Learning from Data Journeys

Sabina Leonelli

Abstract The introduction discusses the idea of data journeys and its characteristics as an investigative tool and theoretical framework for this volume and broader scholarship on data. Building on a relational and historicized understanding of *data as lineages*, it reflects on the methodological and conceptual challenges involved in mapping, analyzing and comparing the production, movement and use of data within and across research fields and approaches, and the strategies developed to cope with such difficulties. The introduction then provides an overview of significant variation among data practices in different research areas that emerge from the analyses of data journeys garnered in this volume. In closing, it discusses the significance of this approach towards addressing the challenges raised by data-centric science and the emergence of big and open data.

1 Introduction: Data Movement and Epistemic Diversity

Digital access to data and the development of automated tools for data mining are widely seen to have revolutionized research methods and ways of doing research. The idea that knowledge can be produced primarily by sifting through existing data, rather than by formulating and testing hypotheses, is far from novel; and yet, developments in information technology and in the financing, institutionalisation and marketization of data are making "data-intensive" approaches more prominent than ever before in the history of science. This is perhaps most blatant in the emphasis placed by both the public and private sectors on the production and exploitation of "big" and "open" data – in other words, on the creation, dissemination and aggregation of vast datasets to facilitate their re-purposing for as wide a range of goals as possible.[1]

[1] As exemplified by the Open Science and Innovation policy of the European Commission (European Commission 2016).

S. Leonelli (✉)
Department of Sociology, Philosophy and Anthropology & Exeter Centre for the Study of the Life Sciences (Egenis), University of Exeter, Exeter, UK

Alan Turing Institute, London, UK
e-mail: s.leonelli@exeter.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_1

The promise of big and open data is tied to two key factors. One is their *mobility*: the value of data as prospective evidence increases the more they travel across sites, since this makes it possible for people with diverse expertise, interests and skills to probe the data and consider whether they yield useful insight into their ongoing inquiries.[2] The other is their *interoperability*, that is the extent to which they can be linked to other types of data coming from a variety of diverse sources.[3] It is through linkage techniques and tools that data become part of big data aggregates, which in turn function as empirical platforms to explore novel correlations, power machine learning algorithms and ask ambitious and innovative questions.

This volume interrogates the conditions for data movement, and the ways in which data mobility and interoperability can be achieved, from the viewpoint of the history, philosophy and social studies of science. What is already clear from the growing scholarship on data is that this requires enormous resources, apposite technologies and methods, and high levels of human ingenuity - which is why in the world of research as in many other parts of society, online databases, data visualization tools and data analytics have become indispensable to any form of research and innovation.[4] This insight runs counter the hyped public discourse around the supposedly intrinsic power of big data and the related expectation that, given a lot of data, useful and reliable discoveries would follow. And yet, even recognising that mobilizing data requires resources is not enough to understand how they can be effectively used as sources of evidence. Stocking up on skills and tools from data science, information technology and computer engineering does not suffice for knowledge production. The critical issue is how to merge such expertise and solutions with existing domain-specific knowledge embedded in evolving social contexts, thus developing methods that carefully and creatively tailor data-intensive approaches to the study of specific targets and the achievement of given goals. In other words, transforming data into knowledge requires more than some generalist algorithms, clustering methods, robust infrastructure and/or clever apps: it is a matter of adapting (and sometimes creating) mathematical and computational tools to match the ever-changing characteristics of the research targets, methods and communities in question – including their political and economic context.

To highlight this, the volume brings together in-depth case studies that document the motivations and characteristics of the existing variety of data practices across

[2] Data mobility has been associated to the rise of a "fourth revolution" in knowledge production that is affecting all aspects of society (Hey et al. 2009; Kitchin 2014; Wouters et al. 2013; Floridi 2011). I argued that extensive data mobility is a defining characteristic of data-centric science, which also captures the historical novelty of this approach to data (Leonelli 2016).

[3] This is widely recognized in data science itself, where interoperability is viewed as one of the four crucial challenges to so-called "FAIR" data (that is, data which are "findable, accessible, interoperable and reusable"; Wilkinson et al. 2016). See also extensive ethnographic research on interoperability conditions by Christine Borgman and collaborators (e.g. Edwards et al. 2011; Borgman 2015) and the Exeter data studies group (e.g. Leonelli 2012; Tempini and Leonelli 2018), among others.

[4] See for example the inaugural issue of the Harvard Data Science Review (Meng 2019), in which these factors are all highlighted as integral components of data science.

research fields, locations, projects, objectives and lines of inquiry. This provides readers with insight into the salient circumstances affecting data interpretation, be they scientific, technological, political and/or social – and thus with concrete grounding to consider *how such variety originates, how it affects whether and how data are moved and re-used, and with which implications for the knowledge being generated – and its social roles*.

Data production and use within different areas of research have long been defined by highly distinctive histories, methods, objects, materials, aims and technologies. Such diversity is a key challenge to any attempt to articulate the general characteristics and implications of data-intensive science, and indeed there is arguably no single characterisation that can fit all the different ways of working subsumed under that umbrella. Leading research organisations, science academies and science policy bodies have repeatedly argued that when it comes to data practices, "one size does not fit all" and it is thus damaging to apply the same guidelines and standards for data management across different fields, research situations and long-standing traditions.[5] In a similar vein, historians have documented various forms of big data production and interpretation across space, time and disciplinary boundaries[6]; and researchers in the social and information sciences have documented the diverse ecosystems underpinning research in biology, biomedicine, physics, astronomy and the social, environmental and climate sciences – and pointed to differences in data types and standards, preferred instruments, norms and interests as having an enormous impact on the effectiveness of strategies to analyse large datasets brought together from different sources.[7]

How does such diversity affect the conditions under which data are processed and disseminated for re-use across different research environments? This is the question at the heart of this volume. Answering this question implies, first of all, understanding how data practices (ranging from the design of data collection to data processing and interpretation) adapt to specific situations, while also arching back to long-standing methodological traditions and norms. It also involves understanding how data actually move from one setting to another, what it takes for that movement to occur and what conceptual, material and social constraints it is subject to. Such understanding is particularly relevant in our age of distributed global networks, multidisciplinary collaboration and Open Science, where the pooling and linking of data coming from different fields, topics and sources constitutes at once a tantalising opportunity and a significant challenge. Without the ability to track how data change themselves and their environment as they move across contexts, it is impossible to strategize, innovate or even just document data practices and their

[5] See for instance the OECD (2007), Boulton et al (2012), the Global Young Academy (2016), the Open Science Policy Platform (2018) and the European Commission (2017). The whole working agenda of the Research Data Alliance is also based around the recognition of field-specific data requirements. I have discussed the epistemic foundations for this view in Leonelli (2016).

[6] For instance see Blair (2010), Aronova et al. (2018), Daston (2017).

[7] Among prominent contributors: Geoff Bowker (1994 and subsequent works), Paul Edwards (2010), Rob Kitchin (2014), Borgman (2015).

effects – also making it hard to assign responsibility for mistakes, misunderstandings or wilful deceptions in the use of data as evidence for decision-making.

Tracking data movements and explaining their direction and implications cannot be done solely through quantitative methods. Bibliographic analyses are of limited use since the vast majority of researchers, despite grounding their research on the consultation of databases, are not in the habit of documenting their searches or cite their data sources with precision when writing up results. The re-use of data is most commonly acknowledged in the form of a citation to a journal article providing a specific interpretation of the data. Where data are sourced from a repository rather than a published paper, citation is less reliable (also because some repositories do not provide stable identifiers for their datasets, so data users would cite the whole repository rather than the specific entry of interest); and the pivotal role played by data infrastructures in facilitating the re-use of data remains largely hidden.[8] Moreover, the number of infrastructures, technologies and standardisation tools developed to process and mobilise data is growing exponentially, generating vast and interdependent networks of resources which are extremely hard to map and describe even for the practitioners involved. One of the reasons for this growth is the insistence by researchers working within different traditions to tailor their data practices and related tools as closely as possible to their existing methods and commitments. This requirement makes sense given that such methods and commitments have been adapted over centuries to the study of the specific characteristics of phenomena of interest, and yet makes it difficult for researchers to agree on common standards and norms. This reluctance, coupled with a project-driven, short-term funding system, encourages an uncontrollable and unsustainable proliferation of resources for the management and analysis of data, with hundreds of databases emerging every year in relation to the same research field. As is often the case when scores of information resources haphazardly multiply and intersect, this proliferation results in obfuscation: each tool for data mobilisation becomes a black-box whose effects on the wider landscape are impossible to quantify without a thorough qualitative assessment.[9] The expanding network of variously interlocked data resources and infrastructures is thus not only hard to trace, but opaque in its impact on knowledge generation.

The investigative approach used in this volume builds on extensive research on the history of different fields, the qualitative study of the practices and ethos characterising the research communities in question, and consideration of how such history affects: (1) the norms, strategies and behaviours utilized when collecting, sharing and processing data, including measuring frameworks and specific instruments and skills; and thus (2) the outputs of research, which may include knowledge claims but also technologies, methods and forms of intervention. Through the in-depth investigation of case studies, we follow different stages of data movements,

[8] This has made it very difficult to quantify the impact of data infrastructure on research, and thus their value (Bastow and Leonelli 2010; Pasquetto et al. 2017).

[9] For detailed studies on this phenomenon, see Mongilli and Pellegrino (2014), Pasquale (2015), Egyedi and Mehos (2015), Ebeling (2016), Leonelli (2018a).

ranging from the planning that precedes data production to various ways in which data are mobilised and re-purposed, often with the goal of providing "actionable" knowledge. The volume as a whole constitutes a (undoubtedly partial, yet rich) sample of the variety of data practices to be found in different portions of the research world. At the same time, the volume exemplifies a coherent overarching approach to the investigation of data movements and their implications, which is ideally suited to analysing the diverse conditions under which data are handled, understanding the reasons underpinning such diversity, and identifying nodes of difference and similarity in ways that can help develop best practice. This approach, which we call the study of "data journeys", is what this introductory chapter aims to systematically review and articulate.

To this aim, this chapter is structured as follows. I first discuss the very notion of data and provide a conceptualisation of data epistemology that proves particularly suitable to the emphasis on data mobility and interoperability: the historicized and relational view of *data as lineages* (Sect. 1). I then discuss the idea of data journey both as a way of theorising data movement and as a methodological tool to investigate it (Sect. 2). I emphasise how data movements often transcend institutional boundaries and evade – or even reshape -- traditional conceptions of division of labour in science, thus making categories such as 'disciplines' and 'research fields' descriptively and normatively inadequate. The fluid nature of data journeys makes them challenging to identify and reconstruct, and yet it is the very opportunity to articulate and explicitly tackle those challenges that makes data journeys into useful units of analysis to map and compare the situations and sets of practices through which data are mobilised and used (Sect. 3). As a demonstration, I reflect on some significant differences and similarities among data practices that emerge from the analyses of data journeys garnered in this volume (Sect. 4). In closing, I discuss the significance of this approach towards addressing the scientific, political, economic and social challenges raised by data-centric science and the emergence of big data. This body of work does not sit easily with the current political and economic push towards universal adoption of big and open data as motors of research and innovation (Srnicek 2017, Mirowski 2018). Recognizing the diversity of data journeys and related practices explains the difficulties involved in governing and standardizing big and open data, and highlights the considerable resources and the breadth of expertise involved in re-using data in ways that are sustainable, reliable and trustworthy.

2 Mutability and Portability: Data as Lineages

When attempting to define what data are and how they contribute to the production of knowledge, reference to the Latin etymology of the term 'datum' - meaning "that which is given" - is unavoidable. This volume takes one aspect of this etymology very seriously: the reference to the *public life of data* as objects that can be physically moved and passed around (whether through digital or analogue means), so as

to be subject to scrutiny by people other than those involved in their creation. Data are mobile entities, and their mobility defines their epistemic role. Hence, for any object to be identified and recognised as datum, it needs to be portable.

This is not a new position. An early proponent was Bruno Latour in his seminal discussion of how data produced during fieldwork are subsequently circulated (Latour 1999). Latour, however, added that while data are defined by their mobility, their epistemic power derives from their immutability - their capacity to stay the same and thus to be taken as a faithful and stable document of the specific moment, place and environment in which they were created. In this interpretation, data are static products of one-off interactions between investigators and/or the parts of the world under investigation: while phenomena change over time, the data that document them are fixed.

This volume was born of a different premise: that this impression of fixity, often associated to the idea of data as "given", is highly misleading. In virtually all of the cases discussed in this volume, data are everything but stable objects ready for use. What makes data so powerful as sources of evidence is rather their *mutability*: the multiple ways in which they are transformed and modified to fit different uses as they travel across space, time and social situations. In order to serve their evidential function, data need to be adapted to the various forms of storage, dissemination and re-use over time and space to which they are subjected. Hence the mobility of data depends on their capacity to adapt to different landscapes and enter unforeseen spaces. As they travel around, data undergo frequent modification to fit their new environments. They acquire or shed components, merge with other data, shift shape and labels, change vehicles and companions, and such transformations prove essential to their usability by different audiences and purposes. As Mary Morgan (2010) noted in relation to the travels of facts, data are therefore best viewed as *mutable mobiles*. The more they travel, the more they shift shape to suit their new circumstances, and as a result prove tractable and effective in serving new goals.

This conceptualisation of data immediately poses a series of conceptual and methodological problems. Do data retain some integrity while they travel? How do we make sense of data as objects that remain identifiable while changing characteristics, shape and format throughout their journeys? And when do data cease to be data and become something else? The chapters of this volume answer these questions in the form of stories of data birth, regeneration, loss and even death. These stories highlight the extent to which what is used as data by a given group at a given moment in time and space may not retain that function at a later time, either because the group shifts attention to other objects as sources of evidence or because the journey to new research situations fails.

One way to frame these stories and their significance for data epistemology is to adopt a *relational view of data*, within which the power to represent and thus document specific aspects of the world is not intrinsic to data in and of themselves, but rather derives from situated ways of in which data are handled (such as specific forms of modelling and interpretation).[10] This is not to say that the physical features

[10] I discuss the relational view of data in detail in Leonelli (2016, 2018a).

of data objects – what colour and consistence they are, what marks they bear, and perhaps most crucially, whether or not they resemble (and in which respects) given aspects of the world – do not matter. Quite the opposite: the material properties of data as objects play a pivotal role in enabling and constraining specific practices of assemblage, dissemination and interpretation. And yet, they are not the only constraint on modelling and theorising. Other significant factors include the technologies, materials, social settings and institutions involved in facilitating or impeding data travel. For example, the photograph of a child has physical properties that make it a potentially useful source of evidence in a study of human physical development, but this potential can only be realised under a series of conditions that include: the availability of comparable data (say pictures of other children, pictures of the same child at different times, or other types of data on the child such as her height and family history); the extent to which the resolution and format of the photograph fit the requirement imposed by the computational tools used in the analysis; and the opportunity to access relevant metadata (such as the age and location of the child, which however constitute sensitive data whose circulation and use are strictly regulated). What data can be evidence for - what representational value is ascribed to them - thus depends on their concrete characteristics *at the time of analysis* as well as the specific situation in which data are being examined.

The relational view of data makes them into historical entities which – much like organic beings – evolve and change as their life unfolds and merges with elements of their environment. Building on this biological metaphor, I propose to conceptualize data as *lineages*: not static objects whose significance and evidential value are fixed, but objects that need to be transformed in order to travel and be re-used for new goals. The metaphor may appear to break down when observing that the plasticity of organisms and their ability to adapt to new environment are essential conditions for their survival, while data seem perfectly able to live a long life without requiring any modification. Typical examples are the contents of archives, musea, repositories and other establishments whose goal is often understood to consist of the long-term preservation of artefacts in their original state. In response to this objection, my contention is that what these establishments preserve are not data, but rather objects which may or may not be used as data (or data sources); and that as soon as the effort is made to use such objects as data or acquire data from them (for example, through measurement), they are at least minimally modified to fit the ever-evolving physical environments and research cultures within which they are valued and interpreted.[11] Using an archaeological artefact or an organic specimen as datum and/or data source, for instance, may involve touching it and moving it around – operations that are likely to affect the object itself, particularly if it is fragile and/or

[11] A very significant difference between data and organisms may consist of the locus of agency, with data depending on the agency of humans for their "evolution" as components of inquiry, while organisms arguably possess some degree of self-organisation. This introduction is no place for a lengthy exploration of these ideas, which are the subject of a manuscript in preparation by Leonelli and John Dupré.

very old, and be conducted differently depending on what instruments researchers are using to document the characteristics of the object.[12]

Thus again, the use of objects as data requires portability and mobility, which in turn beget mutability - for instance when exposing data to new technologies, bringing them to new user communities, and articulating how they may fit new strands of inferential reasoning. The archaeological artefacts discussed by Alison Wylie are a perfect case in point, with her chapter illustrating how the ways in which these materials are manipulated – and traces are extracted from them – changes in parallel to shifting conceptual, institutional and technological contexts of analysis. Both her case and the case of art authentication discussed by Coopmans and Rappert powerfully show how the very value of artefacts as data sources depends on mobilisation and transformation, since if complete consensus was reached on what exactly these objects represent, there would be no incentive to continue to use them as part of a line of inquiry.

By the same token, several chapters in the volume demonstrate the enormous efforts and resources involved in keeping data objects and their evidential value stable over time – from the development and updating of standards and classificatory categories, as discussed by Edmund Ramsden in the case of data about housing and Jean-Paul Gaudillière and Camille Gasnier in relation to health data, to the development of consensus around the interpretive commitments used in data infrastructures (e.g. the biomedical "knowledgebases" analysed by Alberto Cambrosio and colleagues) and the establishment of benchmarks and practices through which data uses can be documented and assessed, as described by Wendy Parker for weather data and Götz Hoeppe for astronomical observations. It is no coincidence that what Cambrosio and colleagues document is the gradual disappearance of data from clinical spaces in favour of established, situated interpretations of those data. Within knowledgebases, the question of what makes data such in relation to any one clinical situation is eschewed in favour of a more practical and actionable reference to agreed interpretative claims.

While other conceptualisations of data may well fit the study of data journeys,[13] the relational view of data as lineages does in my view illustrate the significance of focusing on data movements to understand the role and status of data within research. This approach shifts analysts' attention towards understanding what makes data more or less stable and usable, the epistemic – but also affective, institutional, financial, social - value imputed to the objects used as data across different situations of inquiry, and the extent to which such objects retain or lose integrity and material properties. It thus challenges facile understandings of data as the "raw" materials of science, which have long been critiqued within philosophy and the social sciences,[14] and yet remain attractive to those who like to understand the

[12] See for example Wylie (2002) and Shavit and Griesemer (2011).

[13] Another useful conceptualization, which also emphasizes the significance of studying data as mobile and mutable objects but places emphasis on the socio-material rather than the conceptual conditions of travel, is that proposed by Bates et al. (2016).

[14] As epitomized by the effectively titled book edited by Lisa Gitelman (2013), *Raw Data is an Oxymoron*, and recalled by Helen Longino, a prominent participant in these debates, in the afterword of this volume.

research process as a straightforward accumulation of facts. All the contributions to this volume exemplify how using data as evidence is everything but straightforward, and sophisticated methods, resources and skills are required to guarantee the reliability of the empirical grounds on which knowledge is built.

3 Data Journeys as Units of Analysis

Data journeys can be broadly defined as designating the *movement of data from their production site to many other sites in which they are processed, mobilised and re-purposed.* "Sites" in this definition do not need to refer to geographical locations, though this is often the case: they also encompass temporal locations and diverse viewpoints (whether motivated by different theoretical commitments, expertise and know-how, or by political, social and ethical views).

As a conceptualisation of the research process, the idea of data journeys is a direct counterpoint to the distinction between "hypothesis-driven" and "data-driven" modes of research. Data journeys provide a framework within which to identify and investigate the various ways in which theoretical expectations shape the travel of data and the various vehicles and resources used to support that travel, regardless of whether the data were originally generated to test a given hypothesis. Indeed, focusing on data journeys facilitates the identification and exploration of data movements regardless of whether they are part of the same line of inquiry or methodological approach. Data produced to test a hypothesis are no less likely to travel than data produced for explorative purposes: in both cases, the data are tied to a specific frame of analysis (whether this is conceptual, as in the case of a given hypothesis, or methodological, as in the case of the tools used to collect and/or generate data), and work is required to move them away and beyond that frame. The chapter by Teira and Tempini discusses how data produced by a randomised clinical trial – the posterchild for hypothesis-driven research – do not typically travel beyond the trial itself unless legal protection of patient confidentiality and the commercial sensitivity of the data is in place, as well as institutions and infrastructures to curate the data appropriately (see also Tempini and Leonelli 2018). The difficulties involved in pharmaceutical data journeys become evident when attempting to merge such data with electronic health records gathered for goals different than that of testing. Focusing instead on data whose very history exemplifies the practice of data collection without a predetermined target, James Griesemer demonstrates how the circulation and appropriate mining of the outputs of sequencing experiments also requires the adoption of a complex set of strategies and resources.[15]

Indeed, the metaphor of the "journey" is powerful because, just like many human journeys, data journeys are enabled by infrastructures and social agency to various

[15]The very history of the development of institutional and technological means for sharing sequencing data within and beyond biology illustrates this well (see for example Stevens 2013, Hilgartner 2017 and Maxson et al. 2018).

degrees and are not always, or even frequently, smooth.[16] A useful way to think through the significance of adopting this metaphor is to consider what it can mean for journeys to be successful. Sometimes journeys are perceived as successful when they consist of an item or person following a given itinerary towards a pre-selected point of arrival, by means of existing vehicles and infrastructures. In this interpretation, successful journeys will require meticulous planning and/or dependable and easily accessible infrastructures, which can secure the pathways through which data can be displaced (much in the same way as humans managing a business trip without complications by travelling a well-serviced highway in a dependable car). Well-established and meticulously curated databases, such the biological ones discussed by William Bechtel in his chapter, can sometimes serve as such predictable, controlled travelling tools.

In other cases, the success of a journey will not depend on adherence to an itinerary or even a pre-determined destination, but rather on: the effects of the journey on its protagonists and/or their surroundings; the ability of a given vehicle to mobilise data in the first place; the extent to which data are welcomed and used in new environments; and/or the degree to which the purpose and destination of the journey changes *en route*. This is an interpretation of the idea of journey that relates less to physical displacement and more to intellectual development and learning, whereby one travels to explore, discover and "find meaning". Rachel Ankeny's discussion of the construction of medical case reports is a good example of the hopes and uncertainties built into developing vehicles for data, in a situation where the future uses and potential itineraries of such reports (and thus what counts as data within them) are largely unpredictable. The whole point of this form of data dissemination is to encourage as wide a range of future travel and interpretations as possible.

No matter what the success of a journey is taken to imply, its achievement is prone to the unavoidable serendipity involved in any type of displacement as well as the heightened risks typically associated with travel. Using data journeys as a unit of analysis for data practices and their outcomes helps to identify and evaluate such risks, including questions relating to error in the data (for instance when data are copied inaccurately), misappropriation, misinterpretation and loss – and the relation between such risks and the physical and social characteristics of data objects and their travelling vehicles. Gregor Halfmann's chapter on the transformation of samples into data stresses the precarious transitions involved in datafying the environment, but also the epistemic significance of the material links between the practices of data collection and further data dissemination and use. Once those material links weaken, for instance in cases where digital data have long been stored, formatted, shared and manipulated through various types of databases and related software, it becomes imperative to establish clear benchmarks for what data are reliable in relation to specific uses – and yet, as discussed both by Parker in relation to climate

[16] See also McNally et al. (2012), Lagoze (2014), Bates et al. (2016), among others.

science and Tempini in relation to public health, such benchmarking proves increasingly challenging to design as data journeys grow in length and complexity.[17]

More generally, using data journeys as a theoretical framework helps to consider and examine the relationship between different types of data structures (their physical characteristics as mutable objects) and data functions (their prospective use as evidence). What types of data - and forms of data aggregation - best afford what interventions and interpretations? And to which extent the physical characteristics of data constrain possible goals and uses? Many chapters in this volume focus on numerical data formats and their ability to aggregate and lend themselves to computational and statistical techniques, which in turn facilitates their travel and their reinterpretation for many purposes. Other chapters stress how images and samples lend themselves to different types of manipulations, with their rich material properties making them prone to a large variety of interpretation and also, possibly, to a broad evidential scope. While it has long been recognised that quantification has an important role to play in inferential reasoning, attention to data journeys rather than specific instances of data highlights the epistemic role played by data traditionally regarded as "qualitative".

Similar considerations apply to characteristics often associated to "big data" (Kitchin and McArdle 2016). Take, for instance, the idea of *volume* and the related notion of scale. Griesemer's and Mary Morgan's chapters both emphasise the importance of different kinds of data collectives and groups – such as datasets – to the travels of individual data points (or datums, in Morgan's provocative terms). As they point out, the mining of big data often involves: the merging of datasets of differing scales and sizes, whose components were collected through diverse frameworks; and choices about how such data collectives should be linked or otherwise compared are a fundamental component of data journeys. Another key property associated to big data is *velocity*, and again the study of data journeys enables analysts to interrogate this not just in relation to data production, but to the full arch of data mobilisation and re-purposing. What is the role of speed in data journeys? What impact does higher or lower speed of mobilisation have on the reliability of datasets, the amount of uncertainty and trustworthiness assigned to them, and the extent to which they can be reproducible? While the speed at which data travel may not matter much to their prospective re-use, the speed at which data vehicles, infrastructures and algorithms are developed to facilitate such fast travel matters a great deal. Lack of investment and strategy around data travels implicitly supports a naïve and unrealistic view of data as "speaking for themselves", which could compromise the extent to which data that have been mobilised can reliably interpreted as evidence. A case in point is Koray Karaca's data construction at CERN, where what constitutes a reliable and travel-worthy dataset from any one experiment (collision event) is decided through the automated implementation of models in a fraction of

[17] For lengthier discussions of quality assessment in distributed data systems, see Floridi and Illari (2014), Cai and Zhu (2015) and Leonelli (2017).

a second, but the computational, theoretical and methodological resources that make such a quick decision process possible require immense foresight, adequate theoretical models, a highly sophisticated experimental apparatus and constant calibration work. Similarly, Hoeppe illustrates cases of fast data travel in astronomy while also stressing the importance of explicit reflection on assumptions, norms and standards used during such journeys towards evaluating existing data interpretation.

4 The Significance of Articulating Data Challenges

Regardless of what perspective one has on the nature and roles of data, tracking data journeys is a fruitful methodological tool to investigate what happens to *data* themselves, rather than instruments, methods, claims, epistemic communities, repertoires, epistemic regimes. Attempts to follow and reconstruct data journeys are experiments in identifying components of research that are of direct relevance to data, rather than, as more usual within theory-centric approaches to knowledge development, considering data in order to understand theories and models. In this sense, we take inspiration from the *infrastructural inversion* articulated by Geoffrey Bowker and Susan Leigh Star, with its encouragement to "recognize the depths of interdependence of technical networks and standards, on the one hand, and the real work of politics and knowledge production on the other" (Bowker and Star 1999).[18] What data journeys do is place the spotlight firmly on to data themselves and the implications that infrastructures – among many other forces, expectations and material settings - have on their interpretation.

I already stressed how this approach enables analysts to step beyond a rigid conceptualisation of "disciplinary" knowledge spaces, communities and tools. Data are fascinating research components partly by virtue of their ability to transcend boundaries. The explosion of data journey sites reflects the disruptive power of data with respect to institutional and disciplinary boundaries. Data are collected, circulated and re-used within and beyond the scientific world, across different publics and for widely diverse purposes – think only of crowdsourcing and citizen science as an example of data crossing over various types of research and decision-making in both the private and the public sector. Most significantly, data travels often play an important role in challenging and re-shaping institutional, disciplinary and social boundaries, thus acting as the ultimate "boundary objects" with the ability to construct, destroy and/or re-make boundaries (Star and Griesemer 1989). The approach is exceptionally well-suited to studying the vertiginous development of ever more complex data science tools and infrastructures whose interdependencies and impact on knowledge production require unpacking and investigation. In my own experience of studying data journeys, I found a high level of interest in my results from

[18] See also Bowker (1994) and Star and Ruhleder (1996).

researchers and curators themselves, who are the first to acknowledge how hard it is for any one agent in the system to acquire an overarching view of how data travels. Such an overarching view is arguably impossible to achieve: data journeys, as narratives that bring together various parts of a journey and highlight its implications for (at least some parts of) knowledge production and society, may well constitute the next best thing.

By the same token, many of the advantages so far identified in the adoption of data journeys as a unit of analysis also constitute major challenges, at once conceptual and methodological, which all contributors to this volume had to face. Most obvious is the problem of *when journeys stop*. It is difficult to delimit a data journey, given both the variety of data uses that can derive from the publication of one dataset, and the current explosion of digital data infrastructures. Networks of data infrastructures and related uses can quickly become so complex as to be impossible to localise and track. This difficulty is compounded by the mutable and aggregate nature of data themselves, which makes data even more difficult to follow whenever they are recombined to constitute new aggregates (as discussed in Tempini's, Griesemer's and Morgan's chapters); and the problem of identifying who counts as a "user" of data at different points of a data journey (is it anybody who handles the data, for instance, or is it only those to interpret the data for purposes associated to knowledge-production?).

These issues cannot be settled in any general, abstract manner. As exemplified by the chapters of this volume, solutions to these challenges turn out to be highly situated, and the very opportunity to clearly articulate these challenges constitutes an advantage of adopting data journeys as units of analysis. Nevertheless, they ended up taking similar forms across chapters, thus giving rise to a coherent set of methodological preferences which all contributors converged upon, which I now briefly list:

- *Questioning "fixed" locations*: attention to data journeys involves purposefully looking beyond a specific location in time or space – whether this is conceptualised as a specific project, institution, system or even research field – and questioning what defines and constitutes a situation of inquiry at every step of the way and in clear relation to the goals of the groups involved;
- *Focusing on non-linear, multiple narratives:* reflecting the non-linear nature of data journeys themselves and the several strands of data practice (and related conceptualisations, goals and skills) that may end up animating the travels of a single dataset;
- *Utilizing detailed case studies* to explore and contrast the local characteristics of the data practices in question, for instance through ethnographies and historical reconstruction, thus recognising that the devil in data journeys is in the specific conditions under which movement happens;
- *Engaging with practitioners*: because of the importance of details and of familiarity with context, an embodied understanding of the skills, techniques and goals involved at different moment of a data journey provides a strong platform for interpretation and for assessing the extent to which the chosen cases act (or

not) as representatives for wider concerns and attitudes. The study of data journeys tends to be "in medias res", with science scholars often working alongside, and sometimes collaboratively, with data practitioners.

- *Meddling with other disciplinary lenses*: all contributors to this volume worked from a specific disciplinary/methodological perspective and yet engaged in frequent dialogue with scholars with different skills and goals (including other contributors of this volume), with the aim to heighten awareness of the many dimensions of data journeys and their implications for conceptualizations of data-intensive science. While this may not amount to fully fledged interdisciplinarity, it does call attention to the significance of interest in a multi-disciplinary approach, where historical and philosophical findings inform social scientific studies (and vice-versa).[19]
- *Attention to reflexivity:* ways in which each author carves out case study and identifies data journey is itself important to explicitly discuss, since it has strong repercussions on analysis and it always itself dependent on the analyst's own goals and vantage point. The position of the author depends partly on their own skills, preferences, aims and institutional position, and partly on the characteristics of the groups and data uses that they investigate. Unavoidably, engagement with data journeys typically requires tackling and confronting these issues in ways that make sense given one's interests and situation. Making one's perspective as explicit as possible in the narration of these stories is therefore important.[20]

Taken together, these methodological commitments constitute an overarching approach to the study of data journeys which facilitates the identification and study of common challenges, while at the same time maintaining the ambiguities and generative tensions that virtually all scholars engaged in data studies have identified as constitutive of the epistemic power of data.

5 Nodes of Difference and Similarity

While the range of data practices within this volume makes it impossible to offer a straight comparison between cases on the basis of their disciplinary provenance, some topics do emerge as crucial elements of data mobility across all chapters. In this section, I reflect on ways in which such elements can be used as nodes to identify and reflect upon differences and similarities among data journeys.

Perhaps the most obvious one, which resonates with existing scholarship and my remarks so far on the laboriousness of data journeys, is the *significance of cleaning*

[19] I discussed the value of bringing together philosophical, historical and sociological perspectives to study the management of data within bioinformatics in Leonelli (2010).

[20] The methodological and conceptual demand for reflexivity is discussed in most detail within Hoeppe's chapter.

and processing practices to the interpretation of data. The principles guiding data cleaning can change dramatically across areas, often due to the preferences developed by research communities dealing with different types of data, phenomena and research goals. This is illustrated in Boumans' and Leonelli's comparison between business cycle analysis in economics, where simplicity is regarded as a virtue, and plant phenomics in biology, where simplicity is viewed as potential oversimplification. The tools and methods used to clean data also range widely. In the cases discussed by Tempini and by Parker, attention falls on digital means of filtering data, where a given data format is preferred because it is compatible with existing software and related models. It is notable that despite pertaining to different research areas (environmental and climate science respectively), both examples concern situations where finding technical ways to share heterogeneous and geographically dispersed data is a priority. A different approach consists of identifying standards that can help to systematize vast amounts of data by narrowing down what counts as data in the first place, a phenomenon clearly illustrated by attempts to use biological, medical, socio-economic and environmental data for public health purposes documented in Ramsden's, Morgan's and Gaudillière's and Gasnier's chapters. Yet another take on data cleaning is to proritize circumstances of data use over the characteristics of the data objects in and of themselves, as exemplified by Hoeppe's study of what he calls "architectures of astronomical observations"; or to focus on the effects of data cleaning on a given audience, as illustrated by the selection of data points as markers of authenticity claims for artworks discussed by Rappert and Coopmans.

Visualisation and its power to stabilise data patterns and related interpretations is another theme to emerge strongly from the study of data journeys. Müller-Wille and Porter's cases, both of which concern the study of inheritance to determine recurrence of traits (respectively taken to denote race and mental illness) in specific populations, illustrate how the deployment of tables to visualise data is instrumental towards identifying patterns through which data are organised and understood – and crucially, to make such patterns robust over time even to changes in the underpinning datasets. Bechtel's discussion of network diagrams in contemporary biology provides another case where the patterns generated by a visualisation become themselves data to be disseminated and interpreted, thus engendering a data journey where movement and reuse are dependent on the tractability and interoperability of visualisations rather than of original sequencing data. Another take on sequencing data is provided by Griesemer, who emphasises the grouping of data into datasets as another type of patterning obtained through visualising tools such as Excel spreadsheets and computational interfaces, which transforms specific data ensembles into stable targets for investigation.

Visualisation tools play a central role in data journeys because data are often unwieldy and hard to amalgamate, homogenize or even coordinate. A key reason for this, particularly for data produced for research purposes, is that data are generated through instruments, techniques and methods that are finely tuned to the study of specific phenomena. Hence another node emerging from this volume is the relation between data and the world: that is, the *significance of the target system and its rela-*

tions to humans. The biological world, for instance, has long been perceived as consisting of "endless forms most beautiful" that require tailored research approaches. As discussed in Halfmann's chapter, the study of marine organisms tends to differ dramatically from that of trees, mammals and fungi, not to speak of the ubiquitous microbes whose activities intersect and underpin all other forms of life. This radical methodological pluralism results in myriads of data types, formats and labels, and resistance to overarching attempts at standardisation (as exemplified by Leonelli's plant phenomics).[21] The environmental sciences similarly need to tackle ever-transforming, unique ecosystems, and the biomedical and social sciences follow suit with the additional complications brought by the looping effects involved in humans studying humans – such as the capacity of practices of data classification to change the very phenomena that they identify, as in the case of categories of mental illness which Ian Hacking (2007) usefully described as "interactive kinds". At the same time, within these sciences the role of values and social priorities in guiding data production and interpretation tends to be particularly pronounced, with a desire for actionable knowledge structuring the choice of strategies and vehicles for data journeys and sometimes resulting in adherence to narrow standards for the sake of achieving socially relevant goals (as demonstrated by the chapters of the volume related to public health, including Ramsden, Gaudillière and Gasnier, Teira and Tempini, Morgan, and Cambrosio and colleagues). By contrast, the targets of natural sciences such as astronomy, physics and geology may be no less variable than the biological ones, but are generally perceived to be more independent from human experience (Daston and Lunbeck 2011). The sky thus works, in Hoeppe's terms, as a stable object which can be observed and re-observed across time; while in Koraka's discussion, the collision events studied in particle physics are assumed to be representative of the behaviour of all fundamental particles, regardless of location and circumstances – a commitment that simplifies the process of data amalgamation from different runs of an experiment.

Even where the target of data are assumed to be relatively homogeneous, however, data practices can differ on the basis of the *degree of entanglement perceived to exist between data and the instruments through which they are generated* (which may include conceptual tools like theories and models, or material tools like measuring or experimental apparatus). Within particle physics, the generation of data is deeply informed by theoretical models and the specificities of a highly complex experimental apparatus, as illustrated by Karaca's analysis of data acquisition procedures used at CERN. Similarly, Parker discusses the data-model symbiosis characterising much work in the climate sciences. It is hardly possible to thing about data as "raw" in such cases. The temptation to consider data as raw products of a situated interaction with nature arises more consistently in relation to biological and astronomic work, though even there the idea of 'observing' as a value-neutral, observer-independent activity is quickly dispelled. Rather than focusing on whether

[21] This in turn, somewhat paradoxically, makes it hard to estimate and research the very phenomenon of biodiversity (Müller-Wille 2017).

or not data are treated as raw documents of nature, contributors to the volume found it easier to examine stages of data processing and the extent to which certain traces are being transformed and modified in transit.[22] This is where the journey metaphor comes in useful, highlighting the value that certain kinds of data types, format and related practices of management and processing of data objects have, and how it can differ across communities and stages of travel. The question of "what constitutes raw data?" becomes "what typologies of data processing are there, and what do they achieve within different types of inquiry?"

The relation between *data and materials* such as samples, specimens and preparations deserves a special mention here, partly because it has attracted less attention than other aspects (both in the sciences and in science studies), but also because this is where we find some of the starkest discipline-related differences between data journeys. For archaeologists, for instance, materials are crucial anchors for inquiry, made even more important by their scarcity. Within the biological and biomedical sciences, samples are hard to obtain once data have been digitised and shared via databases. Even in cases where they are collected (such as biobanks, natural history museums or seed banks), samples are depletable and thus hard to access and reuse – and of course living organisms develop and evolve, making it hard to stabilise their characteristics so that they can act as a fixed reference point. Within social sciences such as economics and sociology, it is even harder to hold on to a material sample as populations are constantly transformed.

The *management of change and temporality within and beyond data infrastructures* can itself be considered as a node in the analysis and comparison of data journeys. We discussed how data are transformed through mobilisation, and how the target systems which data are supposed to document are also constantly changing. Notably, change in data and their use as evidence is separate and often disconnected from change in target systems. In other words, the processual nature of data as lineages is out of step with the processual nature of the entities that data are supposed to document: "data time" is not the same as "phenomena time" (Griesemer and Yamashita 2002, Leonelli 2018b). This mismatch can be highlighted or downplayed when ordering, visualizing and interpreting data as representations of specific phenomena – that is, when developing data infrastructures, data mining algorithms and models. This is a problem for (automated and complex) systems for big data analysis, where situated assessment of data provenance and the specific date on which data were originally collected is often unfeasible or side-stepped (Shavit and Griesemer 2009; Leonelli and Tempini 2018). The vast majority of data infrastructures and mining tools assume a static definition of knowledgebase, with no systemic provisions made for capturing change in target systems or in the data themselves. The reification processes involved here prove particularly pernicious when producing visualisations of data that build on each other at increasing levels of abstraction, as in the case of networks where creating links can be relatively simple but can make looking 'back' to the relation between networks and target systems fiendishly difficult.

[22] On the tracking of traces, see Rheinberger (2011).

All these considerations point to a final node of difference and similarity across data journeys, which is *the grounds on which those involved grant legitimacy and trustworthiness to the data*. This is where the cases within the volume show perhaps the greatest degree of variety, with multiple norms and concerns emerging in relation to different data uses. Wylie shows how belief in archaeological data can be warranted through frequent reanalysis of materials and triangulation of existing data with data obtained through new instruments and methods. The cases of Müller-Wille, Porter and Bechtel show visualisation tools adding legitimacy and longevity to biological data that would otherwise be highly contested, while Ramsden shows the links between the adoption of standards, the portability of the data and the degree to which they are accepted and used as grounds for public health decisions. Attitudes to data ownership, governance and authorship can also contribute to evaluations of data credibility, with concerns around ethics and security playing a particularly strong role in the travels of sensitive personal data (as shown in Teira and Tempini's discussion of the different roles that government may take in regulating the dissemination and reuse of medical records). The ways in which data journeys themselves are institutionalised, and the status of institutions themselves, are of course crucial to assessments of trustworthiness. Data regimes become reified and actualised through different types of platforms (Keating and Cambrosio 2003), repertoires (Ankeny and Leonelli 2016), market structures (Sunder Rajan 2016) and moral economies (Daston 1995, Pestre 2003, Strasser 2011), which shape the various ways in which data are valued, including their role as sources of evidence.

6 Conclusion: Why Study Data Journeys?

The approach to data journeys that I sketched here helps to trace the relations between the goals guiding different types of data use and the methodological, epistemic, cultural and political commitments favoured within those situations as they develop and transform over time. This may not be as satisfactory as a straightforward list of components essential to all data journeys or universal conditions under which data are likely to be reused – but the experiences of authors researching data movements, within and beyond this volume, indicate that such a straightforward list may not exist. This finding chimes with the failure of scientific attempts to find universal standards and guidelines for data interoperability and reuse, which resulted in the top global organisations focusing on data curation and dissemination (including the Research Data Alliance, CODATA, the European Open Science Cloud and the Digital Data Curation Centre) backing a discipline-specific approach, within which diversity in epistemic cultures is taken as the starting point for devising data management practices, rather than as an obstacle to overcome to make data travel. The studies contained in this volume point to a yet more radical approach: rather than discipline-specific, the communalities and differences in data journeys emerge

as *use-specific*, and thus dependent on the goals, commitments and tools available to those seeking to extract meaning from data within specific situations.

It could be objected that the focus on data journeys as units of analysis, being so strongly steeped in history, necessarily constitutes a "a posteriori" view of what already happened, which cannot provide insight into current and future events - particularly given the unpredictability of journeys themselves. It is not a coincidence that the best examples of data re-use in this volume come from historical work from the nineteenth and twentieth century. For the more contemporary data journeys documented in this volume, most of which are still ongoing, it may even be too soon to tell about re-use. This should not come as a surprise, given the deep link between the epistemic value of data and their mobility. When conceptualising data themselves as mutable mobiles, data management and use are by definition moving targets, and any attempt to narrate data use necessarily turns away from its present dynamics. This does not mean that the study of data journeys cannot offer lessons for the future. Quite the opposite: this approach provides a way to pose the fundamental normative question, "what are data journeys good for?"

Asking this question is crucial at a time in which reliance on the "power of big data" permeates public discourse. The possibility to bring lots of data together is often hailed as a force for good, capable of revolutionizing the third sector (for instance through the personalisation of service provision) and fixing virtually any social and environmental problem, ranging from pollution to inequality. Focusing on the challenges and strictures of data travel is an excellent antidote to such hype. Understanding the conditions under which data come to be used, including the various stages and processes through which that use is made possible, shines a light on the costs and opportunities involved in data mobility. Data journeys need to be reconstructed and studied with equal attention to technical and to social aspects, thus displaying the extent to which value judgements and financial incentives intersect with scientific goals and technological innovation. This is key to contemporary debates around data storage, protection, security and use, as well as the meaning of openness and fairness in information sharing and the development of artificial intelligence. How are big (and small) data transformed into scientific knowledge, with what implications, and how can the reliability of such knowledge be assessed?[23] Who do data journeys benefit and who do they damage, when and how? Answering these questions requires delving in both the technical and the social worlds of data, thus identifying conceptual and material commitments and their repercussions in terms of who is included, excluded or ignored by such knowledge-making processes. By embodying this type of analysis, this volume exemplifies the value of bringing scholarship from history, philosophy and social studies of science to bear on issues of central concern to contemporary science and science policy.

[23] On the social challenges posed by the use of big data, see for instance the seminal work of dana boyd (e.g. 2012).

References

Ankeny, Rachel A., and Sabina Leonelli. 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.

Ankeny, Rachel A. this volume. Tracing Data Journeys Through Medical Case Reports: Conceptualizing Case Reports Not as "Anecdotes" but Productive Epistemic Constructs, or Why Zebras Can Be Useful. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Aronova, Elena, Christine von Oertzen, and David Sepkoski. 2018. Introduction: Historicizing Big Data. *Osiris* 32 (1): 1–17.

Bates, Jeanne, Y.-W. Lin, and P. Goodale. 2016. Data Journeys: Capturing the Socio-Material Constitution of Data Objects and Flows. *Big Data & Society* 3 (2): 205395171665450.

Bechtel, William. this volume. Data Journeys Beyond Databases in Systems Biology: Cytoscape and NDEx. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Blair, Ann. 2010. *Too Much to Know: Managing Scholarly Information Before the Modern Age*. New Haven/London: Yale University Press.

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

Boulton, Geoffrey, P. Campbell, B. Collins, et al. 2012. *Science as an Open Enterprise*, The Royal Society Science Policy Centre Report 02/12. London: The Royal Society Publishing.

Boumans, Marcel, and Sabina Leonelli. this volume. From Dirty Data to Tidy Facts: Clustering Practices in Plant Phenomics and Business Cycle Analysis. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Bowker, Geoffrey C. 1994. *Science on the Run: Information Management and Industrial Science at Schlumberger, 1920–1940*. Cambridge, MA: MIT Press.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: The MIT Press.

boyd, dana. 2012. Critical Questions for Big Data. *Information, Communications Society* 4462: 37–41.

Cai, Li, and Yangyong Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal* 14: 2.

Cambrosio, Alberto, Jonah Campbell, Etienne Vignola-Gagné, Peter Keating, Bertrand R. Jordan, and Pascale Bourret. this volume. 'Overcoming the Bottleneck': Knowledge Architectures for Genomic Data Interpretation in Oncology. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Coopmans, Catelijne, and Brian Rappert. this volume. Data Journeys in Art? Warranting and Witnessing the 'Fake' and the 'Real' in Art Authentication. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Daston, Lorraine. 1995. The Moral Economy of Science. *Osiris* 10: 2–24.

———. 2017. *Science in the Archives*. Chicago, IL: Chicago University Press.

Daston, Lorraine, and Elisabeth Lunbeck. 2011. *Histories of Scientific Observation*. Chicago, IL: Chicago University Press.

Ebeling, Mary F.E. 2016. *Healthcare and Big Data. Digital Specters and Phantom Objects*. New York: Palgrave Macmillan.

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

Edwards, Paul N., M.S. Mayernik, A.L. Batcheller, et al. 2011. Science friction. Data, Metadata, and Collaboration. *Social Studies of Science* 41 (5): 667–690.

Egyedi, Tineke M. and Donna C Mehos. 2015. *Inverse Infrastructures*. EE.

European Commission. 2016. *Open innovation, open science, open to the world – A vision for the future*. Directorate-General for Research and Innovation. http://bookshop.europa.eu/en/open-innovation-open-science-open-to-the-world-pbKI0416263/. Accessed 9 Sept 2019.

———. 2017. *Incentives and Rewards to Engage in Open Science Activities*. Thematic Report No 3 for the Mutual Learning Exercise Open Science: Altmetrics and Rewards of the European Commission. https://rio.jrc.ec.europa.eu/en/library/mutual-learning-exercise-openscience-%E2%80%93-altmetrics-and-rewards-incentives-and-rewards-engage. Accessed January 2020.

Floridi, Luciano. 2011. *The Philosophy of Information*. Oxford: Oxford University Press.

Floridi, Luciano, and Phyllis Illari. 2014. *The Philosophy of Information Quality*. Springer.

Gaudilliere, Jean-Paul, and Camille Gasnier. this volume. From Washington DC to Washington State: The Global Burden of Diseases Data Basis and the Political Economy of Global Health. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Gitelman, Lisa. 2013. *Raw Data' Is an Oxymoron*. Cambridge, MA: MIT Press.

Griesemer, James. this volume. A Data Journey Through Dataset-Centric Population Genomics. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Griesemer, James R., and Grant Yamashita. 2002. Zeitmanagement bei Modellsystemen. Drei Beispiele aus der Evolutionsbiologie. In *Lebendige Zeit*, ed. H. Schmidgen, 213–241. Berlin: Kulturverlag Kadmos. Managing Time in Model Systems: Illustrations from Evolutionary Biology. Published in German in 2005.

Global Young Academy. 2016. *Open Data Position Statement of the Global Young Academy and the European Young Science Academies*. http://globalyoungacademy.net/wp-content/uploads/2016/04/Position-Statement-on-Open-Data-by-the-Young-Academies-of-Europe-and-the-Global-Young-Academy.pdf. Accessed January 2020.

Hacking, Ian. 2007. Kinds of People: Moving Targets. *Proceeding of the British Academy* 151: 285–318.

Halfmann, Gregor. this volume. Material Origins of a Data Journey in Ocean Science: How Sampling and Scaffolding Shape Data Practices. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Hey, Tony, Stewart Tansley, and Kristine Tolle, eds. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Hilgartner, Stephen. 2017. *Reordering Life: Knowledge and Control in the Genomics Revolution*. Cambridge, MA: MIT Press.

Hoeppe, Götz. this volume. Sharing Data, Repairing Practices: On the Reflexivity of Astronomical Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Karaca, Koray. this volume. What Data Get to Travel in High Energy Physics? The Construction of Data at the Large Hadron Collider. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Keating, Peter, and Alberto Cambrosio. 2003. *Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine*. Cambridge, MA: MIT Press.

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. London, UK: Sage.

Kitchin, Rob, and G. McArdle. 2016. What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society* 3 (1): 1–10.

Lagoze, Carl. 2014. Big Data, Data Integrity, and the Fracturing of the Control Zone. *Big Data & Society* 1 (2): 2053951714558281.

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. Cambridge, MA: Harvard University Press.

Leonelli, Sabina. 2010. Documenting the Emergence of Bio-ontologies: Or, Why Researching Bioinformatics Requires HPSSB. *History and Philosophy of the Life Sciences* 32 (1): 105–126.

———. 2012. When Humans Are the Exception: Cross-Species Databases at the Interface of Clinical and Biological Research. *Social Studies of Science* 42 (2): 214–236.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: Chicago University Press.
———. 2017. Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science Journal* 16 (32): 1–11.
———. 2018a. *La Ricerca Scientifica nell'Era dei Big Data.* Meltemi Editore.
———. 2018b. The Time of Data: Time-Scales of Data Use in the Life Sciences. *Philosophy of Science* 85 (5): 741–754.
Leonelli, Sabina. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Leonelli, Sabina, and Tempini Niccolo. 2018. Where health and environment meet: The use of invariant parameters in big data analysis. *Synthese*. https://doi.org/10.1007/s11229-018-1844-2.
Maxson, Kathryn M., Robert Cook-Deegan, and Rachel A. Ankeny. 2018. The Bermuda Triangle: Principles, Practices, and Pragmatics in Genomic Data Sharing. *Journal for the History of Biology* online first.
McNally, Ruth, Adrian Mackenzie, Allison Hui, and Jennifer Tomomitsu. 2012. Understanding the 'Intensive' in 'Data Intensive Research': Data Flows in Next Generation Sequencing and Environmental Networked Sensors. *International Journal of Digital Curation* 7 (1): 81–95.
Meng, Xiao-Li. 2019. Data Science: An Artificial Ecosystem. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.ba20f892. Accessed 9 Sept 2019.
Mirowski, Philip. 2018. The Future(s) of Open Science. *Social Studies of Science* 48 (2): 171–203.
Mongilli, Alessandro, and Giuseppina Pellegrino, eds. 2014. *Information Infrastructure(s). Boundaries, Ecologies, Multiplicity*. Cambridge: Cambridge Scholars Publishing.
Morgan, Mary S. 2010. Introduction. In *How Well Do Facts Travel*, ed. P. Howlett and M.S. Morgan. Cambridge, UK: Cambridge University Press.
Morgan, Mary S. this volume. The Datum in Context: Measuring Frameworks, Data Series and the Journeys of Individual Datums. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Müller-Wille, Staffan. 2017. Names and numbers: 'Data' in classical natural history, 1758–1859. *Osiris* 32 (1): 109–128. https://doi.org/10.1086/693560.
Müller-Wille, Staffan. this volume. Data, Meta Data and Pattern Data: How Franz Boas Mobilized Anthropometric Data, 1890 and Beyond. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
OECD. 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. http://www.oecd.org/science/sci-tech/38500813.pdf Accessed 9 Sept 2019.
Open Science Policy Platform. 2018. OSPP-REC: Recommendations of the Open Science policy platform. https://ec.europa.eu/research/openscience/pdf/integrated_advice_opspp_recommendations.pdf. https://doi.org/10.2777/958647. Accessed January 2020.
Parker, Wendy S. this volume. Evaluating Data Journeys: Climategate, Synthetic Data and the Benchmarking of Methods for Climate Data Processing. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Pasquale, Frank. 2015. *The Black Box Society. The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.
Pasquetto, Irene V., B.M. Randles, and C.L. Borgman. 2017. On the Reuse of Scientific Data. *Data Science Journal* 16: 8.
Pestre, Dominique. 2003. Regimes of Knowledge Production in Society: Towards a More Political and Social Reading. *Minerva* 41: 245–261.
Porter, Theodore M. this volume. Most Often, What Is Transmitted Is Transformed. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Ramsden, Edmund. this volume. Realizing Healthful Housing: Devices for Data Travel in Public Health and Urban Redevelopment in the Twentieth Century United States. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Rheinberger, Hans-Jörg. 2011. Infra-Experimentality: From Traces to Data, From Data to Patterning Facts. *History of Science* 49 (164): 337–348.
Shavit, Ayelet, and James Griesemer. 2009. There and Back again, or the problem of locality in biodiversity surveys. *Philosophy of Science* 76 (July): 273–294. https://doi.org/10.1086/649805.
Shavit, Ayelet, and James R. Griesemer. 2011. Transforming Objects into Data: How Minute Technicalities of Recording 'Species Location' Entrench a Basic Challenge for Biodiversity. In *Science in the Context of Application*, ed. Martin Carrier and Alfred Nordmann, 169–193. Boston, MA: Boston Studies in the Philosophy of Science.
Srnicek, Nick. 2017. *Platform Capitalism*. Cambridge/Malden: Polity Press.
Star, Susan L., and James R. Griesemer. 1989. Institutional Ecology, Translations and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science* 19 (3): 387–420.
Star, Susan L., and Katherine Ruhleder. 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7 (1): 111–134.
Stevens, Hallam. 2013. *Life Out of Sequence: Bioinformatics and the Introduction of Computers into Biology*. Chicago: University of Chicago Press.
Strasser, Bruno J. 2011. The Experimenter's Museum GenBank, Natural History, and the Moral Economies of Biomedicine. *Isis* 102 (1): 60–96.
Sunder Rajan, Kaushik. 2016. *Pharmocracy: Value, Politics and Knowledge in Global Biomedicine*. Durham: Duke University Press.
Tempini, Niccolò. this volume-a. The Reuse of Digital Computer Data: Transformation, Recombination and Generation of *Data Mixes* in Big Data Science. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Tempini, Niccolò. this volume-b. Visual Metaphors: Howardena Pindell, Video Drawings, 1975. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Tempini, Niccoló, and Sabina Leonelli. 2018. Concealment and Discovery: The Role of Information Security in Biomedical Data Re-Use. *Social Studies of Science* 48 (5): 663–690.
Tempini, Niccolò, and David Teira. this volume. The Babel of Drugs: On the Consequences of Evidential Pluralism in Pharmaceutical Regulation and Regulatory Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Wilkinson, Mark D., et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018.
Wouters, Paul, Anne Beaulieu, and Andrea Scharnhorst. 2013. In *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences*, ed. Sally Wyatt. Cambridge, MA: The MIT Press.
Wylie, Alison. 2002. *Thinking from Things. Essays in the Philosophy of Archaeology*. Berkeley: University of California Press.
Wylie, Alison. this volume. Radiocarbon Dating in Archaeology: Triangulation and Traceability. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Sabina Leonelli is Professor of Philosophy and History of Science at the University of Exeter, where she codirects the Exeter Centre for the Study of the Life Sciences (Egenis) and leads the "Data Governance, Algorithms and Values" strand of the Institute for Data Science and Artificial Intelligence. Her research concerns the epistemology and governance of data-intensive science, the philosophy and history of organisms as scientific models and the role of open science in the global research landscape. She has an interest in science policy and served as expert for national and international bodies including the European Commission. She is a Turing Fellow, Editor-in-Chief of *History and Philosophy of the Life Sciences* and Associate Editor of the *Harvard Data Science Review.* Her publications span philosophy, social science, biology, history, data science and science policy and include the monographs *Data-Centric Biology: A Philosophical Study* (2016) and *La Recherche Scientifique à l'Ère des Big Data* (2019). Between 2014 and 2019, she led the European Research Council Starting Grant "The Epistemology of Data-Intensive Science" which supported the development of this volume.