Forthcoming in Philosophy of Science, 2015

What Counts as Scientific Data? A Relational Framework

Sabina Leonelli

Exeter Centre for the Study of the Life Sciences (Egenis) & Department of Sociology, Philosophy and Anthropology, University of Exeter

s.leonelli@exeter.ac.uk

Abstract

This paper proposes an account of scientific data that makes sense of recent debates on data-driven and 'big data' research, while also building on the history of data production and use particularly within biology. In this view, 'data' is a 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. They do not have truth-value in and of themselves, nor can they be seen as straightforward representations of given phenomena. Rather, they are fungible objects defined by their portability and prospective usefulness as evidence.

Wordcount (excluding acknowledgments): 4744

1. Introduction

Current scientific debates on big data, data-centric research and data infrastructures have reignited social and cultural interest in what counts as data, and under which conditions data are transformed into knowledge. I propose a philosophical perspective that makes sense of these developments, while at the same time building on long-standing discussions of theory-ladenness and inferential processes within philosophy, as well as an awareness of the history of scientific data production and use. I consider data as tools for communication, whose main function is to enable intellectual and material exchanges across individuals, collectives, cultures, governments, and – in the case of biology – species, and whose mobility across these groups is a hard-won scientific achievement. This constitutes a novel perspective within the philosophy of science, whose students have so far emphasized the man-made, situated nature of data production and interpretation, but have paid little attention to the ways in which data are disseminated and shared after they are first generated, and thus to the challenges and ingenuity involved in devising ways to make data travel. Moreover, I argue that data are not just quantities, though their quantification and assemblage into groups plays an important role in their dissemination and use. They are, first and foremost, material artifacts; and their physical characteristics, including their format and the medium through which they are conveyed, are as relevant to understanding their epistemic role as their social and conceptual functions. This position reflects a perspective on scientific epistemology that emphasizes its processual and embodied nature, and seeks to understand science by studying what Hans-Jörg Rheinberger calls the "medial world of knowledge-making" – that is, the practices and instruments

through which research is carried out (Rheinberger 2011, 340). Within this view, unraveling the conditions under which data are created and disseminated is crucial to understanding what counts as knowledge in the first place, and for whom; and to assessing the epistemic value of the various outputs of knowledge-making activities, whether they be claims, data, models, theories, instruments, communities and/or institutions.

In what follows, I briefly review relevant views within the philosophy of science and then present my own framework. I define 'data' as a 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. Data thus consist of a specific way of expressing and presenting information, which is produced and/or incorporated in research practices so as to be available as a source of evidence, and whose behavior and scientific significance depend on the context in which it is used. In this view, data do not have truth-value in and of themselves, nor can they be seen as straightforward representations of given phenomena. Rather, data are essentially fungible objects, which are defined by their *portability* and their *prospective usefulness as evidence*.

2. Data in 20th Century Philosophy of Science

Etymologically, the term data is Latin plural for the expression 'what is given'. Consider two examples of what biologists regard and use as data: measures of positions of gene markers on a chromosome (Figure 1) or photographs taken to document different stages of embryological development (Figure 2). These examples show the resonance that the idea of data as 'given' has in the life sciences. They can

be easily construed as a starting point for scientific reasoning about a variety of phenomena, including genome architecture, the patterns of expression of specific genes and their impact on early development. They are traces left by the process of measurement and manipulation of organic samples undertaken in a biological experiment, and as such they are taken to document, as accurately as possible given the instruments used, features and attributes of a natural entity – in this cases, an organism like a plant (Fig. 1) or a chicken (Fig. 2). This is where the idea of 'raw data' comes from. These images are as close as a biologist gets to documenting aspects of a phenomenon of interest in a way that can inform further inquiry, without necessarily attempting to reproduce or represent the phenomenon itself. Figure 1 is particularly interesting in this respect, as it shows how data can be produced so as to convey information about a target object, without bearing any morphological or conceptual similarity to any of its parts. The interpretation of the scientific meaning of these figures typically comes from the viewer, who decides whether to regard them as evidence, and for which phenomena, on the basis of her interests, background knowledge and familiarity with the procedures and materials from which the data were obtained.

The importance of the observer in attributing meaning to these images points to a paradox that seems to characterize the role of data in science, and provides a starting point for philosophical analysis. It consists of the observation that, despite their epistemic value as 'given', data are clearly made. They are the results of complex processes of interaction between researchers and the world, which typically happen with the help of interfaces such as observational techniques, registration and measurement devices, and the re-scaling and manipulation of objects of inquiry for the purposes of making them amenable to investigation. This is the case for data

produced through simulations, which are perhaps most clearly the result of specific conceptualizations (Winsberg 2010), as well as data resulting from experimental processes of manipulation, which involve recourse to complex apparatus and procedures that embody an interpretation of the world – an example of what Giere (2006) calls the perspectival nature of observation. It is also the case for data generated outside the controlled environment of the laboratory. For example, hand-made notes or photographs of a herd of bison, taken by ethologists to document their observations of bison behavior, are conditioned both by the employment of specific techniques and instruments (type of camera, small or large notebook, specific make of pen) and by the interests and position of the observer (Radder 2006).

The tension between viewing data as instances of the world and emphasizing their man-made nature has acted as a thread for philosophical discussions of scientific methods at least since the scientific revolution. For the most part, philosophers have focused their efforts towards debunking the myth of data as given rather than made, noting that humans are too conditioned by their own assumptions and interests to be able to observe the world objectively. Accordingly, scientific methods have been portrayed as efficient means to moderate, and where possible annihilate, such subjectivity – an achievement that presupposes the recognition that what one takes to be a fact about the world may well be a fallacious impression generated by the senses. Philosophers' general mistrust of sense perception is often linked to a portrayal of induction as an untrustworthy method, insofar as it seems to assume that there can be observations so reliable and fact-like that one can infer truthful generalisations about the world from them. A vocal advocate of this view was Pierre Duhem, whose writings inspired later ideas about the "theory-ladenness" of data (the inevitable

influence of theoretical presuppositions on data collection, selection and interpretation; Bogen 2010, Schindler 2013).

The theory-laden, man-made nature of data makes it difficult to conceptualize them as providing objective evidence for theories. Nevertheless, most Anglo-American philosophers writing between the 1950s and the 1990s conceived of data as the ground on which theories are validated, which therefore needs to be reliable and trustworthy. This requirement does not fit well with the realization that data reflect scientists' interests, background knowledge, location, instruments and research strategies. It also runs against the insight that "publicly available data typically cannot be produced except through processes whose results reflect the influence of causal factors that are too numerous, too different in kind, and too irregular in behavior for any single theory to account for them" (Bogen 2010, 18). How can data, understood as an intrinsically local, situated, idiosyncratic, theory-laden product of specific research conditions, serve as confirmation for universal truths about nature?

The formulation of this question underscores a view of scientific knowledge as a set of universally valid claims about the world, in which theories and explanations are conceived as the main product of research, and hence as the scientific elements that are most deserving of philosophical scrutiny. This view has been challenged by the so-called 'practice turn' within the philosophy of science, which emphasizes the fruitfulness of examining actual processes of discovery, rather that their *post facto* reconstruction. In particular, the study of scientific models, whose epistemic role varies depending on their concrete features and their users's interests, demonstrated the philosophical import of studying research practices in detail (e.g. Morgan and Morrison 1999). Even within this approach, however, attention has been focused on processes of data production and on how those affect inferences from data to

knowledge claims, rather than on the status of data as research outputs and the ways in which they are mobilized to foster inquiry. Hence, we are left with a conceptualization that goes a long way towards explaining the formation of patterns and generalizations, but does not help to understand the unique status and research functions of data themselves – a shortcoming made evident by the relative inability of philosophers, so far, to intervene meaningfully in scientific debates on data-intensive methods.

3. Data Movements in the Philosophy of Experiment

The philosophy of experiment constitutes an exception to this general disregard for data as epistemic elements. Ian Hacking contributed a definition of experimental data as *marks* produced by human interactions with research instruments. By focusing on the material circumstances in which data are generated, Hacking's account remains agnostic about the epistemic role that data may play in scientific inquiry, and stresses instead the constraints and opportunities provided by the manifold of formats and shapes in which data are produced in the laboratory – comprising "uninterpreted inscriptions, graphs recording variation over time, photographs, tables, displays" (Hacking 1992, 48). Hacking's work inspired the seminal account by James Bogen and James Woodward (1988) of the relationship between data production and the production of claims about phenomena, which resurrected a conception of data as things that can be straightforwardly observed.

Hans-Jörg Rheinberger (2011) also took inspiration from Hacking in his analysis of data as things that can be stored and retrieved, and are thus made durable. His conclusions differ from Hacking's insofar as he does not view the marks produced by

scientific instruments – which he calls "traces" or "signals" – as an example of data. Rather, he conceives of data as the result of further manipulations performed with the purpose of storing the original marks and making them available to others. As an example, Rheinberger uses the first DNA sequence gel produced by Fred Sanger and collaborators in 1977 to visualize the molecular structure of the DNA sequence of bacteriophage PhiX174. The gel generates discrete stripes of varying lengths on a photosensitive plate, which Rheinberger interprets as *traces* generated by this laboratory technique. The abstraction of these stripes into a chain of symbols standing for the four nucleid acid bases, GATC, is what Rheinberger views as transformation of traces into *data* (ibid., 6-7). This account benefits from the extraordinary success of the use of letters as symbols for nucleobases, which can be easily circulated and whose format has certainly facilitated the success of molecular biology. Rheinberger also builds on Bruno Latour's notion of "chains of inference", which highlights how the establishment of knowledge claims is grounded in the production and movement of "immutable mobiles" – that is, objects that can serve as anchors for knowledge claims thanks to their stability across contexts (Latour 1999).

Both Latour and Rheinberger recognize that the marks produced in the course of research need to be processed in order to travel, and that traveling is crucial to their functioning as evidence. Taken jointly, their accounts contribute a key insight to my analysis, which is the emphasis on the epistemic importance of the mobility of data and the labor required to realize it. What I do not share with them is the emphasis on stability. When traveling from their original context of production to a database, and from there to a new context of inquiry, biological data are anything but stable objects. My work on biological databases (2009, 2010, forthcoming) demonstrates how the procedures involved in packaging data for travel involve various stages of

manipulation, which may happen at different times and may well change the format, medium and shape of data. This is the case even when data are presented in apparently straightforward formats, such as the sequence data used in Rheinberger's example. While the use of letters to indicate nucleobases is among the most recognizable and universally intelligible symbols in contemporary biology, biologists and curators have a vast choice of file formats in which those data could be stored and visualized – for instance, the Staden format presents the letters in succession ('GGTACGTAGTAGCTGCTACGT'), while both the database Ensembl and the sequence repository GenBank uses provides a variety of possible formats depending the methods used to produce sequences and the users' interests.[1] The subsequent travel of sequence data is certainly affected by these choices (Fry 2007). I agree with Rheinberger and Latour that this process affects the ways in which data are used as evidence. Does it warrant a distinction between the traces obtained through scientific intervention and their 'packaged' versions, ready for travel? I think not.

Packaging happens at several stages of data travel, and is often implemented already at the point of data production. As Rheinberger recognizes, increasing amounts of biological data are generated digitally, to make it easier to handle them computationally and submit them to databases. Given the iterativity characterizing the processes through which data are produced and disseminated (Chang 2004, O'Malley 2011), trying to differentiate between the marks produced by scientific inquiry and those obtained through further manipulation seems arbitrary. Scientists engage in data generation in full awareness that the outputs of that activity need to travel beyond the boundaries of their own investigation. This awareness is built into the choice of

[1] See the overview of sequence data formats at http://www.compbio.ox.ac.uk/bioinformatics_faq/format_examples.shtml (accessed February 2014).

instruments used; the recording of procedures and protocols carried out in lab books; and the decisions about how outputs may be stored, shared with peers, fed into models, integrated with other data sources. Scientists recognize the need to package data for travel as an essential requirement for knowledge production, which underlies both the planning of research, the generation of data and their further elaboration. This undermines the idea that we can neatly distinguish between data as traces directly derived from investigation and their further manipulation, abstraction and translation into a variety of formats.

4. Data as Portable Objects

A better option is give up altogether on a definition of data based on the degree to which they are manipulated, and focus instead on the relation between researchers' perceptions of what counts as data and the stages and contexts of investigation in which such perceptions emerge. I propose to view data as any product of research activities, ranging from artifacts such as photographs to symbols such as letters or numbers, which is collected, stored and disseminated in order to be used as evidence for knowledge claims. This does not mean that whoever gathers data already knows how they might be used. Rather, what matters is that observations or measurements are collected *with the expectation* that they may be used as evidence for claims about the world in the future. Hence, any object can be considered as a datum as long as (1) it is treated as potential evidence for one or more claims about phenomena, and (2) it is possible to circulate it among individuals.

This definition frames the notion of data as a *relational category* that can be attributed to any set of objects, as long as they fulfill the two requirements above. What counts

as data depends on who uses them, how and for which purposes. Data can therefore include experimental results as well as field observations, samples of organic materials, results of simulations and mathematical modeling, even specimens. For instance, a genetically engineered mice colony displaying abnormal behavior may be taken to constitute data for claims such as 'mice with X genetic make-up tends to exhibit behavior Y'. Other types of data supporting the same claim will include photographs and videos of the mice, samples of their blood, their genome sequence and observation notes made by researchers to describe their behavior. A research group investigating evolutionary claims, such as 'bacterial populations exhibit evolutionary novelties as a result of multiple mutations in their genomes', might be interested in a different combination of objects as sources of evidence, such as fitness data for multiple generations, genome sequences of ancestral and evolved strains and photographs of the morphology of colonies at various evolutionary stages.

It follows that the same objects may or may not be functioning as data, depending on which role they are made to play in scientific inquiry. This is a significant insight given the contradictions and uncertainties, evidenced in much scientific and policy literature, about how data should be defined and whether their identity changes whenever they shift format, medium or context. Many participants in these discussions think that, despite the multiple types and uses of data across the sciences, debates concerning data-intensive science should be grounded on a context-independent definition of data. In other words, they view data as *representational* entities, which depict a specific part of reality independently of the circumstances under which they are considered. Under this interpretation, what matters when analyzing data is uncovering which aspects of reality they document; and indeed, their epistemic significance stems from their ability to represent such aspects of reality

irrespectively of the context. The definition of data provided by the Royal Society, which casts them as "numbers, characters or images that designate an attribute of a phenomenon" (Royal Society 2012, 12), can be easily interpreted in this vein.

Despite its intuitive appeal, I wish here to oppose a representational interpretation of data on both empirical and conceptual grounds. I do not think that data necessarily "designate an attribute of a phenomenon". This is, firstly, because researchers often produce data without knowing exactly which phenomenon they may document. Data production, particularly for high-throughput biological data, can and does happen simply because scientists have access to a given instrument and/or because they hope that consulting those data may yield new questions or insights on as yet unknown phenomena (Steinle 1997). Secondly, viewing data as representations of a specific attribute restricts their epistemic role to evidence for claims concerning that attribute. What I wish to highlight, instead, is how the same set of data can act as evidence for a variety of phenomena, depending on how they are interpreted – a feature that I take to be central to understanding the epistemic power of data.

Within this relational framework, it is meaningless to ask what objects count as data in the abstract, because data are defined in terms of their function within specific processes of inquiry, rather than in terms of intrinsic properties. The question 'what is data' can only be answered with reference to concrete research situations, in which investigators make specific decisions about what can be used as evidence for which claims. This position is purposefully not intended to help evaluating the motivations that may push scientists to consider specific objects as data (I think that assessing these choices is a matter of scientific, rather than philosophical, competence). Rather, I am interested in using a study of how data are routinely handled by scientists to

understand the circumstances under which certain objects are successfully used as evidence towards claims.

The only characteristic that I see as essentially tied to my definition of a datum is its *portability*. Portability is a crucial precondition for using data as evidence, because the establishment of scientific claims is widely recognized as a social activity that needs to involve more than one individual. No intellectual achievement, no matter how revolutionary and well-justified, can be sanctioned as a contribution to scientific knowledge unless the individual concerned can express her ideas in a way that is intelligible to a community of peers, and can produce evidence that can be exhibited to others as corroborating her claims. Making data travel is therefore a necessary, though not sufficient, condition for their prospective use as evidence. If data are not portable, it is not possible to pass them around a group of individuals who can review their significance and bear witness to their scientific value. This insight links my approach to that proposed by Lorraine Daston and Peter Galison (1992), who described data as portable, workable and "communal"; and Mary Morgan (2010, 2012), who stressed the crucial importance of movement to assessing the value of data as evidence. The cases of data travel examined by these authors, in conjunction with my study of data curation practices (e.g. Leonelli 2010), demonstrate the diverse extents to which data can be made to travel, and the importance of acknowledging efforts to make data portable as scientific achievements in their own right.

The crucial role of portability is also what leads me to characterize data as *material* artifacts, independently of whether they are circulated in a digital form or not. As emphasized by Hacking, whether they are symbols, numbers, photographs or specimens, all data types need a physical medium in which they can be disseminated. This mundane observation has an important philosophical implication, which is that

the physical characteristics of the medium significantly affect the ways in which data can be disseminated, and thus their usability as evidence. In other words, when data change medium, their scientific significance may also shift. This is notable given the diversity of media that data are likely to encounter in their journeys, as I already stressed in the case of sequence data. This finding underscores the man-made quality of data, and thus the difficulties in viewing them as objective sources of evidence. Rather than viewing this as a problem, I welcome this recognition of how deeply the characteristics of data are intertwined with the specific stages of research in which they are used. Recognizing the ingenuity involved in data dissemination illustrates how shifts in the format and media used to disseminate data are not an obstacle to the use of data as evidence, but rather an essential component of their journeys, without which they would not travel as efficiently (or at all).

5. Conclusion: On the Mutability of Scientific Data

I argued that the same set of objects may or may not be functioning as data, depending on whether it is portable and on which role it is made to play in scientific inquiry. One objection to this view may come from philosophers who distinguish between types and tokens of information, where a type denotes a specific form (a dot on a microarray slate, a letter in a sequence, a dark spot in a photograph of tissue) and a token denotes any physical instantiation of such a form, which would obtain whenever the form was reproduced. In this view, while tokens are material objects, types are immaterial forms whose individual instantiations may be concrete, but whose unique identity is intangible. This is one way to account for the difference between the identity of a dataset as unique source of information and the concrete

instantiations of such dataset whenever it is copied through a multitude of media (e.g. Timpson 2013, Ch. 2). Copyright law makes a similar distinction when discussing the difference between an original work and its copies, and separating the intangible forms to which authorship may be attached from the tangible instantiations to which property claims may be attached irrespectively of whether one is the author or not (Biagioli 2014). My response is that while the distinction between token and types is indeed useful for ascertaining authorship or other forms of property claims, it does not hold when evaluating the epistemology of data. The more data travel, the less clear it becomes who counts as their author and/or owner. Consider a biologist who produced a set of microarray data and deliberates on how to disseminate it. She may wishes to be recognized as its original producer (its author), so she may publish it in a data journal or donate it to a database that promises to mention her as the source. This will tie her name to the format and medium in which the data were originally produced. However, once the data leave her lab and start their journeys across screens, printouts and databases around the world, they will be modified and reformatted so as to fit new approaches and landscapes. Depending on what uses the data are eventually put to, and by whom, those modifications may well prove as relevant to making data into valuable evidence as the efforts of the original data producer. What counts as the 'form' or the 'type' of this dataset? Considering the mutability of data that travel, and how the characteristics of their vehicles affect how those data are read, interpreted and re-used, leads me to conclude that the information content of data cannot be separated from their form: the distinction between types and tokens is meaningless in the context of data dissemination and re-use.

Does that mean that every time a dataset is copied or translated into a new format, its scientific significance necessarily changes? This is not the case in my framework,

because formats or media do not determine by themselves the scientific significance of data, but rather this depends on wider context of inquiry in which they are adopted. The need to recognize the epistemological significance of the variability and polymorphism involved in data journeys is why I insist on avoiding a definition of data that would make it possible to identify and discuss them independently of a specific context. There is no such thing as data in and of themselves, as what counts as data is always relative to a given inquiry where evidence is sought to answer, or even formulate, a question. Data are not only modifiable in principle, but are in fact frequently modified during their travels in ways that profoundly affect their ability to function as evidence.

Acknowledgments

This research was funded by a Visiting Scholarship of the Max Planck Institute for the History of Science (project "Sciences of the Archive") and by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 335925 (project "The Epistemology of Data-Intensive Science"). Many thanks to audiences at the 2014 meeting of the German Network for the Philosophy of Biology in Münster, the 2014 PSA/HSS meeting in Chicago and the 2014 ERC Workshop "What is Data-Intensive Science?" in Exeter, where this paper was presented and discussed; and particularly to Lorraine Daston, James Griesemer, David Sepkoski, Mary Morgan, Staffan Müller-Wille, Thomas Reydon and David Sepkoski for useful discussions.

References

Biagioli, Mario. 2014. "Plagiarism, Kinship and Slavery." Theory, Culture & Society January 21 2014. Accessed February 20 2014.

Bogen, James. 2013. "Theory and Observation in Science." In *The Stanford Encyclopedia of Philosophy (Spring 2013 Edition)*, edited by Edward N. Zalta. Accessed February 20 2014.

Bogen, James, and James Woodward. 1988. "Saving the Phenomena." *The Philosophical Review* 97 (3): 303–352.

Daston, Lorraine, and Peter Galison. 1992. "The Image of Objectivity." *Representations* 40: 81-128.

Fry, Ben. 2007. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. Sebastopol, CA: O'Reilly Media.

Giere, Ronald N. 2006. *Scientific Perspectivism*. Chicago, IL: The University of Chicago Press.

Hacking, Ian. 1992. "The Self-Vindication of the Laboratory Sciences." In *Science as Practice and Culture*, edited by Andrew Pickering, 29-64. Chicago, IL: The University of Chicago Press.

Latour, Bruno. 1999. "Circulating Reference: Sampling the Soil in the Amazon Forest." In *Pandora's Hope: Essays on the Reality of Science Studies*, by Bruno Latour, 24-79.

Leonelli, Sabina. 2009. "On the Locality of Data and Claims About Phenomena." *Philosophy of Science,* 76, 5: 737-749.

Leonelli, Sabina. 2010. “Packaging Small Facts for Re-Use: Databases in Model Organism Biology.” In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, edited by Peter Howlett and Mary S. Morgan, 325-348. Cambridge, UK: Cambridge University Press.

Leonelli, Sabina. Forthcoming. *Life in the Digital Age: A Philosophical Study of Data-Centric Biology*. Chicago: University of Chicago Press.

Morgan, Mary S. 2010. “Travelling Facts.” *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, edited by Peter Howlett and Mary S. Morgan, 3-42. Cambridge, UK: Cambridge University Press.

Morgan, Mary S., and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge, UK: Cambridge University Press.

O'Malley, Maureen A. 2011. "Exploration, Iterativity and Kludging in Synthetic Biology." *Comptes Rendus Chimie* 14 (4): 406–412.

Radder, Hans. 2006. *The World Observed/The World Conceived*. Pittsburgh, PA: University of Pittsburgh Press.

Royal Society. 2012. “Science as an Open Enterprise.” Accessed 14 January 2014. http://royalsociety.org/policy/projects/science-public-enterprise/report/.

Schindler, Samuel. 2013. "Theory-laden Experimentation." *Studies in History and Philosophy of Science* 44 (1): 89-101.

Steinle, Friedrich. 1997. "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* 64: S65-S74.

Timpson, Christopher G. 2013. *Quantum Information Theory and the Foundations of Quantum Mechanics*. Oxford, UK: Oxford University Press.

Winsberg, Eric. 2010. *Science in the Age of Computer Simulations*. Chicago: Chicago University Press.