The Politics of Data: The rising prominence of a data-centric approach to scientific research.

blogs.lse.ac.uk/impactofsocialsciences/2015/10/05/politics-of-data-scientific-data-centrism/

10/5/2015

Today we launch a series of posts on the politics of data. Big data, small data and data sharing will be critically examined by a range of experts, each exploring the implications of the changing data landscape for research and society. In the first piece, Sabina Leonelli *and* Louise Bezuidenhout *argue the study of data itself is an excellent entry point to reflect on the activities and claims associated to the idea of scientific knowledge. How scientists perceive their research environments, what they recognize as strengths and limitations, and what in these environments pose material or social challenges to data engagement all influence how data travels.*

Scientific research has long involved efforts to generate, analyse and interpret large masses of data, as illustrated by the major data collection and curation efforts characterising 17th century astronomy and metereology and 18th century natural history. Thus, current big data may well have remarkable volume, velocity, variability, variety and veracity, but they are not a new phenomenon within science, and many disciplines have developed sophisticated strategies to cope with overflows of information. At the same time, contemporary manifestations of big data have distinctive features that relate to the technologies, institutions and governance structures of the contemporary scientific world.

For instance, this approach is typically associated to the emergence of large-scale, multi-national networks of scientists; to a strong emphasis on the importance of sharing data and regarding them as valuable research outputs in and of themselves, regardless of whether or not they have yet been used as evidence for a given discovery; the institutionalization of procedures and norms for data dissemination through the Open Science and Open Data movements, and policies such as those recently adopted by Research Councils UK and key research funders such as the European Research Council, the Wellcome Trust and the Gates Foundation; and the development of instruments, building on digital technologies and web services, that facilitate the production and dissemination of data with a speed and geographical reach as yet unseen in the history of science.
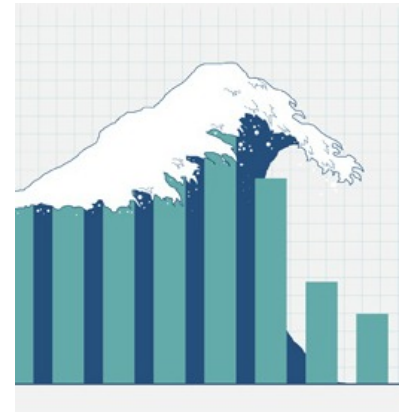
Image credit: W. H. O., Stanley O. Foster M.D., M.P.H,USCDCP (CC0 Public Domain)

This peculiar conjuncture of institutional, socio-political, economic and technological developments have considerably increased international debate over processes of data production, dissemination and interpretation within science and beyond. This level of reflexivity over data practices is arguably the most novel and interesting aspect of contemporary debates over big data. What we are witnessing is thus not the emergence of a wholly new research paradigm dealing with hitherto unseen types of data, but rather the rising prominence of a *data-centric approach* to scientific research, where concerns over data sharing and use in the long term take precedence over immediate attempts to analyze data.

Thus conceptualized, data centrism raises fundamental epistemological issues, which are deeply intertwined with the political challenges posed by big data. What are data, and how are they transformed into meaningful information? What is the status of so-called raw data with respect to other sources of evidence? What constitutes good, reliable data, and how can this be assessed? Should there be restrictions to data dissemination, particularly in cases where the ownership of data is disputed? What role do theory and materials play in data-intensive research, and does it make sense to disseminate big data in the absence of information about data provenance and/or the samples on which data were originally obtained? What difference do the scale (itself a multifaceted notion), technological sophistication and institutional sanctioning of widespread data dissemination make to discovery and innovation? Philosophical analysis can help to address these questions in ways that inform both current data practices and the ways in which have been conceptualized within the social science and humanities, as well as by policy bodies and other institutions.

Scientific research is often presented as the most systematic set of efforts in the contemporary world aimed to critically explore and debate what constitutes acceptable and sufficient evidence for any given belief about reality. The very term 'data' comes from the Latin 'given', and indeed data are meant to document as faithfully and objectively as possible whatever entities or processes are being investigated. And yet, data collection is always steeped in a specific way of understanding the world and constrained by given material and social conditions, and the resulting data are therefore marked by the historical circumstances through which they were generated: what constitutes trustworthy or sufficient data changes across time and space, making it impossible to ever assemble a complete and intrinsically reliable dataset. Furthermore, data are valued and used for a variety of reasons, including as sources of evidence, tokens of exchange and personal identity, signifiers of status and markers of intellectual property; and myriads of data types are produced by as many stakeholders, from citizens to industry and

governmental agencies, which means that what constitutes data, for whom and for which purposes is constantly at stake.

This landscape makes the study of data into an excellent entry point to reflect on the activities and claims associated to the idea of scientific knowledge, and the implications of existing conceptualisations of various forms of knowledge production and use. This is exemplified by ongoing research at the University of Exeter on data handling practices amongst scientists in both the developed world and Low and Middle Income Countries (LMICs). As such research illustrates, what constitutes knowledge, and a 'scientific contribution', varies enormously depending not only on access to data, but also on what is regarded as relevant data in the first place, and what capabilities any research group has to develop, structure and disseminate their ideas. This research has involved a large empirical component involving scientists from a number of laboratories in the UK and in African countries, carried out under the leadership of Professor Brian Rappert and with funding by the Leverhulme Trust. In interviews, scientists were asked to discuss their research, their management of data, how they disseminated their data, and what online data they re-used. From these interviews it became evident that there were a range of material and social aspects of their research environment that played significant roles in their overall data engagement activities.

Interviews carried out in sub-saharan Africa yielded particularly significant results in terms of the political and scientific dimensions of big data sharing. In contrast to most current discussions on data, the issues highlighted in these interviews were often the small, innocuous characteristics of laboratory life in low-resource environments. Thus, issues such as teaching loads, availability of up-to-date computers and software, the age of the research equipment used to produce the data and similar issues featured significantly in how scientists talked about their own research – how they conducted their own research, how they viewed the resultant products of this research, and what they recognized as "data" to share and re-use. Such research clearly demonstrates the importance of scrutinizing all processes involved in data engagement and to recognize the role that research environments play in not only the creation of data, but also their selection, presentation and dissemination. How scientists perceive their research environments, what they recognize as strengths and limitations, and what in these environments pose material or social challenges to data engagement all influence what data travels in or out of any research context.

Such observations raise interesting questions for data studies, as the limitations of overarching descriptions of data or data sharing practices become apparent. Indeed, studying how LMIC scientists engage with the data they create and re-use draws considerable attention to both the material and social aspects of data sharing. The types of data shared and valued, the longevity of these data, and the pathways through which they are disseminated and re-used all have complicated relationships to the research environments in which they are utilized. In consequence, homogenized perceptions of key issues such as what data are, how raw data differs from processed data, and how data ownership can be understood reveal their limitations.

This is part of a wider series on the Politics of Data. For more on this topic, also see Mark Carrigan's Philosophy of Data Science interview series and the Discover Society special issue on the Politics of Data (Science).

Image credit top left: opensource.com (CC BY-SA)

Note: This article gives the views of the author, and not the position of the Impact of Social Science 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 Authors

Sabina Leonelli *is Associate Director of the Exeter Centre for the Study of the Life Sciences (Egenis), Associate Editor of the journal History and Philosophy of the Life Sciences, and Principal Investigator of the ERC Starting Grant, [DATA_SCIENCE]: the Epistemology of Data-Intensive Science. Last year she completed a monograph titled "Life in the Digital Age: A Philosophical Study of Data-Centric Biology" in which her general perspective is articulated*

(under review with Chicago University Press). She can be found on Twitter @sabinaleonelli *and the research group* @DataScienceFeed*.*

Louise Bezuidenhout *(University of Exeter and University of Witwatersrand, South Africa) is an empirical ethicist whose research focuses on issues relating to responsibility and openness in the life sciences. The majority of her research focuses on life science research in developing countries – particularly sub-Saharan Africa. She is fascinated by how and why developing country scientists are represented (or not) in international life science ethics discourse and whether STS research and policy development can be made more sensitive to reflect the pressures of these research environments.*