Submitted to Data Science May 28, 2017

Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology

Sabina Leonelli
Exeter Centre for the Study of the Life Sciences & Department of Sociology, Philosophy and Anthropology, University of Exeter, UK; School of Humanities, University of Adelaide, Australia; s.leonelli@exeter.ac.uk

Abstract (233 words)

This paper reflects on the relation between international debates around data quality assessment and the diversity characterising research practices, goals and environments within the life sciences. Since the emergence of molecular approaches, many biologists have focused their research, and related methods and instruments for data production, on the study of genes and genomes. While this trend is now shifting, prominent institutions and companies with stakes in molecular biology continue to set standards for what counts as 'good science' worldwide, resulting in the use of specific data production technologies as proxy for assessing data quality. This is problematic considering (1) the variability in research cultures, goals and the very characteristics of biological systems, which can give rise to countless different approaches to knowledge production; and (2) the existence of research environments that produce high-quality, significant datasets despite not availing themselves of the latest technologies. Ethnographic research carried out in such environments evidences a widespread fear among researchers that providing extensive information about their experimental set-up will affect the perceived quality of their data, making their findings vulnerable to criticisms by better-resourced peers. These fears can make scientists resistant to sharing data or describing their provenance. To counter this, debates around Open Data need to include critical reflection on how data quality is evaluated, and the extent to which that evaluation requires a localised assessment of the needs, means and goals of each research environment.

Keywords (6): data quality, research assessment, peer review, scientific publication, research methods, data generation

Introduction: Open Data and the Assessment of Data Quality in the Life Sciences

Much of the international discussion around Open Science, and particularly debates around Open Data, is concerned with how to assess and monitor the quality and reliability of data being disseminated through repositories and databases (Science International 2015, Cai and Zhu 2015). Finding reliable ways to guarantee data quality is of great import when attempting to incentivise data sharing and re-use, since trust in the reliability of data available online is crucial to researchers considering them as a starting point for – or even just complement to – their ongoing work (Ossorio 2011, Royal Society 2012, Borgman 2012, Leonelli 2016, Digital Science 2016). Indeed, the quality and reliability of data hosted by digital databases is key to the success of Open Data, particularly in the wake of the "replicability crisis" recently experienced by fields such as psychology and biomedicine (Open Science Collaboration 2015, Allison et al 2016), and given the constant acceleration of the pace at which researchers produce and publish results (Pulverer 2015). However, the wide diversity among the methods, materials, goals, techniques used in pluralistic fields such as biology, as well as the diverse ways in which data can be evaluated depending on the goals of the investigation at hand, make it hard to set common standards and establish international guidelines for evaluating data quality (Cai and Zhu 2015). Attempts to implement peer review of the datasets donated to digital databases are also proving problematic, given the constraints in resources, personnel and expertise experienced by most data infrastructures, and the scarce time and rewards available to researchers contributing expertise to such efforts. This problem is aggravated by the speed with which standards, technologies and knowledge change and develop in any given domain, which makes it difficult, time-intensive and expensive to maintain and update databases and related quality standards as needed.

This paper examines the relation between international discussions around how to evaluate data quality, and the existing diversity characterising research work within the life sciences, particularly in relation to biologists' access to and use of instruments, infrastructures and materials. Since the molecular bandwagon took off in Europe and the US in the 1950s, the majority of resources and attention within biology has been dedicated to creating methods and technologies to study the lowest levels of organisations of organisms, particularly genomics (Testa and Nowotny 2011, Müller-Wille and Rheinberger 2015). This trend is now reversing, with substantial interest returning to the ways in which environmental, phenotypic and epigenetic factors interact with molecular components (Barnes and Dupre 2008, Dupre 2012, Müller-Wille and Rheinberger 2017). However, countries which adopted and supported the molecular approach – also including Japan, China and Singapore – continue to set the standards for what counts as 'good science' worldwide. In practice, this means that the technologies and methods fostered by top research sites in these countries – such as, most glaringly, next generation sequencing methods and instruments - are often taken as exemplary of best laboratory practice, to the point that the use of software and machines popular in those locations is widely used as proxy for assessing the quality of the resulting findings.

This situation turns out to be problematic when considering the sophisticated relationship between the goals and interests of researchers at different locations, the specific characteristics of each target system in biology, and the methods devised to study those systems. These factors may vary and be combined in myriad ways, giving rise to countless different ways to conduct and validate research, and thus to assess the quality of relevant data. It is also troubling when considering research environments that do not have the financial and infrastructural resources to avail themselves of the latest software or instrument, but which are nevertheless producing high-quality data of potential biological significance – because of the materials they have access to, their innovative conceptual or methodological approach, or their focus on questions and phenomena of little interest to researchers based elsewhere. All too often, researchers working in such environments are afraid that lack of access to the latest technologies will affect the quality and reliability of their data, and will make their findings vulnerable to criticisms by better-resourced peers. These fears can result in researchers being unwilling to share their data and/or to describe the specific circumstances and tools through which they were obtained, thus making it impossible for others to build on their research and replicate it elsewhere.

Against this background, this paper defends the idea that *debates around Open Data can and should foster critical reflection on how data quality can and should be evaluated, and the extent to which this involves a localised assessment of the challenges, limitations and imperfections characterising all research environments*. To this aim, I first reflect on existing models of data quality assessment in the life sciences and illustrate why the use of specific technologies for data production can end up being deployed as a proxy for data quality. I then discuss the problems with this approach to data quality assessment, focusing both on the history of molecular biology to date and on contemporary perceptions of technological expectations and standards by researchers in both African and European countries. I stress how technologies for data production and dissemination have become markers for researchers' identity and perception of their own role and status within their fields, in ways that are potentially damaging both to researchers' careers and to scientific advancement as a whole. This discussion is based on observations acquired in the course of ethnographic visits to biological laboratories in Wales, Britain, the United States, Belgium, Germany, Kenya and South Africa; extensive interviews with researchers working on those sites conducted between 2012 and 2016; and discussions on Open Data and data quality carried out with African members of the Global Young Academy (GYA) as part of my work as coordinator for the Open Science working group (https://globalyoungacademy.net/activities/open-science/).[i] I conclude that it is essential for research data to be evaluated in a manner that is localised and context-sensitive, and Open Data advocates and policies can play a critical role in fostering constructive and inclusive practices of data quality assessment.

Existing Approaches to Research Data Quality Assessment

Data quality is a notoriously slippery and multifaceted notion, which has long been the subject of scholarly discussion. A comprehensive review of such debates is provided by Phyllis Illari and Luciano Floridi (2014), who highlight how the various approaches available, while usefully focusing on aspects such as error detection and countering misinformation, are ultimately tied to domain-specific estimations of what counts as quality and reliability (and for what purposes) that cannot be transferred easily across fields, and sometimes even across specific cases of data use. This does not help towards the development and implementation of mechanisms that can guarantee the quality of the vast amounts of research data stored in large digital repositories for open consultation. Data dissemination through widely available data infrastructures is characteristic of the current Open Data landscape, and fits the current policy agenda in making research results visible and potentially re-usable by anybody with the skills and interest to explore them. This mode of data dissemination relies on the assumption that the data made accessible online are of sufficient quality to be useful for further investigation. At the same time, data curators and researchers are well-aware that this assumption is problematic and easy to challenge. This is, first, because no data type is 'intrinsically' trustworthy, but rather data are regarded as reliable on the basis of the methods, instruments, commitments, values and goals employed by the people who generate them (Cai and Zhu 2015); and second, because while it possible to evaluate the quality of data through a review of related metadata, this evaluation typically require expert skills that not all prospective data users possess or care to exercise (Leonelli 2016).[ii]

The problems involved in continuing to develop large research data collections without clear quality benchmarks is widely recognised by academies, institutions and expert bodies involved in Open Data debates, and debates over data quality feature regularly in meetings of the Research Data Alliance, CODATA and many other learned societies and organisations around the world. While it is impossible to summarize these extensive debates within the scope of this paper, I now briefly examine six modes of data quality evaluation that have been widely employed so far within the sciences, and which continue to hold sway while new solutions are being developed and tested.

The first and most common mode of data quality evaluation consists of *traditional peer review* of research articles where data appear as evidence for scientific claims. The idea here is that whenever scientific publications are refereed, reviewers also need to assess the quality of the data used as evidence for the claims being made, and will not approve of publications grounded on untrustworthy data. Data attached to peer-reviewed publications are therefore often assumed to be of high quality and can be therefore be openly disseminated without problems. However, there are reasons to doubt the effectiveness of this strategy in the current research environment. This only works for data extracted from journal publications, and is of little use when it comes to data that have not yet been analysed for publication – thus restricting the scope of databases in ways that many find unacceptable, particularly in the current big data landscape where the velocity with which data are generated has dramatically increased, and a key reason for open dissemination of data is precisely to facilitate their

interpretation. It is also not clear that peer review of publications is a reliable way to peer review data. As noted by critics of this approach, traditional peer review focuses on the credibility of methods and claims made in the given publication, not on data per se (which are anyhow often presented within unstructured 'supplementary information' sections, when they are presented at all; Morey et al 2016). Reviewers are not usually evaluating whether data could usefully be employed to answer research questions other than the one being asked in the paper, and as a result, they provide a skewed evaluation. This could be regarded as an advantage of peer review, since through this system data are always contextualised and assessed in relation to a particular research goal; yet, it does not help to assess the quality of data in contexts of dissemination and re-use. Thus, data curators in charge of retrieving and assessing the quality of data originally published in association with papers need to employ considerable domain-specific expertise to be able to extract the data from existing publications and making them findable and usable. An example of this is the well-known Gene Ontology, whose curators annotate data about gene products by mining published sources and adapting them to common standards and terminology used within the database, which involves considerable labour and skill (Leonelli et al 2011, Blake et al 2015).

Indeed, a second mode of data quality assessment currently in use relies on *evaluations by data curators in charge of data infrastructures*. The argument in this case is that these researchers are experts in data dissemination – they are the data equivalent of a librarian for traditional manuscripts – and are therefore best equipped to assess whether or not the data considered for online dissemination are trustworthy and of good enough quality for re-use. Hence, in the Gene Ontology case cited above, curators not only select which data are of relevance to the categories used in the database, but also assign "confidence rankings" to the data depending on what they perceive as the reliability of the source – a mechanism that certainly assigns considerable responsibility for data quality assessment to those who manage data infrastructures. This solution works reasonably well for relatively small and well-financed data collections, but fails as soon as the funding required to support data curation ceases to exist, or the volume of data becomes so large as to make manual curation impossible. Also, this type of data quality assessment is only as reliable as the curators in charge, especially in cases where data users are too far removed from the development and maintenance of databases to be able or willing to give feedback and check on curators' decisions.

A third mode of data quality assessment is thus to *leave decisions around data quality to those who have generated the data in the first place*, which avoids potential misunderstandings between data producers, reviewers and curators. Again, this solution is not ideal. For one thing, existing databases have a hard time getting data producers to post and appropriately annotate their own data (cases such as PomBase, where over half of the authors of relevant papers post and annotate datasets themselves, are far and few between, and typically occur in relatively small and close-knit communities where trust and accountability are high; McDowell et al 2015). Furthermore, whatever standards data producers are using to evaluate the quality of their data, it will unavoidably be steeped in

the research culture, habits and methods of their own community and subfield, as well as the goals and materials used in their own research. This means that data producers do not typically have the ability to compare different datasets and evaluate their own data in relation to data produced by other research environments, as would be required when assembling a large data infrastructure. Whenever data leave their context of production and enter new contexts of potential re-use, new standards for quality and reliability may well be required, which in turn demands for external assessment and validation from outside the research environment where data were originally generated.

A fourth method for data quality assessment consists in the employment of *automated processes and algorithms*, which have the potential to reduce dramatically the manual labour associated with data curation. There is no doubt that automation facilitates a variety of techniques to test the validity, reliability and veracity of data being disseminated, particularly in the context of data linkage facilities and infrastructures (Kambatla et al 2014, Primiero 2014). However, such tools typically need to make substantive general assumptions about what types of data are most reliable, which are hard to defend given the user-related nature of data quality metrics and their dependence on the context and goals of data assessment. An interesting model for the development of future data quality assessment processes within the life sciences is provided by the many Quality Assessment Tools used to evaluate clinical data in biomedical research, though that approach relies again on the exercise of human judgement, which in turn results in contentious disparities in its application (e.g. Stegenga 2014).

As a fifth option, there have been attempts to *crowdsource quality assessment* by enabling prospective data users to grade the quality of data that they find available on digital databases. While this method holds great promise, it is hard to apply consistently and reliably in a situation where researchers receive little or no credit for engaging with the curation and reuse of existing data sources, and providing feedback to data infrastructures that may enhance their usefulness and long-term sustainability. As a result of the lack of incentive to participate in the curation of Open Data, most databases operating within the life sciences receive little feedback from their users, despite the (sometimes considerable) effort put into creating channels for users to provide comments and assess the data being disseminated. Moreover, it is perfectly possible that users' judgements differ considerably depending on their research goals and methodological commitments.

Given the difficulties encountered by the methods listed above, researchers involved in data quality assessments (for instance, related to data publication or to the inclusion of data into a database) may recur to a sixth, unofficial and implicit method: the *reliance on specific technologies for data production as proxy markers for data quality*. In this case, specific pieces of equipment, methods, materials are taken to be intrinsically reliable and thus to enhance – if not guarantee – the chance that data produced through those techniques and tools will be of good quality. Within the life sciences, prominent examples of such proxies include the use of next generation sequencing machines and mass

spectrometry in model organism biology, microbiomes and systems biology; light-producing reporter genes produced by reputable companies in cell and developmental biology; and de novo gene synthesis and design/simulation software in synthetic biology. These tools are strongly embedded in leading research repertoires within biology, and are extensively adopted by laboratories around the world (Ankeny and Leonelli 2016). They are typically easy to verify, with well-established protocols in place and little additional expertise or labor needed, giving rise to what philosopher Ulrich Krohs calls "convenience experimentation" (Krohs 2012). And they are typically a good fit for existing Open Data infrastructures and formats, which are often developed alongside such technologies as part of the same repertoire (as in the case of sequencing data; Leonelli and Ankeny 2015).

What Technology, for Which Purpose?

It could be argued that researchers in the life sciences have long been dependent on instruments for data classification and interventions on organisms, and that given the crucial role of such tools in knowledge production, reference to the use of technologies as a proxy for data quality is epistemically justified – particularly when this metric is used in conjunction with other evaluation procedures, such as those described above. In this section, I counter this position by pointing out that it takes no account of the powerful market forces at play in the provision and dissemination of (often extremely expensive) research technologies, and the distortions that this involves when it comes to evaluating what counts as an ideal research environment – and thus as "best practice" – in biological research.

The power and size of the industrial complex devoted to the development and mass production of research technologies has grown exponentially since the 1950s, in parallel with the growth of the scale and size of biological research worldwide; and with it, the costs, marketing and competition around research tools have spiraled up (Rajan 2006). The production of lab equipment is now big business particularly in the United States and Europe, with the top 25 companies accounting for 23.6 billion dollars in sales in 2015 alone (Thayer 2016). This explosion in the market, alongside the priority accorded to technologies that could capture digitally data pertaining to the molecular level of organization of organisms, ended up fueling a perception of sequencing tools and related equipment as an essential part of any biological investigation, whose utilization lends credibility to research results. The monopoly held by the companies Affymetrix and Illumina over the production of genetic assays and microarray data which endured from the mid-1990s to the late 2000s when competitors emerged, is but one example of the way in which competitive marketing has made its way in the best funded labs around the world, and thus into researchers' ideal of what a perfect research setting needs to look like (Rogers and Cambrosio 2007, Research and Markets 2016). To keep up their revenue, technology providers have a strong incentive as well as the means to set standards for what count as acceptable data in any one area, by pushing the idea that using their tools guarantees high-quality data. The abundant advertisement of lab equipment to be found in any international science journal, including

leading publications such as *Nature* and *Science*, bears testament to this phenomenon; as do the large spaces allocated to the marketing of research technologies within any respectable international congress in the life sciences. Thus, market forces introduce incentives for biological labs to possess specific pieces of machinery that are not necessarily linked to achieving research excellence, but rather to the desire to be able to use standards and specifications of data formats that are promoted internationally through the marketing of these technologies.

Given this situation, it is not surprising that the use of technology as proxy for data quality continues to occur among editorial boards, research institutions and funders, and international research consortia who have the power to determine what counts as "good" research practice, including what counts as data quality. This is acknowledged by biologists working in UK and US labs that I have interviewed over the last few years. Even in very well-equipped laboratories at established and well-funded research institutions, researchers complained to me about their access to instruments and related materials. Most notably, when interviewed on practices of data production, dissemination and re-use, researchers displayed insecurity and discomfort around the state of their equipment and of their ability to use it. For instance, I encountered statements of unease around:

- instruments and materials that their lab did not possess and which the researcher in question did not view as essential to her research, but whose use was requested as ulterior confirmation of her findings by the reviewers of the journals in which she had tried to publish;
- the extent to which the use of the equipment at hand was being maximised for the benefit of research. For example, many UK-based research groups interviewed over their use of high-throughput technologies for data production expressed worries around the level of technical skill required to use those tools, the proficiency with which lab members were operating the technology, whether their lab was making the most of such tools;
- the extent to which possessing a given piece of equipment may constitute a competitive disadvantage, but forcing researchers to choose specific research directions in order to make sure that the investment made in the machines is justified. This trend is most evident and best documented in the case of genomic sequencing, a technology whose development required a high level of investment by governmental agencies – an investment on which funders expect to see returns, thus pushing researchers to capitalise on the resulting genomic data (e.g. Hilgartner 2017);
- the fast-moving technological developments in the relevant field, which makes even very well-established and visible research groups fearful of being left behind or unaware of the latest instruments and techniques on offer (see also Levin et al 2016).

Such widespread insecurity and fears in relation to research environments in the life sciences is not surprising, given the variety of equipment on offer, the high level of technical skill required to use it, the high costs involved in assembling

and maintaining an internationally recognised research lab, and the constantly evolving market. Even within well-resourced labs based in prominent and rich institutions, researchers rarely have access to all the technology that they view as potentially relevant to their various projects; and worries around being "locked-in" a given technology, and/or unable to use it in the most fruitful way, are widespread across highly provisioned research environments. Such worries have arguably grown in parallel to the increased emphasis on transparency and accountability recommended by Open Science guidelines, and the related explosion of replication experiments pointing to the irreproducibility of many supposedly well-established results. These developments have an enormous potential to improve scientific methods and communication strategies, by eliciting a healthy and necessary preoccupation with producing high-quality, well-justified, intelligible and re-usable results. At the same time, it is important to recognise that Open Science guidelines and replicability requirements also undermine the implicit trust among peers that so far characterised many areas of biological inquiry, with several researchers confiding to me that they fear being found wanting by colleagues and worry constantly about whether their laboratory set-up and related skills will be recognised as sufficient and well-suited to their line of inquiry.

Implications for Low-Resourced Research Environments

Within high-resourced research environments, there are many mechanisms in place to mitigate the potentially harmful implications of this breakdown in trust, and to turn Open Science requirements into an opportunity to develop common standards of best practice. First, researchers working in well-funded labs have the means and opportunity to constantly exchange personnel, visits and equipment (and related reagents and materials) with each other, so as to learn from each other and work collaboratively to maintain quality standards in their field. Secondly, researchers based in internationally visible and powerful institutions are in a good position to propose specific (uses of) technology as gold standard for their peers, and have the resources to adapt quickly to emerging repertoires, instruments and trends. Furthermore, such researchers typically have access to at least some well-recognised equipment, which they can make accessible to staff from other labs in exchange to access to other tools.

These strategies do not always work in the context of an increasingly diverse and globalized research workforce, and particularly not in research locations which are not easily reachable because of their geographical location, and/or where there are stark inequalities in access to technologies, related infrastructures and materials, and internationally visible and acknowledged collaborative networks. Many biologists are based in contexts where access to the latest and most expensive technology is not guaranteed, financially viable or even relevant - for instance, because research focuses on areas such as morphology, physiology, developmental biology, botany, immunology and ethology, where access to the most recent genome sequencer may not matter since the production of molecular data may not be the focus of inquiry. Whether or not it affects research practice and outcomes, lack of access to the latest equipment can make

researchers insecure on several fronts, including: what they do not have access to, and how important it may be for their work and/or adherence to international expectations; technical skills that they may lack; and the very reliability and quality of their data, regardless of whether that depends on having the latest equipment. These are similar fears and insecurities to those experienced by researchers working in high-resourced environments. And yet, researchers in low-resourced environments often do not have access to the kinds of buffer available to their better-equipped colleagues, with severe consequences for their publication strategies. In interviews conducted with researchers in South Africa and Kenya in 2014, for instance, it was clear that insecurity around data production methods and access to technology has a strong impact on researchers' self-confidence and wish to have visibility, share data and publish work internationally (Bezuidenhout et al 2016, 2017; Bezuidenhout this issue; Rappert this issue).

Such findings are not unique nor should they be particularly surprising: scholars in Science and Technology Studies and anthropology have long stressed the role of technology as a marker for identity politics particularly in the African continent (e.g. Ferguson 2006). As starkly illustrated recently by work such as Damien Droney's in Ghana (Droney 2014), Julie Livingston in Botswana (Livingstone 2012), Joanna Crane in Uganda (Crane 2013) and Abena Dove Osseo-Asare across West and East Africa (Osseo-Asare 2014), popular culture associates being a scientist with owning spectacular equipment, and this perception filters down to researchers themselves. Equipment is the most visible and concrete marker of wealth in a lab, and it is often interpreted as a signal of the extent to which a research environment in a low-income country can aspire to produce research comparable in quality and significance to that produced by a high-resourced lab. Technology thus becomes a marker for inclusion and a symbol of being part of the Western world in some way – taking distance from the identity of "African scientist" which many researchers find cumbersome and problematic in their dealings with international publishing outlets, funders and institutions. This contributes to the already unequal championing of home-grown scientific approaches and techniques vis-a-vis methods, concepts and questions imported from the Global North, despite the existence of research areas that are less dependent on expensive machinery and more on elements commonly found across low-resourced environments, such as manpower, expertise and access to specific locations or natural resources (Kelly and Lezaun 2017).

These considerations, which of course apply more widely than African science and potentially include all research conducted in low-resourced environments, bring me to the conclusion that using references to specific technology as proxies for data quality has at least four problematic implications:

1. *It may act as an incentive to reduce diversity and creativity in research approaches, by encouraging standardization and the use of the same techniques and technologies regardless of the research context.*

This situation is troublesome when considering the sophisticated relationship between the goals and interests of researchers at different locations, the specific characteristics of each target system in biology, and the methods devised to study those systems – factors which vary widely and can be combined in myriads of ways, giving rise to countless different ways to conduct and validate research. It is also problematic when considering research environments that cannot avail themselves of the latest software or instrument, but which are nevertheless producing high-quality data of potential biological significance – sometimes because of the materials they have access to, sometimes because of their innovative conceptual or methodological approach, sometimes because they are targeting questions and phenomena of little interest to researchers based elsewhere.

2. *It leads to widespread mistrust and fear of openness, particularly when it comes to the sharing of research data.*

All too often, researchers working in low-resourced environments are afraid that lack of access to the latest technologies will affect the quality and reliability of their data, and will make their findings vulnerable to criticisms by better-resourced peers. Disparity in access to technologies also affect the speed and efficiency with which data being shared are analysed, giving researchers based in well-equipped labs the opportunity to analyse and publish on data produced in low-resourced conditions much faster than the original data producers (under the current evaluation regimes, which privilege publication of papers over data production, this is equivalent to being scooped). These fears can result in researchers being unwilling to share their data and/or to describe the specific circumstances and tools through which they were obtained.

3. *It reinforces systematic disadvantage among labs that do not have access to expensive resources.*

This may be the result of researchers' own reluctance to acquire international partners who could question their methods, and/or to disclose their set-ups (as in the previous point). It may also arise due to the insidious power that assumptions around what counts as a good research environment have within academic structures, evaluation panels and editorial boards of international journals. It is no secret that researchers located in highly reputable institutions have less trouble having their papers accepted for peer review at top-level journals such as *Nature* and *Science*. Similarly, many national policies explicitly ask researchers to emulate the working practices of what are typically regarded as scientific leaders at top Western institutions.

4. *It encourages misunderstandings and miscommunication between research data producers and users.*

People who do not articulate the differences between their environments, or feel compelled to minimise them in the name of implicit good standards for "best practice", are at risk of miscommunications and misunderstandings, leading to

breakdown in collaborations, problems in interpreting results and difficulties in replicating experiments.

Conclusion: Fostering critical engagement with data quality

In this paper, I pointed to data quality assessment as crucial to international research collaboration and advancements in the age of Open Data. At the same time, I warned that the push to Open Data, which involves an increasing emphasis on standard data formats and tools for data sharing, is affected by the extensive commercialisation of lab equipment and technologies for data dissemination. These elements risk to create a situation where data quality is assessed on the basis of the technologies being employed, rather than the fit between data, methods, materials available and research questions being asked. By contrast, research strategies are typically fine-tuned to the specific questions that researchers wish to pursue and to the phenomena that they wish to study, and such fine-tuning is conducive to research outputs that are credible, well-justified and innovative in their approach and significance. There is thus a wide variety of models for what may count as 'best practice', 'adequate data stewardship' and 'good research environments', whose relevance depends on the specific situations of inquiry in which researchers operate. A molecular biology lab with latest equipment based at Harvard or Cambridge needs not be the standard against which all research set-ups around the world are set, and should certainly not be implicitly taken to play that role. What type of experimental set-up fits which research project is a contextual matter, depending on lots of factors including the research questions and approach that is taken, the expertise of the researchers in question, the social dynamics within the group and its international collaborations, and the institutional support, infrastructures and materials available to researchers.

This does not mean that researchers working under very different conditions should not talk with each other and exchange tips for improving their environment and working habits. Quite the contrary: acknowledging diversity is an important step towards making such conversation more meaningful and fruitful, as long as this involves challenging the presumption (often unjustified, as the research above demonstrates) that researchers working within the same field actually mean the same when using similar terminologies, and should be constrained in the same ways regardless of the specificities of their working environment.

It is imperative that researchers, policy-makers and funders engaged in debates around data quality take these dimensions into account, particularly when thinking about implementing Open Data practices in low-resourced research environments. The sharing of data typically relies on the ability to use sophisticated data formats and digital data infrastructures, and thus to keep up with the fast pace of technological change associated to such data sharing tools and standards. This becomes problematic given the importance to store and disseminate multiple data formats, non-digital sources (as in 'old-fashioned' paper archives) and data produced by different versions of same software, which

helps to embrace the variety of work carried out in the sciences globally (including both low-resourced laboratories and the so-called 'long tail of science'). Also, it is crucial to enable researchers to develop their projects whether or not they avail themselves of the latest technology, and hence to consider and assess when such technology is needed, and for which purposes.

Thus, Open Data initiatives should be aware of the implications of endorsing specific types of technologies (whether hardware, software or specific laboratory instruments) as markers of research quality. Debates around Open Data should include explicit and field-specific reflections around the relation between data, research instruments and methods, where researchers clearly articulate their assumptions on what constitutes 'best practice', who sets a model for such work, and whether such assumptions are realistic and warranted in light of their own research experiences. This type of articulation is a precious tool for research advancement, since it would encourage confrontation and dialogue at the international level around what quality standards are desirable for data, and with respect to which uses and research goals. These reflexive exercises could be of great value in an ever-globalised and diverse scientific landscape, where the specificity of locations, methods and interests characterising each research community needs to be documented as essential meta-data. In the absence of such critical engagement, Open Data guidelines risk to dismiss or obscure researchers' situated knowledge and practices (as well as the diversity of fundamental research carried out around the world, Rochmyaningsih 2016), and instead appeal to politically charged and potentially damaging assumptions about what constitutes 'best practice'.

Acknowledgments

This research was funded by the European Research Council grant award 335925 ("The Epistemology of Data Science"), the Leverhulme Trust Grant number RPG-2013-153 ("Beyond the Digital Divide"), and the Australian Research Council, Discovery Project DP160102989 ("Organisms and Us"). The author gratefully acknowledges input from the Data Studies group at Exeter, and particularly Louise Bezuidenhout, Brian Rappert and Ann Kelly as co-team members of the Leverhulme Project; participants to the "Pacing Science" workshop that took place at the University of Exeter in May 2016, where this paper was first presented, and particularly Linsey McGoey and Simon Hodson; members of the Global Young Academy "Open Science" and "Global Access to Research Software" working groups, especially Abdullah Shams Bin Tariq and Martin Dominik; members of the Knowledge/Value research network, particularly Kaushik Sunder Rajan and Kris Peterson; and the dozens of researchers who spared time and effort to discuss their methods, outputs and working conditions with me and my colleagues.

Ethics and Consent

The interviews and ethnographies used as empirical source in this paper have received approval by the Social Science Ethics Committee of the University of Exeter, in relation to projects "Beyond the Digital Divide" and "The Epistemology of Data-Intensive Science". All participants signed a consent form. Their contributions are anonymised and their confidentiality is fully respected.

References

Allison, D B et al. 2016 A tragedy of errors. *Nature*, *530*, 27–30.

Ankeny, RA and Leonelli, S 2016 Repertoires: A Post-Kuhnian perspective on scientific change and collaborative research. *Studies in the History and the Philosophy of Science: Part A* 60: 18-28. http://dx.doi.org/10.1016/j.shpsa.2016.08.003

Ankeny, RA and Leonelli, S 2015 Valuing data in postgenomic biology: How data donation and curation practices challenge the scientific publication system. In Richardson, S and Stevens, H (eds) *Postgenomics*. Duke University Press, pp.126-149.

Barnes, B and Dupré, J 2008 *Genomes and What to Make of Them*. Chicago, IL: The University of Chicago Press.

Bezuidenhout, L 2017 (this issue) *Data Science*.

Bezuidenhout, L, Rappert, B, Leonelli, S, Kelly, A 2016 Datasets for beyond the digital divide: Sharing research data across developing and developed countries. *figshare*. https://dx.doi.org/10.6084/m9.figshare.3203809.v1

Bezuidenhout, L, Leonelli, S, Kelly, A and Rappert, B 2017 Beyond the digital divide: Towards a situated approach to Open Data. *Science and Public Policy*. http://dx.doi.org/10.1093/scipol/scw036

Blake, JA, Christie, KR, Dolan, ME, Drabkin, HJ, Hill, DP, Sitnikov, LND et al. 2015 Gene Ontology Consortium: Going forward, *Nucleic Acids Research* 43 (D1): D1049–56. doi:10.1093/nar/gku1179

Borgman, Christine L 2012 The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology,* 63(6): 1059–1078.

Cai, L and Yangyong Z 2015 The challenges of data quality and data quality assessment in the Big Data era. *Data Science Journal* 14: 2. doi:10.5334/dsj-2015-002.

Calude, CS and Longo, G 2016 The deluge of spurious correlations in Big Data, *Foundations of Science*, 1–18. doi:10.1007/s10699-016-9489-4.

Crane, J T 2013 *Scrambling for Africa: AIDS, Expertise and the Rise of American Global Health Science*. Ithaca, N.Y.: Cornell University Press.

Droney, D 2014 Ironies of laboratory work during Ghana's second age of optimism, *Cultural Anthropology*, *29*(2), 363–384. Doi: 10.14506/ca29.2.10

Dupré, J 2012 *Processes of Life*. Oxford, UK: Oxford University Press.

Ferguson, J 2006 *Global Shadows: Africa in the Neoliberal World Order*. Durham, N.C.: Duke University Press.

Floridi, L and Illari, P 2014 *The Philosophy of Information Quality*. Synthese Library 358. Cham, CH: Springer.

Kambatla, K, Kollias, G, Kumar, V and Grama, A 2014. Trends in big data analytics, J*ournal of Parallel and Distributed Computing* 74(7): 2561-2573. http://dx.doi.org/10.1016/j.jpdc.2014.01.003.

Kelly, A and Lezaun, J in press, 2017 The wild indoors: Rooms spaces of scientific inquiry. *Cultural Anthropology*.

Krohs, U 2012 Convenience experimentation, *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (2012) 52–57. http://dx.doi.org/10.1016/j.shpsc.2011.10.005

Lagoze, C 2014 Big Data, data integrity, and the fracturing of the control zone. *Big Data and Society*, 1(2). https://doi.org/10.1177/2053951714558281

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

Leonelli, S 2016 Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production, *Philosophical Transactions of the Royal Society: Part A*. 374: 20160122. http://dx.doi.org/10.1098/rsta.2016.0122

Leonelli, S 2014 What difference does quantity make? On the epistemology of Big Data in biology, *Big Data and Society*, 1 (1). https://doi.org/10.1177/2053951714534395

Leonelli, S, and Ankeny, RA 2015 Repertoires: How to Transform a Project into a Research Community. *BioScience* 65(7): 701-708. https://doi.org/10.1093/biosci/biv061

Leonelli, S, Diehl, AD, Christie, KR, Harris, MA and Lomax, J 2011 How the Gene Ontology evolves, *BMC Bioinformatics*, 12:325 DOI: 10.1186/1471-2105-12-325

Levin, N, Leonelli, S, Weckowska, D, Castle, D, and Dupré, J 2016 How Do Scientists Understand Openness? Exploring the Relationship between Open Science Policies and Research Practice. *Bulletin for Science and Technology Studies* 36(2): 128-141. http://dx.doi.org/10.1177/0270467616668760

Livingston, J 2012 *Improvising Medicine: An African Oncology Ward in an Emerging Cancer Epidemic*. Durham, N.C.: Duke University Press.

Mayer-Schönberger, V and Cukier, K 2013 *Big Data: A Revolution that Will Transform How We Live, Work and Think.* London, UK: John Murray.

McDowall, M D, Harris, MA, Lock, A, Rutherford, K, Staines, DM, Bähler, J, Kersey, PJ, Oliver, SG, and Wood, V 2015 PomBase 2015: Updates to the Fission Yeast Database, *Nucleic Acids Research* 43 (Database issue): D656-61. doi:10.1093/nar/gku1040.

Morey, RD, Chambers, DC, Peter J. Etchells, Christine R. Harris, Rink Hoekstra, Daniël Lakens, Stephan Lewandowsky, et al. 2016 The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review, *Royal Society Open Science* 3 (1): 150547. doi:10.1098/rsos.150547.

Müller-Wille, SW and Rheinberger, HJ 2017 *The Gene: From Genetics to Postgenomics.* Chicago, IL: The University of Chicago Press.

Müller-Wille, SW and Rheinberger, HJ 2012 *A Cultural History of Heredity*. Chicago, IL: The University of Chicago Press.

Ossorio, P 2011 Bodies of Data: Genomic Data and Bioscience Data Sharing, *Social Research* 78 (3): 907–32.

Osseo-Asare, AD 2014 *Bitter Roots: The Search for Healing Plants in Africa*. Chicago University Press.

Pulverer, B 2015 Reproducibility Blues. *EMBO Reports* 34(22): 2721-2724. 10.15252/embj.201570090

Rappert, B 2017 (this issue) *Data Science*.

Rajan, KS 2006 *Biocapital*: *The Constitution of Postgenomic Life.* Duke University Press.

Research and Markets 2016 *The World Market for Microarrays.* Report. http://researchandmarkets.com/report/7fhv5g/the_world_market

Rochmyaningsih, D 2016 The developing world needs basic research too, *Nature* 534 (7605): 7–7. doi:10.1038/534007a.

Rogers, S and Cambrosio, A 2007 Making a new technology work: The standardization and regulation of microarrays, *Journal of Biology* 80: 165–78.

Testa, G and Nowotny, H 2011 *Naked Genes*. MIT Press.

Thayer, A 2016 Top Instrument firms in 2015, *C&EN* 94(17): 32-35. http://cen.acs.org/articles/94/i17/Top-instrument-firms-2015.html

---

[i] The empirical research for this paper was carried out by me within research sites in Wales, Britain, the United States, Germany and Belgium, and by Louise Bezuidenhout within sites in South Africa and Kenya (for more details on the latter research and related methods, see the paper by Bezuidenhout in this special issue). Given the sensitive nature of the interview materials, the raw data underpinning this paper cannot be openly disseminated; however, a digested and anonymized version of the data is provided on Figshare (Bezuidenhout et al 2016).

[ii] It has also been argued that data quality does not matter within big data collections, because existing data can be triangulated with other datasets documenting the same phenomenon, and datasets that corroborate each other can justifiably be viewed as more reliable (Mayer-Schöneberg and Cukier 2013). Against this view, myself and others pointed out that triangulation only works when there are enough datasets that document the same phenomenon from different angles, which is not always the case in scientific research (see e.g. Leonelli 2014, Calude and Longo 2016).