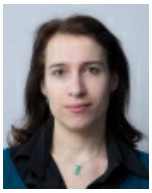
Without urgent action big and open data may widen existing inequalities and social divides



The juncture of big and open data informs areas as diverse as artificial intelligence, agriculture, and public health, and promises to transform our ability to tackle global challenges. However, ***Sabina Leonelli*** *highlights three major concerns over how big and open data are currently managed: the unsustainable nature of the digital data landscape; the quality and credibility of the data themselves; and how data sources currently represent only privileged – typically English-speaking – individuals and communities, with little representation from less visible and more vulnerable groups. These challenges can be overcome, but to do so requires significant investment in key data governance priorities.*

I would like to tackle the role of big and open data in contemporary society, and the well-justified fear that the development of related digital technologies and artificial intelligence may widen existing inequalities and social divides. The term "big data" is typically associated with the idea that lots of data, about anything and everything, can now be rapidly produced, stored, and disseminated in digital form – and that this enables new ways of searching, analysing, using, and reusing these data, often well beyond the contexts in which the data were originally generated.

Big data are most valuable when they move around, travel to new sites, and are interpreted in relation to different questions and problems. In this sense, the rise of big data (and related infrastructures and expertise) is strongly related to the rise of open science, a movement that takes advantage of digital platforms and communication technologies to revolutionise how knowledge is created, circulated, and assessed around the globe. The juncture of big and open data is informing areas as diverse as artificial intelligence, agriculture, and public health, and promises to transform human abilities to tackle global challenges. At the same time, it also has the potential to profoundly undermine the legitimacy, credibility, and trustworthiness of scientific expertise, and expand the already overwhelming divide between those who benefit from digital technologies and those who are losing out.

Over the last decade, my research focused on investigating how data are disseminated and reused across a wide variety of contexts, both within and beyond science, and in several different locations, including both low-income and high-income countries. My research group did this largely through qualitative social science and humanities methods, by interviewing data producers, stewards, and users at length about their experiences and perceptions, and studying their practice and history. Building on this research, I want to highlight three major concerns associated with the ways in which big and open data are currently managed.

Image credit: Mahdis Mousavi, via Unsplash (licensed under a CC0 1.0 license).

Date originally posted: 2018-02-14

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/02/14/without-urgent-action-big-and-open-data-may-widen-existing-inequalities-and-social-divides/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

The first is the unsustainable nature of the digital data landscape, and the urgent need to find business models that can support data storage, sharing, and analysis. Open science is neither quick nor cheap. It requires digital and material infrastructures that are only effective when they are efficiently maintained and regularly updated. And yet there is no clarity at the moment on who should shoulder the costs, for how long, and how this should be coordinated locally and internationally.

The second concern is around the quality and credibility of the data themselves and the processes used to transform those data into knowledge. In the era of fake news, it is more important than ever to provide credible evidence for knowledge claims, yet most data collections and online repositories lack effective systems of review and quality control. Similarly, there is little scrutiny of the assumptions and bias built into data mining algorithms used for data retrieval and analysis. The global reach and multiple locations of the myriad interconnected databases currently hosting big and open data makes it hard to trace accountabilities for how these systems were built and how they shape data interpretation. The push to recycle existing data may also provide a disincentive for creative, novel research, and foster a culture of conservatism in research and innovation.

The third concern is around the extent to which big and open data are reinforcing existing social divides, for instance by favouring data sources that represent only privileged individuals and communities. The vast majority of large research databases display “tractable” data produced by rich, English-speaking groups, with very little representation from less visible and more vulnerable groups.

I believe it is possible to overcome these challenges, but to do so requires urgent action. Substantial investment must go into data governance and stewardship. A key priority is supporting fairness in data handling – for instance, through the identification of exclusions and inequalities built into data pipelines, investment in adequate data expertise and skills, and the development of intelligent and ethical strategies for data sharing and the development of algorithms (including, but not limited to, following the FAIR principles).

Other priorities for data governance include:

- the development of adequate data expertise and the understanding that whether and how data should be open needs to be decided on a case-by-case basis
- mechanisms to promote the quality and trustworthiness of data sources and analytic tools
- sustainable data infrastructures (and related expertise); that is, resources that are interoperable, long-term, internationally coordinated, and publicly accountable
- creative solutions to global challenges in dialogue with relevant publics
- critical scrutiny of research across different audiences.

This blog post is based on the author’s intervention in the “Digital Transformations” plenary of the World Science Forum 2017 in Jordan (video available here).

Note: This article gives the views of the author, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the author

Sabina Leonelli *is Professor of Philosophy and History of Science and Co-Director of the Exeter Centre for the Study of the Life Sciences (Egenis) at the University of Exeter. She serves as Open Science expert for the European Commission and is working on a monograph on the impact of big data on research. Her book Data-Centric Biology appeared in 2016 with Chicago University Press. She can be found on Twitter @sabinaleonelli and the research group @DataScienceFeed.*

Date originally posted: 2018-02-14

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2018/02/14/without-urgent-action-big-and-open-data-may-widen-existing-inequalities-and-social-divides/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/