdades na sociedade e impulsionam as relações de poder:

[…] data need to be understood as an essential part of infrastructures, which should be owned by the public just as artificial intelligence built on top of them; conceptualized in this way, data would form public commodities which can be used by several agents simultaneously, but private corporations would have to pay for their use.

Levantam uma questão muito relevante: o eixo anglo-americano prevalece em termos linguísticos. Esta homogeneização linguística repercute-se em todas as línguas minoritárias que arriscam ser votadas ao esquecimento. "The unusual will become unused and ultimately unusable" aplica-se tanto ao conhecimento alternativo como a todos os elementos que se encontram fora do paradigma da homogeneização, incluindo as variâncias linguísticas.

Os autores citam Postman (1993) que dá como exemplo uma passagem da obra de Huxley (1932) "Admirável Mundo Novo", referência ficcional muito atual, como ilustração da submissão de todas as formas de vida cultural à soberania da técnica e da tecnologia.

O capítulo final reaviva a atenção sobre o ecossistema que envolve os dados e os os seus atores na criação e partilha do conhecimento. Os autores pretendem, com esta provocação "empowering both researchers and their

subjects to make the development of big data approaches to knowledge creation more humane". Retirar o empoderamento dos dados isolados do seu contexto, relembrando que não são omnipotentes e omniscientes sem o elemento humano, sem o contexto, sem a riqueza das emoções humanas.

Como nota final, parece-me que atingem claramente o objetivo de provocação e questionamento da supremacia dos dados, no modo como vemos e interpetamos o mundo e suscitam perguntas essenciais e urgentes. No entanto, creio que enriqueceria o seu conteúdo ao fornecerem uma visão estrutural mais lata, onde fossem identificados os elementos positivos desta era tecnológica, aliando assim as suas propostas e questões à realidade atual. A consciência da hegemonia dos grandes grupos económicos, empresariais, governamentais como ameaça a um bem comum que deveria pertencer ao público, a todos nós, o conhecimento sob todas as suas formas que advém da multiplicidade de dados, deve ser acompanhada de propostas exequíveis a breve termo, face aos poderes atualmente instaurados de difícil penetração. A negociação deve estar aliada à contestação, para evitar ruturas que empoderem ainda mais as entidades com meios para aceder e tratar grandes quantidades de dados, obscurerecendo os pequenos núcleos. Caminhar dentro do paradigma para realizar transformações, optar por um discurso de pequenos passos que levem à mudança, aliando as perspetivas positivas com as negativas do *big data*. O cidadão individual, as pequenas comunidades, os pequenos projetos científicos, as culturas epistémicas de menor visibilidade desmoralizam rapidamente e incorrem até no risco de serem "esmagadas" se optarem por um confronto de contornos demasiado inflexíveis.

A obra está acessível em acesso aberto in *https://www.bloomsburycollections.com/book/the-trouble-with-big-data-how-datafication-displaces-cultural-practices/*

Bibliografia e leituras recomendadas

Borgman, C. L. (2007). *Scholarship in the digital age*. MIT Press.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

Borgman, C. L. (2020). Big Data, Little Data, or No Data? Why Human Interaction with Data is a Hard Problem. *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 1. https://doi.org/10.1145/3343413.3377979

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: data, diversity, and digital

libraries. *International Journal on Digital Libraries, 16*(3-4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

Dourish, P., & Gómez Cruz, E. (2018). Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society*, *5*(2), 1-10. https://doi.org/10.1177/2053951718784083

Edmond, J., Horsley, N., Lehmann, J., & Priddy, M. (2022). *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. Bloomsbury Academic. https://doi.org/10.5040/9781350239654

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005). Scientific data management in the coming decade. *ACM SIGMOD Record*, *34*(4), 34-41. https://doi.org/10.1145/1107499.1107503

Gray, J. (2007). Jim Gray on eScience: A Transformed Scientific Method. In S. T. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Dataintensive scientific discovery*. Microsoft Research.

Griffin, S. (2013). New Roles for Libraries in Supporting Data-Intensive Research and Advancing Scholarly Communication. *International Journal of Humanities and Arts Computing*, *7*(supplement), 59-71. https://doi.org/10.3366/ijhac.2013.0060

Huxley, A. (1932). *Brave new world*. Garden City Pub. Co.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

Postman, N. (1993). *Technopoly: The Surrender of Culture to Technology*. Penguin Random House.

Regan, C. J. (2012). Review: The Fourth Paradigm by Tony Hey, Stewart Tansley, and Kristin Tolle. *InterActions: UCLA Journal of Education and Information Studies, 8*(1). https://doi.org/10.5070/D481011836

Saracevic, T. (1999). Information science. *Journal of the American Society for Information Science, 50*(12), 1051-1063. https://doi.org/10.1002/(SICI)1097-4571(1999)50:12<1051::AID-ASI2>3.0.CO;2-Z

Tóth-Czifra, E. (2022). *The Trouble With Big Data: Insights from Jennifer Edmond, Jörg Lehmann, Mike Priddy and Nicola Horsley*. DARIAH Open – Open scholarly practices in the arts and humanities. https://dariahopen.hypotheses.org/1336

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197-208. https://doi.org/10.24908/ss.v12i2.4776

Walton, M. (1989). *The Deming management method*. Mercury Books.