Final Version Accepted in December 2016 for publication in Science, Technology and Human Values, Special Issue "Data Shadows: Knowledge, Openness and Absence" (2017)

Introduction

Data Shadows: Knowledge, Openness and Absence

Sabina Leonelli, s.leonellli@exeter.ac.uk

Brian Rappert

Gail Davies

Exeter Centre for the Study of the Life Sciences, University of Exeter, UK

Current discourse around data, and particularly that associated with "Big" and "Open" data (e.g. Kitchin 2013; Science International 2015), is infused with the importance of the available, the pre-existing, the present. Across domains as diverse as scientific research, investigative journalism and financial services, data are typically regarded as "givens": things that are, and that can be marshalled into evidence for knowledge claims. The conception of openness, as seminally phrased by the Open Knowledge Foundation, presupposes the potential to make use of that which already exists: "Open data and content can be freely used, modified, and shared by anyone for any purpose" (Open Knowledge Foundation 2014). In this formulation, data are not only taken to be unobstructed and accessible, but are also conceptualized as discrete units that can be easily identifiable, are stable in their format and content, and can be moved across a range of contexts.

However, data are not always stable, ready-made objects. Even data produced by large experimental apparatus or digital tracking services, such as sequencing data in biology or health data produced by smartphone apps, require processing in order to be usable as evidence, and such processing requires resources, relevant expertise and appropriate technologies, which in turn can affect the medium in which data are circulated as well as the information that they are taken to carry (Edwards 2010; Edwards et. al 2011; Leonelli 2016). Furthermore, getting data to move is often dependent on the viability of data as a form of commodity, that is, as something that can be exchanged, sold, acquired and used in a variety of ways and purposes, but in an exclusory manner (Sunder Rajan 2006, Farquhar and Sunder Rajan 2014). This in turn creates complex pathways for data to travel. How they travel depends on the value of datasets for different stakeholders. These determinations often result in data being out of reach for at least some of their potential publics.

This special issue highlights how the preoccupations with making data present can be usefully analyzed and understood by tracing the related preoccupations with what is unavailable, inaccessible or absent, which unvaryingly but often implicitly accompany debates about data and openness. We intend absence as an umbrella descriptor to refer to how data are missing, incomplete, unreliable, ignored, unwanted or untagged. The reasons for such states are many. Data can be hard to capture, store, perceive and disseminate, depending on their format, the technologies available for processing, and the degree of commitment and capital underlying these efforts. They can be imagined, willed and strategized about, without being accessible or even obtainable in some ready fashion. And sometimes, rather than providing evidence for what is there, they provide evidence for what is not. Paying attention to what is *not* given, and how such absences infuse efforts to make

use of data, provides a fruitful analytic lens through which to make sense of contemporary information landscapes, and of the varied and multifaceted expectations arising from and inspiring those landscapes.

The papers in this special issue exemplify this approach by examining attempts to locate and handle data of various kinds in research areas as diverse as seismology, economic modelling, biology, and archaeology, across different historical periods and geographical locations. What brings these cases together is a sensibility to the diverse and conflicting attitudes of researchers toward the idea of "making data open," and related difficulties in identifying what constitutes data in the first place, for which purposes and under which material and social conditions. The papers thus contribute to an emerging historical and sociological scholarship on the extent to which notions of "presence" and "absence" come bundled together as part of efforts to make data open (e.g. Hilgartner 2012, Rappert 2015). At the same time they provide a critical framework for understanding when and how certain objects and artifacts come to be viewed as data, and in relation to which practices, skills, interests and accountabilities. Examining data that are not there, not readily available, and/or not usable towards proving claims or fostering discoveries brings us to confront the significance of what is not typically recognized as knowledge — what is invisible, tacit, ignored, denied, expected, forbidden, private, inaccessible, unknown or unexplored.

When looking at the diversity of ways in which data can be approached as "present" or "absent," one inevitably comes across varied forms of inter-relation among these notions. What is absent clearly depends on what is present, as exemplified by situations where a given dataset is posited as missing or incomplete because of a lack of time or resources required to obtain it. The gathering of data also invariably produces a sense of what has yet to be

obtained, particularly given the impossibility (long noted by philosophers) to ever capture all the possible attributes of a given phenomenon. This understanding of data absence as yearning or potential becomes even more pronounced in the face of unfulfilled promises of "comprehensiveness" and "completeness" of Big Data (Leonelli 2014). For a researcher at a loss for how to analyze the hundreds of existing datasets of potential relevance to a given investigation, or a curator tasked with sorting countless stacks of untagged data, what is manifestly there can imply a sense of what will remain effectively out of reach. For an activist reading through official documents with visibly redacted text, the glaring gaps in what is made available can foster a conviction that matters of some significance must lie under black ink—whether or not this is in fact the case. And for any given dataset that is being disseminated, there may not be enough information (say, due to the unavailability of original samples or appropriate metadata) to be able to effectively interpret it as evidence.

It is the variable and multifaceted nature of the relation between absence and presence that brought us to focus on the notion of *data shadows*. It is of course possible to interpret the phenomenon of shadowing in a linear fashion, where light represents knowledge (or at least the means to it) and shadows thereby its absence. However, as exemplified by any museum exhibit or art gallery, shadows can be used to hide things as well as to make things more clearly noticeable, or to emphasize aspects of an object on which one wants to draw attention. Furthermore, it is unclear whether shadows should be regarded as significant entities in their own right, or whether they are better conceptualized as secondary to something else. There is thus an ambiguity and a strategic relationality to shadowing processes that parallels the relational nature of data, and the multiplicity of motives, goals and conditions through which data may be construed as (in)significant, partial or complete, (un)intelligible or (in)accessible. Whatever sort of want of illumination characterizes a shadow, it is not the

kind that necessarily blocks out from view whatever it envelopes. Similarly, claims around the absence of data are often accompanied both by a clear vision of what data would be desirable and why, and by demands for additional activities, resources or skills to bring such data into the light. The idea of data shadows is thus meant to disrupt a simplistic and straightforward understanding of data as either "there" or not, and instead to foster critical questions around why, how, for whom and when data are perceived as available, portable, and/or meaningful. The concept of data shadows also has an aesthetic element, drawing attention to data visualization practices and how the patterning of light and shade can act to clarify, obscure and attract speculation at the same time (Coopmans 2014; Davies 2014; Smith 2015).

Pursuing the study of data as shadows is challenging because it highlights the empirical, methodological and conceptual challenges involved in defining, identifying and tracking data. The overall approach adopted here is to move away from conceptualizing data as immutable commodities wedded to a particular form, setting and set of expectations. Instead, we focus on the shifting value and embodiments of data, their situated, time-dependent qualities, and the extent to which their perceived significance as forms of evidence, commodities, and/or tokens of personal identity relates to their availability, format and use. In other words, we attend to the negotiations regarding what counts as data, for whom, when, where, why—and how this changes, and what is regarded as missing in such processes. As part of this approach, these papers also embrace different timespans, ranging from contemporary practices and debates (Levin and Leonelli) to mid-20th century historical episodes (Aronova) to present day re-analyses of past activities (Wylie, McGoey). This mixture helps to develop the sense in which "the same" datasets may be highlighted as available or not, desirable or unwanted, significant or insignificant at different moments, with

varying results both synchronically (across different contexts at the same time) and diachronically (within the same context over an extended period of time). This resonates with existing scholarship on how the organization, formatting and visualization of data matter to their subsequent analysis and to whether they are used or not (e.g., Bowker 2000; Hine 2006; Borgman 2015; Leonelli 2016). It also builds on studies of ignorance (Gross and McGoey 2015; Proctor and Schiebinger 2008), what is impossible to know (Gross 2012; Wynne 1992), and what is inaccessible or secret (Rappert 2009, 2015; Balmer 2013), by examining the circumstances and implications of shadowing for current discourse around data openness and disclosure.

As a more detailed introduction to the individual contributions within this collection, we prefer to avoid a simple overview of their core arguments—which you can access from the abstracts—and instead ponder two central questions for all of them: why do these papers engage with data practices, and how do they do so? The very consideration of data as objects of study requires some degree of reflexivity, with analysts needing to be particularly alert to their own goals and the ways in which they position themselves and their "evidence" with respect to the matters under scrutiny. Methodologically, this poses the thorny issue of which kinds of skills and background knowledge are desirable - or even required - in order to understand and interpret data practices within highly specialized fields. These papers show how the analysis of data use can involve varying levels of technical awareness and interaction with data users, as well as diverse disciplinary expectations and genealogies. The authors' long-term engagement with their area of interest, and their diverse disciplinary backgrounds ranging from sociology to anthropology, philosophy and history, ensure a multifaceted and rich examination of how data are used as evidence, and the implications that such practices have for knowledge-making processes.

The first of our papers, Alison Wylie's "Old Data made New Again: How Archaeological Evidence Bites Back," assesses epistemic strategies used by archaeologists to tackle concerns about what is missing, absent, invisible from the historical record, and, more specifically, the ways in which data collected in the past (so-called "legacy data") can be re-analyzed and re-used for different purposes, including to challenge the very theories that they once inspired. Central to her argument are the temporal conditions often faced in the field; namely, archaeologists must not only locate and then make sense of what has been obscured in time, but they do so in conditions in which each generation of archaeologists is dependent on the research traditions, tacit skills, professional expectations, and agendas of those that came before them. As these matters are not necessarily easily available for inspection, unacknowledged presumptions and understandings can be incorporated within what is taken to be "data" from the past. Wylie delves into several examples of archaeological practices to explain how such intergenerational dependency does not necessarily undermine or delimit claims made on the basis of such evidence, but rather can foster practitioners' ability to develop new knowledge. To do so, she builds on her extensive experience in the historical study of archaeological work, which enables her to pick cases of particular poignancy for documenting the tight interplay between the ways in which data are conceptualized, the material properties and affordances of different types of data, the conditions under which such different data types are put to use as evidence for knowledge claims, and the extent to which they can be accessed and shared among communities of practice. Legacy data have "purchase" precisely because they are regularly reconsidered and re-contextualized by groups of researchers with different goals and commitments. In line with her philosophical work on pluralism and the "epistemology of things" in scientific practice (Wylie 2002, 2006), Wylie thus illustrates how diversity in approach and intent, and its impact on what is taken to be

significant data, strengthens the epistemic fruitfulness of practices of data analysis, rather than undermining it.

In "Earthquake Data Prospecting and Mannar's Cold War," Elena Aronova advances a sense of how questions about what counts as worthwhile data can be tied to the purposes for which data is marshalled as evidence. She does that through a detailed historical comparison of contrasting visions for seismology within and between national contexts as well as over time, taking the international political frame of the Cold War as her prime context, and the Soviet amateur seismologist Vladimir Mannar as her central character. Her account thus centers on the actions of one individual, using them as an entry point into a complex nexus of scientific innovation, governmental intervention and the overarching relevance of international relations to shaping seemingly microscopic events and decisions (such as, what forms of data are preferred for documenting seismic activity, and who participates in their collection and management). For Mannar, the start of the Cold War, with its as yet uncertain and undefined social and regulatory structures, provided an opportunity to propose and advance a form of citizen-based seismology—one that would contrast with its Western counterpart fields by marshalling citizens to produce evidence directed toward predicting earthquakes. Such "socialist seismology" challenged the idea of an expert-driven research agenda and called into question what kinds of equipment should be used and thereby which data should be produced. The development of atomic and nuclear weapons by the East and West, however, would radically transform the terrain in which disputes about the reconciliation of openness, professionalism and ideology unfolded. During the 1950s and 1960s, Soviet seismology developed into a technologically sophisticated form of "Big Science" bound with military requirements and restrictions, with experts and large investments reasserting themselves as sources of authority and legitimation. This meant a decided shift away from the equipment,

infrastructure, working style and aims of the science envisioned by Mannar, and it was an important precursor for the large-scale, globalized and highly regimented efforts of data collection that have come to define contemporary science (see Davies et. al 2013). Aronova builds her investigation on her substantial expertise in analyzing the history of big environmental science, as well as her cultural and linguistic positioning as a Russian scholar working in the United States and Europe.

Linsey McGoey's article "The Elusive Rich" is also grounded in historical research, but it moves away from the focus on one case to considering shifts in data practices across the history of a whole discipline, with the aim of illustrating how disciplines can become blinkered to matters that were once central preoccupations. Her target is the treatment of measurements of wealth in economics. This she tackles first by historical overview and then by an analysis of contemporary debates, charting the contests over the relevance of the notion of the marginal productivity of income distribution—a move made possible by her interdisciplinary background in both the sociology and the history of economics. By tying the fading attention to questions about the legitimacy of income generation, and the resulting patterns of income inequality, to limitations in data collection and analysis in mainstream economics, she further shows how a particular and restricted understanding of wealth distribution has emerged - one wherein the wealthiest have been rendered invisible in statistical calculations. Her long-dureé approach enables her to document how this understanding developed over time into a self-reinforcing cycle wherein what data are available to feed models, which again constrain data collection and interpretation (see also Morgan 2012). The stakes in this process of producing practices of ignoring are profound: even as debates about inequality abound in the social sciences and public life, central

questions about what constitutes legitimate wealth creation remain out of both scholarly and public discussion, partly due to the lack of evidence.

Continuing with contemporary debates but turning to micro-institutional practices, Nadine Levin and Sabina Leonelli use interviews and ethnographic data to reconstruct the painstaking, question-begging, and tension-ridden means whereby researchers in biology and bioinformatics seek to make their science and data "open." By stressing the importance of making research as accessible, transparent and re-usable as possible, the Open Science movement seeks to enforce the accountability of science to the public, to increase equality and participation, and to foster its ability to generate new knowledge (Hey et al. 2009; Royal Society 2012). Underling this imperative to be "open" is the notion that making such materials available will increase productivity, responsibility, reproducibility, and innovation in science. Based on research untaken with UK-based scientists who are themselves active proponents of Open Science efforts, as well as deep and ongoing engagement with contemporary policy meant to foster and regulate the uptake of openness by researchers, Levin and Leonelli argue that the processes of making open are hard to discipline or streamline. These processes are contingent on the specifics of laboratory research practices and highly indeterminate in their possible forms, a situation that makes it unfeasible to regard "openness" as a definite good in and of itself. Efforts directed towards the achievement of openness entail acts of valuing wherein the question of what to make available depends on what types of scientific labor, materials and resources are deemed important and what types are not. The struggles involved in making such judgments, and in establishing who is ultimately responsible, raise wide-ranging questions about the constitution of research infrastructures and professional beliefs that govern the valuation and distribution of data. As one aspect of their argument, Levin and Leonelli suggest how the emphasis on access,

dissemination, and usability of data is made dependent on the viability of forms of commercial capture.

How these papers differ in conceptualizing and using empirical evidence—in other words, how they identify and employ data in support of their arguments—is notable and deliberate. By bringing together these approaches, we highlight the potential fruitfulness of various approaches to data collection, as well as the importance of bringing comparative and historical considerations and research into an evaluation of scientific practices. Each of these papers exemplifies a mode of engagement with synchronic and diachronic dimensions of specific uses of data, thus illustrating the variety of ways in which comparison and history can enter STS analysis. Each also exemplifies the importance of reflexivity in using data to write about data. At the very least, keeping an eye on the genealogy of data practices and their contextual nature can help STS scholars to reflect on their own understandings of empiricism and the extent to which methodological choices resonate with their interests, preferences and sensitivities in exploring social worlds.

In advancing themes, such as the generative and consequential potential of absence or the distributed yet path-dependent nature of the movement of data, it is tempting to conclude this introduction by slipping into an asymmetric empiricist language wherein the data and information given in these articles is itself treated as readily available, pre-existing, the present. The use of data in seismology, archaeology, biology and economics is here presented as historically contingent and ambiguous, but what about the claims made by researchers operating under the broad banner of Science and Technology Studies? One way to answer this question would be to insist that our own analysis is based on highly corroborated claims and observations that are just…well... plainly there as you can read for yourself. This

collection then could be positioned as "filling important gaps" in STS scholarship and elsewhere, and providing new firm ground on which to analyze data shadows in the future.

We as editors would like to resist such a move. Our attention to why the authors engage with data practices and how do they do so is meant to underscore how their endeavors are situated in relation to certain goals and commitments as well as particular types of data, while at the same time lending presence, credibility and visibility to the very data that they discuss. In positioning concerns with data practices and policies in relation to their own experience and expertise, the commentaries by Rachel Ankeny, Brian Balmer, Carlo Caduff, and Sally Wyatt offer further arguments along these lines. More than just closing with a negative refrain about the cautions of doing social research, we end with a call for STS research to attend to the relevance of data shadows in their own analyses. In the spirit of acknowledging difficulties of doing, we suggest that what is needed are ways of acknowledging and articulating the salience of absence and the shadowy nature of the evidence we utilize in our work—with the analyses in the special issue themselves suggesting the possibilities, contradictions and complications that "opening up" can entail.

Acknowledgments

The authors gratefully acknowledge the Humanities, Arts and Social Science Fund of the University of Exeter, the European Research Council (award number 335925), a ESRC/AHRC/Dstl project titled "The Formulation and Non-formulation of Security Concerns" (ES/K011308/1) and the University of Chicago for sponsorship of the workshop

"Dark Data" at the University of Exeter in December 2014, within which the papers in this collection were first presented and discussed. The meeting was the final event in the Knowledge / Value Seminar Series, led by Kaushik Sunder Rajan, which focused on historical and emergent relationships between epistemology and value (www.knowledge-value.org/kv5). We thank all participants, and particularly Jennifer Cuffe, Joe Dumit, Mike Fisher, James Griesemer, John Kelly, Sharon Traweek, Niccolo Tempini and Neal White, for excellent discussions and feedback on the ideas presented in this editorial. Our greatest debt is to Kaushik Sunder Rajan for convening such a wonderful group of scholars to explore how questions of both epistemology and value are now at stake across diverse practice and sites. SL also acknowledges the European Research Council (award number 335925) for funding her research. Finally, we thank the editorial team of ST&HV for their assistance and support in developing the special issue.

Authors' Biographies

Sabina Leonelli is Associate Professor in Philosophy and History of Science and Co-Director of the Exeter Centre for the Study of the Life Sciences, where she leads the Data Studies research strand (www.datastudies.eu). Her research focuses on the philosophy, history and sociology of data-intensive science, especially the research processes, scientific outputs and social embedding of Open Science, Open Data and Big Data. She published widely in philosophical, STS and biology journals, and is the author of *Data-Centric Biology: A Philosophical Study* (2016).

Brian Rappert is Professor of Science, Technology and Public Affairs at the University of Exeter. His long term interest has been the examination of the strategic management of

information; particularly in the relation to armed conflict. His books include *Controlling the Weapons of War: Politics, Persuasion, and the Prohibition of Inhumanity*; *Biotechnology, Security and the Search for Limits*; and *Education and Ethics in the Life Science*. More recently he has been interested in the social, ethical, and political issues associated with researching and writing about secrets, as in his book *Experimental Secrets* (2009) and *How to Look Good in a War* (2012).

Gail Davies is Professor in Human Geography at the University of Exeter. Her research charts the changing geographies of laboratory animal science and seeks to develop innovative ways of supporting policy and engaging publics with complex issues in science and technology. She is currently interested in how the experimental practices of collaboration and interdisciplinarity, across diverse bodies from the UK Animals in Science Committee to the artist group The Office of Experiments, bring different issues to light and shade.

References

Balmer, Brian. 2013. *Secrecy and Science: A Historical Sociology of Biological and Chemical Warfare*. Farnham: Ashgate.

Borgman, Christine L. 2015. *Big Data, Little Data, No Data*. Cambridge, MA.: MIT Press.

Bowker, Geoffrey C. 2000. "Biodiversity Datadiversity." *Social Studies of Science* 30(5): 643-683.

Coopmans, Catelijne. 2014. “Visual Analytics as Artful Revelation.” in Coopmans, C., Vertesi, J., Lynch, M, and Woolgar, S. (eds), *Representation in Scientific Practice Revisited* Cambridge, MA: MIT Press.

Davies, Gail, Emma Frow and Sabina Leonelli. 2013. “Bigger, Faster, Better? Rhetorics and practices of Large-Scale Research in Contemporary Bioscience.” *BioSocieties* 8(4): 386-396.

Davies, Gail. 2014. “Searching for GloFish®: Aesthetics, Ethics, and Encounters with the Neonbaroque.” *Environment and Planning A* 46(11): 2604-2621.

Edwards, Paul N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Edwards, Paul N., Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L.Borgman. 2011. "Science Friction: Data, Metadata, and Collaboration." *Social Studies of Science* 41(5): 667-690.

Farquhar, Judith and Sunder Rajan, Kaushik. 2014. "Introduction." Special Issue on Knowledge/Value: Information, Archives, Databases. *East Asian Science, Technology and Society* 8(4): 383-89.

Gross, Mathias. 2012. “'Objective Culture' and the Development of Nonknowledge: Georg Simmel and the Reverse Side of Knowing”. *Cultural Sociology* 6(4): 422 – 437

Gross, Mathias and Linsey McGoey (eds.) 2015. *Routledge International Handbook of Ignorance Studies*. London: Routledge.

Hey, Tony et al (eds) 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm

Hilgartner, Steven 2012. "Selective Flows of Knowledge in Technoscientific Interaction: Information Control in Genome Research" *British Society for the History of Science* 42(2): 267-280.

Hine, Christine. 2006. "Databases as Scientific Instruments and their Role in the Ordering of Scientific Work." *Social Studies of Science* 36(2):269-298.

Kitchin, Rob. 2013. *The Data Revolution*. London: SAGE Publishers.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: Chicago University Press.

Leonelli, Sabina. 2014. "What Difference Does Quantity Make? On the Epistemology of Big Data in Biology." *Big Data and Society* 1: 1-11.

Morgan, Mary. 2012. *The World in the Model*. Cambridge, UK: Cambridge University Press.

Open Knowledge Foundation. 2014. *Open Science Definition*. http://opendefinition.org/, accessed 8 October 2014.

Proctor, Robert L. and Linda Schiebinger (eds). 2008. *Agnotology: The Making and Unmaking of Ignorance*. Stanford, CA: Stanford University Press.

Rappert, Brian. 2009. *Experimental Secrets: International Security, Codes, and the Future of Research*. New York: University Press of America.

Rappert, Brian. 2015. 'Sensing Absence' In Brian Rappert and Brian Balmer (eds) *Absence in Science, Security and Policy* London: Palgrave.

Royal Society. 2012. *Science as an Open Enterprise*. London: Royal Society.

Science International. 2015. *Open Data In a Big Data World*. Accessed September 2016.

Smith, Wally. 2015. "Technologies of Stage Magic: Simulation and Dissimulation" *Social Studies of Science* June 45: 319-343. DOI: 10.1177/0306312715577461

Sunder Rajan, Kaushik. 2006. *Biocapital*. Durham: Duke University Press.

Wylie, Alison. 2002. *Thinking from Things: Essays in the Philosophy of Archaeology*. Berkeley, CA: University of California Press.

Wylie, Alison. 2006. "When Difference Makes a Difference: Introduction." *Episteme: Journal of Social Epistemology* 3.1-2: 1-7.

Wynne, Brian. 1992. "Misunderstood Misunderstandings: Social Identities and Public Uptake of Science." *Public Understanding of Science* 1: 281-304.