Sabina Leonelli
Niccolò Tempini
Editors

Data Journeys in the Sciences

Springer Open

Data Journeys in the Sciences

Sabina Leonelli • Niccolò Tempini
Editors

Data Journeys in the Sciences

Springer Open

Editors
Sabina 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

Niccolò Tempini
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

ISBN 978-3-030-37176-0 ISBN 978-3-030-37177-7 (eBook)
https://doi.org/10.1007/978-3-030-37177-7

Cover image produced by Michel Durinx [centimedia.org] on the basis of extracts from Holland, W.J. (1922) The butterfly book; a popular guide to a knowledge of the butterflies of North America. Garden City, N.Y., Doubleday (available from the Biodiversity Heritage Library, DOI: https://doi.org/10.5962/bhl.title.5524).

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface: A Roadmap for Readers

What is the point of data in research? Philosophers and methodologists have long discussed the use of data as empirical fodder for knowledge claims, highlighting, for instance, the role of inductive reasoning in uncovering what data reveal about the world and the different ways in which data can be modelled and interpreted through statistical tools. This view of data as a fixed, context-independent body of evidence, ready to be deployed within models and explanations, also accompanies contemporary discourse on Big Data – and particularly the expectation that the dramatic increase in the volume of available data brings about the opportunity to develop more and better knowledge. When taking data as ready-made sources of evidence, however, what constitutes data in the first place is not questioned, nor is the capacity of data to generate insight. The spotlight falls on the sophisticated algorithms and machine learning tools used to interpret a given dataset, not on the efforts and complex conditions necessary to make the data amenable to such treatment. This becomes problematic particularly in situations of controversy and disagreement over supposedly "undisputable" data, as in the case of current debates over the significance of climate change, the usefulness of vaccines and the safety of genetic engineering. Without a critical framework to understand *how* data come to serve as evidence and the conditions under which this does or does not work, it is hard to confront the challenges posed by disputes over the reliability, relevance and validity of data as empirical grounds for knowledge claims.

In this volume, we move decisively away from the idea that what counts as data – and in turn, how data are presented, legitimized and used as evidence – can be given for granted and that finding the correct interpretative framework is all that is required to make data "speak for themselves". We focus instead on the strategies, controversies and investments surrounding decisions around what researchers identify and use as data in the first place: in other words, the myriad of techniques, efforts, instruments, infrastructures and institutions used to process and mobilize data so that it can actually serve as evidence. No matter how "big" data are, the road from data to knowledge remains complex and full of obstacles. The contributions collected in this book highlight a wide spectrum of activities involved in handling data – including data collection, aggregation, cleaning, dissemination, publication, visualization

and ordering – with the aim to study the opportunities and constraints posed by such activities on how data are eventually interpreted. Hence, the initial question about what role data play in research morphs into two further queries: What conditions are required to identify data in the first place and to make them usable as evidence? And what implications does data processing carry not just for the content of the knowledge being produced but for the extent to which that knowledge can ground interventions in the world and inform political, scientific, social, economic debate?

The contributions to this volume help readers to ponder these questions by guiding them into the thick web of entanglements involved in making data move across time, space and social context. Readers are asked to accompany data in their journeys from their material *origin* through human interactions with the world (which range from the collection of objects to the generation of traces and measurements) to their *dissemination* across various forms of aggregation (datasets, data series, indicators) and vehicles (databases, publications, archives) and ultimately to their *use* as evidence for claims.[1] During these journeys, data experience many different types of encounters – with other data, diverse groups of users, specific infrastructures and technologies and political, economic and cultural expectations – which affect and shape the data themselves and their prospective usability. Far from underestimating the politics and power of data, which so many contributors to the emerging field of critical data studies have so effectively highlighted, we seek to document how such politics is embedded, reified and/or revised in the technical and epistemic work that structures everyday research practices. Delving into stories of how data travel involves seeing data as entities that can, and often do, change their properties in response to their environment and relations – and whose travels are often choreographed and regulated to achieve a variety of (sometimes incompatible) goals. What comes to be seen as datum at any one point in time is itself the result of a journey; and far from being linear and well-organized, the journey is often full of detours and unpredictable changes, largely due to the diverse and complex social networks and contexts responsible for making data move.

Unsurprisingly, the study of data in motion generates a wealth of insights that are not easily systematised in one thread. When first imagining this volume, I had envisioned a straightforward comparative work, which would examine differences and similarities across data practices in the biological, biomedical, environmental, physical and social sciences. Following 5 years of discussions with my coeditor and the wonderful team of authors assembled here, however, the illusion that data could be disciplined and contained in this way has been shattered. Data journeys transcend and defy disciplinary boundaries both in the methods used to track and analyse them and in the domains in which data come to be seen as valuable. Of course, as many of these chapters forcefully illustrate, epistemic cultures and context-specific norms shape and direct the journeys and uses of data. And yet, even when looking at highly discipline-specific cases (such as the analysis of genomic data in population biology or observations in astronomy), we found surprising parallels and intersections with other social and epistemic worlds – and a wealth of opportunities for data initially

[1] For an extended discussion of data journeys as a theoretical and methodological tool, see the introduction to the volume.

collected for a specific, supposedly self-contained purpose to journey further, towards new settings and unforeseen uses.

To convey this unpredictability, the structure of the volume is organized along what we imagine to be the *stages* of a data journey, with each cluster of chapters discussing the skills, methods, activities and norms that may be typically associated to those stages across a wide range of research areas. The first section is devoted to the *origins* of data: from the choice and interaction with material samples of the world from which data will be extracted (Halfmann) to the role of theories and instruments in data generation (Karaca) and the questions involved in choosing a vehicle for data to start their travels towards new interpretations (Ankeny). The second section examines the ways in which data are brought together for analysis and specifically the practices, standards and tools involved in data *cleaning* (Boumans and Leonelli), *clustering* (Morgan and Griesemer) and *visualizing* (Bechtel) – with a strong emphasis on the challenges and opportunities presented by the aggregation of data coming from different sources towards novel uses and interpretations. The third section explores the circumstances and implications of data *sharing*, paying attention particularly to the tight intersection between decisions about who can access the data and criteria used to evaluate and regulate their quality and reliability (both within research communities, as discussed by Parker in relation to climate data and Hoeppe in relation to astrological observations, and within broader policy and governance circles, as considered by Teira and Tempini in the case of patient data). The fourth section considers data *interpretation* and highlights the ways in which commitments to analytic techniques, instruments and concepts (Wylie) as well as decisions around what is considered to be data (Tempini) and metadata (Müller-Wille) may need to be transformed in order for data to be fruitfully used or reused within new situations. The fifth and final sections juxtapose different cases of data journeys to raise questions about what *fruitful data use* may actually consist of: first, by focusing on the procedures used to make data and related claims actionable, credible and accountable to the various types of publics and goals involved in data journeys, ranging from clinical settings (Cambrosio et al.) to public health (Ramsden) and related policies (Gaudilliere and Gasnier) and, second, by questioning the very narrative of authentication and discovery that often underscores the use of data as evidence (Rappert and Coopmans).

Reading the volume in this order will help those interested in the full arch of data journeys to understand the specificities associated to different stage of travel and the deep interrelations and intersections across them. This is but one type of variety encountered in the qualitative study of data movements, however. I want to mention another six for readers to consider before delving in and deciding how to engage with this book.

- **Variety of research domains**
  While many of the most fascinating data journeys are not contained within traditional research domains, the authors assembled within this volume have been purposefully approached for their deep, long-term engagement with specific research areas, so that the volume could encompass a wide range of disciplinary areas and illustrate the many sources of variety among research areas and epis-

temic communities. Readers interested primarily in *biology* could focus on the chapters by Bechtel and Griesemer, who consider different stages in the journeys of genomic data, and then look to the chapters by Tempini, Cambrosio et al., Müller-Wille and Ramsden to witness how biological data move into the *biomedical, policy and public health* domains. Students of biomedicine should prioritize the chapters by Ankeny, Tempini, Ramsden, Cambrosio et al. and Gaudilliere and Gasnier; those interested in the *environmental* sciences should start with Halfmann (oceanography) and Parker (climate science); those interested in the *physical* sciences should read Koraka (particle physics) and Hoeppe (astronomy); and for the *social and historical* sciences, the spotlight shifts to chapters by Morgan (economics), Wylie (archaeology) and Rappert and Coopmans (art authentication). The chapter by Boumans and Leonelli exemplifies the attempt to compare data practices across two very different domains: economics and plant science.

- **Variety of relations between data and other research components**
  A key question arising when interrogating the nature of data is the extent to which their role and characteristics differ from the ones associated to metadata, materials, models, apparatus and infrastructures. Several contributors to the volume address this issue directly. Halfmann starts this thread by questioning the relationship between data and *samples*, which has been rarely discussed within data studies so far. Hoeppe and Karaca focus on the entanglements between data and *instrumentation*, particularly in cases – such as Hoeppe's telescopes and Karaka's particle accelerator – in which whole research communities are formed around highly complexity and expensive apparatus. Parker discusses instead the relationship between *models* and data in the climate sciences and the extent to these two types of scientific objects are co-constructed and unavoidably intertwined. Griesemer, Müller-Wille, Tempini, Porter and Bechtel focus on the role of *infrastructures* in data visualization and the extent to which choices made in order to make data widely accessible and searchable affect their interpretation – but also what gets to count as data and metadata. And last but not least, Morgan, Teira and Tempini, Ramsden and Boumans and Leonelli consider the development and role of *standards* and *measurement frameworks* alongside the travel and clustering of multiple datasets, as crucial components of both the technical and the institutional motors for the travel of data.
- **Variety of data vehicles**
  A closely related concern is the question of how data actually travel, which, borrowing the terminology devised by Morgan (2010) to discuss travelling facts, can be usefully characterized as a question around vehicles. As mentioned above, many of the chapters in the volume discuss the characteristics of *databases* and related *search engines*, thus contributing to the burgeoning scholarship on data infrastructures pioneered by Bowker (1994), and rightly seen to be central to understanding how data move and land in new epistemic spaces. In addition to data infrastructures, the volume considers well-trodden but no less vivid vehicles, such as *case reports* – heavily descriptive narratives which Ankeny highlights as fruitful in identifying, capturing and ordering data for future analysis.

Bechtel's analysis of the travel of genomic data beyond databases to various tools of network analysis points instead to the complexities of interlocking data infrastructures that build upon each other, a situation in which *software* itself comes to play a crucial role as data vehicle – as explicitly discussed by Tempini's reflection on the travels of digital data. Less intuitively and perhaps more controversially, particular forms of *governance*, such as the monitoring of global health by the United Nations and the shifts in pharmaceutical regulation by the Food and Drug Administration, can themselves constitute effective vehicles for data journeys, as considered in the chapters by Gaudilliere and Gasnier and Teira and Tempini.

- **Variety of grounds for legitimacy**
  The question of what makes data reliable, legitimate and trustworthy is another key issue underpinning several of the chapters, both because of its importance to understanding the role of data as empirical evidence and because it is often a central concern for the protagonists of the case studies discussed by volume contributors. What does it mean for data to be fit for purpose? In other words and paraphrasing a seminal discussion by Clarke and Fujimura (1992) on the epistemic roles of tools in biological research, what count as the "right" data for the job, and how do we verify the credibility of data interpretations? Perhaps most striking in this respect is Wylie's investigation of the shifting grounds through which different generations of archaeologists have assessed the legitimacy and significance of carbon dating as a method for data collection and (re)interpretation. Similarly focused on intergenerational understandings of data, Müller-Wille discusses how physiological and sociological data on a controversial issue such as race managed to retain credibility for over a century, while Parker analyses benchmarking practices in relation to climate data sourced at very different locations and times (and thus hard to align and homogenize). A different approach consists of understanding how changes to the very properties of datasets – and the metadata that accompanies them – can be credibly framed as strategies to increase the usefulness and reliability of data as evidence. This question is confronted by Morgan in relation to data aggregation and Tempini with reference to their computational handling, while Cambrosio et al. focus on the shift from talk of "data" to talk of "knowledge" in medical information infrastructures, which is tied to the emergence of consensus around what "levels of evidence" are needed for clinical interventions and which sources of knowledge can be trusted.
- **Variety of data types**
  Perhaps most striking, particularly to readers used to think of data as computable numbers, is the breadth of data types considered in this volume. While some authors (e.g. Morgan) focus specifically on the properties of numerical data, it soon becomes clear that the objects identified and used as data within research are not limited to the quantitative results of measurement practices. Observations, both in textual and graphical forms, are common in medicine (Ankeny), astronomy (Hoeppe) and the life sciences (Boumans and Leonelli), where images and diagrams function both as containers of data (e.g. Griesemer) and as data in and of themselves (Bechtel, Cambrosio et al). The transformations involved in digi-

tizing analogue objects (Halfmann) and making them amenable to different types of computation (Tempini) complicate easy distinctions between quantitative and qualitative data. Different data properties and formats are instead often in a historical continuum, related to each other by specific technologies and techniques employed at different times to extract various forms of insight. At the same time, the preference for numerical data that can be easily aggregated – such as the assessment of economic performance (Boumans and Leonelli) or national compliance with given objectives (Gaudilliere and Gasnier) – can skew the analysts' attention, with significant implications for what kinds of knowledge become established (Porter 1995). The extent to which data are amenable to visualization is also a crucial determinant of mobility (Porter).

- **Variety of methodological approaches to the study of data**
  The volume puts philosophical, historical and sociological methods of research in dialogue with each other, thus bringing together different styles and disciplinary approaches to the study of data movements. All contributors are conversant with different disciplinary approaches, which they merge to consider data journeys from a qualitative viewpoint – a somewhat unavoidable choice given the importance of understanding motivations, goals and historical circumstances in order to track data and reconstruct their travels. This multidisciplinarity is a key characteristic of the volume and the result of the authors' own commitment to dialogue across fields as well as the extensive conversations held during the 5 years in which the volume was assembled – exemplified most directly by the comparison of data cleaning practices in economics and plant science coauthored by Boumans and Leonelli. The emphasis and argumentative style of authors does, at the same time, reflect the differences in their expertise, which could also be used as an entry point for readers. Authors with a stronger background in *history* provide vivid narratives of data moving across long time periods and multiple geographic sites, thus fostering an understanding of the long dureé of data journey and the enabling or constraining role played by institutions such as the American Public Health Association (Ramsden) and the Institute for Health Metrics and Evaluation (Gaudilliere and Gasnier) and political debates like those surrounding the notion of race (Müller-Wille). Authors rooted in *social studies of science* provide ethnographic forays into the goals, expectations and social organization of researchers, which helps to better understand apparently straightforward practices such as observation in astronomy (Hoeppe) and the interpretation of biomarkers in clinical practice (Cambrosio et al); and those more *philosophically oriented* delve deep into the technical, material and conceptual tools employed to structure, order and analyse data, thus highlighting the epistemic role of, for example, samples (Halfmann), experimental apparatus (Koraka), digital formatting (Tempini), visualization tools (Bechtel, Griesemer), evaluation practices (Parker, Boumans and Leonelli, Teira and Tempini) and ordering or narrative devices (Ankeny, Morgan).
- **Variety of data politics**
  Because the epistemic work underpinning data processing and movement is unavoidably value-laden, our authors' own political commitments around key data-related concerns are also in evidence within each piece. These commit-

ments are not uniform, not so much due to overt disagreement over the same issues but rather in the sense of authors being interested in different forms of politics. Both Ramsden's analysis of public health data journeys and Teira and Tempini's work on electronic health records focus on the potential for inequality and governmental exploitation of such data to implement specific forms of social control. Griesemer is more concerned with how social and racial representation is handled through the travel of genomic data and what groups are excluded or included by database structures. Morgan worries about the diverse measuring frameworks through which different types of datasets are clustered together and the resulting unevenness and potential loss of meaning when using such diverse clusters as indicators – as with the Sustainable Development Goals of the United Nations. Wylie is similarly interested in questions of legacy and in the accountability of temporal and methodological discontinuities in the handling and interpretation of data as evidence (in archaeology and beyond). Rappert and Coopmans grapple with questions of trust and authority in delivering judgements over data interpretation. And a whole set of authors, including Ankeny, Boumans and Leonelli, Tempini and Halfmann, worry about the opacity of data-handling processes that have a strong and yet underacknowledged effect on how data are then used and interpreted. All of these concerns are deeply political and have significant implications for ongoing debates around, for example, the trustworthiness of Big Data as source of evidence and the potential for inequality and exploitation underpinning open data policies.

These roadmaps are by no means exhaustive but hopefully provide at least a sense of the breadth and import of the material presented in this volume. We encourage our readers to find their own approach to the chapters and let themselves be challenged by these wide-ranging, diverse and sometimes challenging discussions, whose overarching aim is to provide a feel for the sophistication, complexity and epistemic significance of efforts devoted to data mobility within research and beyond.

Exeter, UK
London, UK

Sabina Leonelli

References

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.

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.

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 Geophysics at Schlumberger, 1920-1940*. MIT Press.

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.

Clarke, Adele E. and Joan H. Fujimura. eds. 1992. *The Right Tools for the Job: At Work in Twentieth-Century Life Sciences*. Princeton University Press.

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.

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.

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.

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.

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.

Morgan, Mary S. 2010. Travelling Facts. In: Howlett, P. and M. S. Morgan, *How Well Do Facts Travel? The Dissemination of Reliable Knowledge.* Cambridge University Press, pp. 3-42.

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. 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.

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.

Porter, Theodore M. 1995 *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.

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.

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, 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.

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.

Acknowledgements

We are deeply grateful to the many researchers across the humanities, social sciences and natural sciences, who took time and effort to interact with us within the context of the 5-year European Research Council project "The Epistemology of Data-Intensive Science". Some of those interactions go back to SL's work in the context of the "How Well Do Facts Travel" Leverhulme Trust project led by Mary Morgan at the London School of Economics and Political Science in 2008, when some of the contributors to this volume first gathered to discuss how research data move within and across contexts and with which implications. The funding by the European Research Council enabled us to work together and to organize four workshops and several sessions at conferences between 2014 and 2019, through which we pursued our common interest in data mobility in a variety of fields. It is impossible to acknowledge every individual who helped us along the way, though a record of these multiple discussions is available on the project website www.datastudies.eu. Here, we wish to give particular thanks to our authors for their long-term commitment to discussing data journeys, their support through the journey it took to bring this book to life and their stellar contributions. We also wish to acknowledge scholars who took the onerous role of discussants or who participated in key project events but whose work does not appear (or, at least, not directly) in this volume: Margunn Aanestad, Elizabeth Arnauld, Elena Aronova, Ruth Bastow, Dan Bebber, Louise Bezuidenhout, Andy Boyd, Gail Davies, John Dupré, Lora Fleming, Luciano Floridi, David Ford, Simon Forbes, Mary Ebeling, Carole Goble, Harriet Gordon-Brown, Sara Green, Steve Hinchliffe, Eva Huala, Susan Kelly, Javier Leuzan, James Lowe, James McAllister, Pilar Ossorio, Nick Provart, Barbara Prainsack, Kaushik Sunder Rajan, Julian Reiss, Federica Russo, David Sepkoski, Nick Smirnoff, David Studholme, Bruno Strasser, Hans-Jörg Rheinberger, David Ribes, Eran Tal, Sharon Traweek, Matthew Vaughn, Robin Williams and Sally Wyatt.

We have benefitted enormously from the unwavering support of our colleagues at the Exeter Centre for the Study of the Life Sciences. In particular, Michel Durinx provided precious assistance in preparing some of the figures in this volume, including its cover; Gregor Halfmann, the now successfully graduated PhD student on the project, shared many insights with us throughout its development; and Chee Wong

has been the best Project Administrator that we could ask for and the organizer of all the Exeter-based events from which the volume emerged.

Christi Lue, Ties Nijssen, Nitesh Shrivastava, Sanjeev KeerthiKumari and others in the production team of Springer assisted us admirably during the publication phase. The bulk of the research work and the whole publishing cost of this volume were funded by the European Research Council Grant Number 335925, for which we are immensely thankful especially at this time of anti-European sentiment in the United Kingdom. In the final year of preparations, we also benefited from support from the Alan Turing Institute, EPSRC Grant EP/N510129/1.

Last but not least, we wish to thank Gemma (NT) and Michel, Luna and Leonardo (SL) for the joy that they brought to our work and their patience with the bumps in the road.

Sabina Leonelli
Niccolò Tempini

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
October 2019

Contents

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.

Part I
Origins: Data Collection, Preparation and Reporting

Material Origins of a Data Journey in Ocean Science: How Sampling and Scaffolding Shape Data Practices

Gregor Halfmann

Abstract This chapter discusses the epistemological relevance of material interactions during the early stages of a data journey. It shows that processes taking place when research technology makes physical contact with the objects targeted in research endeavours shape the subsequent data journeys and the practices of creating scientific knowledge. The chapter is based on a case study of ecological monitoring in ocean sciences and zooms in on the practice of sampling the oceans' ecosystems with mechanical sampling devices that are towed regularly by commercial ships. I propose an understanding of materiality as the integration of physical matter from various sources so as to constitute a new entity, in this case a research sample containing plankton organisms. The material integration is followed by material continuity, the preservation of the sample throughout several if not all stages of the research process without a change of medium. This two-fold understanding is an attempt to ground the notion of "materiality" epistemologically rather than ontologically. As shown with empirical examples, material interactions are the origin of resistances or challenges which unfold throughout the research process as scientists intend to create knowledge by manipulating and analysing physical objects. The scientific practices are shaped by investigating, resolving, and working around these challenges.

1 Introduction

This chapter tracks physical interactions during the creation of research samples and discusses their epistemological significance. On the basis of a case study in ocean science, I argue that interactions between materials of the research technology and of the natural systems studied by scientists shape practices of creating and using scientific data; scientists deliberately study material interactions in order to account for uncertainties and to maintain commensurability of data that have been created decades apart. Understanding the epistemological significance of "materiality" in scientific practices is thus important for studies of data journeys.

G. Halfmann (✉)
Department of Sociology, Philosophy and Anthropology, University of Exeter, Exeter, UK

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_2

A variety of studies in the philosophy and sociology of science are concerned with the material nature of scientific objects and practices. Inducing a 'clash of materials' (Rheinberger 2011: 344) between biological entities and research technologies is central to many experimental practices in the life sciences. Such a clash may lead to the formation of objects, which are described as "material". However, a wide range of objects with fundamentally different formation processes and physical characteristics are used in the life sciences, for example anatomical preparations (Rheinberger 2015), model organisms (Ankeny and Leonelli 2011), or "material models" in the form of species collections in museums (Griesemer 1990). What "materiality" implies for knowledge production has been elaborated by scholars in some cases, showing that material interactions and knowledge production processes are often intertwined, but in a variety of ways.[1] This chapter complements these accounts by tracking the epistemological impacts of material interactions at selected stages of the formation and processing of research samples.

While many scholars have focussed on specific kinds of material objects or material aspects of their case studies, the terms "material" and "materiality" tend to remain rather loosely defined. Quite often, it seems that "material" is used to signal difference or opposition to other classes of objects or processes, which may be labelled "non-material", "virtual", "theoretical", "mathematical", "ideational", or the like. The opposition seems to bear on differences in an entity's physical constitution, stability, or tangibility, but also relates to its ontological status: mathematical theories or ideas certainly differ ontologically from a sampled biological species.

Debates over the meaning of "materiality" have ensued in some cases; for example, Morgan (2003) and Parker (2009) debate how to understand "materiality" with respect to scientific experiments. Parker (2009: 492–93) criticises that computer simulations are not seen as material experiments by many; she further suggests that the emphasis on "stuff" may be misplaced and that epistemologically, the behaviour of a system is more relevant than its ontological characteristics. In science and technology studies, the meaning of materiality has been discussed in relation to a growing interest in ontology; Woolgar and Lezaun (2013: 326) argue that characteristics that may qualify an object as "material" should be treated as practical achievements and "materiality" should therefore be understood as an 'upshot of practices' of a certain kind. These examples show that materiality in scientific practices deserves deeper scholarly consideration; a closer study of materiality may provide classifications involving "material", "non-material", or other types of entities with crucial context and a more solid grounding.

In this chapter, I propose an understanding of materiality as the integration of physical matter from various sources so as to constitute a new entity; the material integration is followed by the preservation of the entity throughout several if not all

[1] For example, the materiality of anatomical preparations results in an 'indexicality' of the object that points to itself rather than representing something else (Rheinberger 2015: 323); standardised material characteristics of model organisms make them usable as 'genetic tools' (Ankeny and Leonelli 2011: 316); the materiality of species collections provide an epistemological robustness to potential changes of theoretical perspective (Griesemer 1990: 83).

stages of the epistemic process without a change of medium. Material integration and material continuity are a two-fold characteristic applicable to objects that scientists create and use as well as to scientific practices. This understanding is an attempt to ground the notion of "materiality" epistemologically rather than ontologically. Empirical examples presented in this chapter show that material interactions are the origin of resistances or challenges, which unfold throughout the research process as scientists intend to create knowledge by manipulating and analysing physical objects; scientific practices are shaped by investigating, resolving, and working around these challenges.

Material integration appears as a characteristic that is applicable to virtually any entity considering formational processes on a biological or chemical level. However, in combination with material continuity, my understanding of materiality leaves aside data practices that involve "jumps" to an entirely different medium. Understanding materiality in research processes requires a focus on preservation and overlaps between different stages of epistemic practices, which can be likened to scholarly accounts of biological reproduction as I discuss later in this chapter.

By focusing on sampling and subsequent research practices, my chapter zooms in—as the title indicates—on the "origins" or the very early stages of a data journey. The beginning of a journey is not necessarily the moment, in which things move physically (or virtually) for the first time. Many would argue that a personal journey begins with thorough planning and smart packing; many choices made at this stage—which route to travel, which shoes to wear—depend on material aspects and conditions such as terrain or expected weather conditions. These conditions create challenges, which shape the actual movement and influence the journey's outcome. The journey, as an unfolding process or development, is enabled, facilitated, or "scaffolded" by these choices and by the artefacts, infrastructures, and agents a traveller has decided to utilise, for example boots, maps, or travel agents. This chapter is not about the data journey per se, but about early stages of an epistemic process; I use the plural form "origins" to account for the difficulty of pointing at one distinct moment, at which the journey begins.[2] A great deal of thinking, planning, and preparing is necessary for data (and for persons) to travel successfully (Leonelli 2016: 40, Learning from Data Journeys); the origins of these preparations, that is the processes and conditions that cause or provoke certain preparatory measures, are scattered across various domains,[3] including, as I intend to show in this chapter, material interactions at the sampling stage.

[2] I use the term "origin" with caution, in particular in relation to material objects; Rheinberger (2011: 338–9) writes that with respect to "traces", which are 'material manifestations' in experimentation before they are turned into representations, an origin does not exist and has never existed. With "origins of a data journey", I intend to highlight a number of processes leading up to the creation of data and the data journey, without implying that a concrete origin in space and time is tangible.

[3] The institutional context of research or the history of a research field, from which research activity is inspired and research methods are passed on, are examples of other domains that introduce restrictions on data practices.

Various stages of processing and manipulation of physical objects are strongly pronounced in my case study, the Continuous Plankton Recorder (CPR) Survey. Research samples containing marine organisms are created by deployment of mechanical sampling devices on commercial ships crossing the oceans. Samples are then analysed in four distinct steps in a laboratory in Plymouth, UK; these include microscopic identification and counting of hundreds of different taxa ranging from single-celled phytoplankton to zooplankton organisms measuring several millimetres. All samples are archived for potential re-analysis in the future. I illustrate with several examples in this chapter that material interactions between the mechanical sampling device and marine organisms require specific practices, which "scaffold" the creation and the interpretation of scientific data.

I understand scaffolding as dynamic structures of conceptualisations, practices, theories, technologies, or personal relationships, which are applied to entities in order to facilitate the development of specific capacities or skills. Wimsatt and Griesemer (2007) have coined the concept of scaffolding in relation to the development of culture and it has since been applied to various domains, including scientific practice. A rich collection of essays (Caporael et al. 2014b) demonstrates the applicability in three very broad domains—evolution, culture, and cognition—and encourages scholars to analyse their own work in terms of scaffolding. An example of its applicability in science is Wylie (2016: 1), who explains how 'interpretive scaffolding' is used in archaeology to determine how material traces of the past can be used as evidence; Wylie points to epistemological consequences of scaffolding by arguing that scaffolding is always provisional and new ways of data interpretation are capable of calling assumptions based on an established scaffold into question.

As the following empirical sections show, the CPR Survey is grounded in the analysis of physical objects probably as much as archaeology; yet, my case is quite different, because the same type of evidence—physical samples containing marine organisms—is created repeatedly over multiple decades. Scientists must then be able to compare old data with new data, which is a common challenge in environmental sciences that study long-term changes of natural systems (Edwards 2010). Besides discussing the material origins of scientific data, this chapter illustrates how the usability of data from different decades is scaffolded by implementing data practices that preserve methodological continuity.

After introducing the CPR Survey and tracking some material interactions and their epistemological implications, I discuss my understanding of "materiality" and elaborate how the material origins of scientific data require different forms of scaffolding and thereby shape data practices.

2 The Continuous Plankton Recorder Survey

The CPR Survey is an ongoing, long-term research programme run by the Sir Alister Hardy Foundation for Ocean Science (SAHFOS) from Plymouth, UK, since 1990 until SAHFOS merged with the Marine Biological Association of the UK (MBA) in

2018.[4] However, the survey itself is much older. The CPR was invented by fisheries ecologist Alister Hardy in the 1920s for the purpose of monitoring zooplankton, the key food source of larval fish (McQuatters-Gollop et al. 2015: 2). The design of the CPR and the steps of sample analysis were developed experimentally in a 'pioneer period' in the 1930s and the early 1940s (Reid et al. 2003: 130). Since the 1950s, the basic methods of sampling and analysis have remained unchanged (Reid et al. 2003: 131–32). With datasets covering more than 70 years, the CPR Survey is one of the longest running time series in environmental and marine science (McQuatters-Gollop et al. 2015: 2). The methodological stability is one of the most important aspects of the CPR Survey; it is vital for its reputation and prestige in the scientific community, but it introduces constraints to scientific practice, as the survey's lab manager David Johns explains:

> "The whole idea is that you keep the methodology the same. You don't want to make any mistakes with methodology, it has got to be the same. We pride ourselves on our 70-year time series, that's what we want." (DR0934: 5)

The CPR Survey has a long and eventful history; it was close to shut down in the 1980s, when long-term marine monitoring programmes in Europe were terminated at an alarming rate (Duarte et al. 1992).[5] Unlike many other programmes, the projected closing of the CPR Survey led to an international initiative strongly supported by the International Council for the Exploration of the Sea (ICES) and the Intergovernmental Oceanographic Commission (IOC) of UNESCO; a rescue fund was put together and established SAHFOS as a charity organisation in 1990 (Reid et al. 2003). SAHFOS' core work was 'the running and safeguarding' of the CPR Survey, according to its former Director (Owens 2015: 2). Running the survey consists of producing data related to plankton distributions from the analysis of samples, which are created through the deployment of CPRs. In addition to this core activity, SAHFOS increasingly engaged in 'ancillary activities and associated science' (Owens 2015: 2).[6]

A CPR is a mechanical filtering device that is towed by commercial ships on their regular shipping routes. Bands of silk inside the CPR filter the seawater and are processed into individual samples measuring around ten by ten centimetres. As of summer 2017 more than 5 million nautical miles have been sampled with CPRs in total and more than 250,000 samples have been analysed.[7] The CPR Survey oper-

[4] At the time of my research, the CPR Survey was still conducted by SAHFOS, and the name therefore appears throughout the chapter and my references. Since April 2018 the CPR Survey is officially run by the MBA and the name "SAHFOS" has now largely disappeared from websites and official statements related to the CPR Survey.

[5] At that time, long-term ecological monitoring 'was considered weak science, akin to stamp collecting' (Reid et al. 2003: 141); around 40% of European monitoring programmes were shut down in the late 1980s (Duarte et al. 1992).

[6] The survey's staff members are involved in the development and testing of new technology, in policy-driven work, or in education and outreach. Several research fellows conduct research in environmental change, molecular ecology, marine biodiversity, sustainable resources, and health and well-being of marine food sources (SAHFOS 2015).

[7] <https://www.sahfos.ac.uk/> [accessed 26 June 2017].

ates mainly in the North Atlantic and the North Sea, where most of the circa 25 regular towing routes are located. All samples are archived and stored in Plymouth for potential re-analysis in the future. Research based on CPR data sets has contributed significantly to the understanding of spatio-temporal dynamics of oceanic plankton and their responses to anthropogenic pressures and climate variability. The data are also used to inform UK and European marine policy-making and management of the seas (McQuatters-Gollop et al. 2015: 2).

In today's ocean science landscape, the CPR Survey is one of the oldest, yet only one of many projects that engage people without scientific credentials or institutional affiliations in sampling or data creation. To meet the economic challenges of sampling the world's oceans on increasingly finer spatial scales and with temporal regularity, a growing number of projects take advantage of recreational and professional seafarers, who regularly interact with the oceans. Picking up the current wave of citizen science and fuelled by technological innovation, marine science is often seen as a prime example of scientific fields with high potential for contributions by citizen scientists (Lauro et al. 2014). The CPR Survey does not refer to itself officially as "citizen science", although a wide range of non-scientists volunteer to make the survey possible. Among them are the captains, chief officers, bosuns, and crews aboard ships, but also ship owners and managers, stevedores, terminal managers, heavy cargo operators, and engineering companies (DR1960: 6).[8] The collaboration is crucial for setting up a ship for towing CPRs and for proper handling and transportation of boxed CPRs in high security areas in the ports' container terminals. For each ship and each tow, the survey relies on a number of volunteers, who make sure that a CPR arrives at the right ship at the right time. The collaborative practice of the CPR Survey has epistemic implications in its own right; most importantly, the geographical scope of the sampling and the CPR data depends on the locations of frequented shipping routes. The social dimension of the CPR Survey, in which research culture meets seafaring culture, offers opportunities for sociological and epistemological research. This chapter, however, focuses on the epistemology of the CPR Survey's material dimension.

3 Material Interactions and their Epistemological Implications

The following sub-sections describe two examples of material processes within the CPR Survey and their epistemological implications. These implications become manifest in data practices such as methods of creating data by sample analysis, but also in the outcomes of those practices, for example in databases and publications.

[8] SAHFOS often used the term "volunteers" to refer to the non-scientists involved in the survey. There is no formal contract with the non-scientists, except for the engineering companies who are commissioned to install davits or blocks on the ships that enable towing of a CPR. The shipping crews, but not the companies or ship owners, are compensated with £60 per tow (DR1960: 6).

The two processes are the deformation of plankton organisms during sampling and interactions between the silk and the organisms.

3.1 Deformation of Plankton Organisms and Identification

A CPR is a steel device that is shaped similar to a bobsleigh and weighs around 90 kg. When a CPR tow is scheduled to begin, crew members use the ship's winch to lower the CPR into the sea. SAHFOS emphasises that the sampling is never to interfere with the ship's normal business; a ship thus never stops or slows down for the deployment of a CPR. The steel body hits the water at up to 20 knots, putting significant tension on the steel wire, the body, and the internal mechanics of the CPR.[9] The wire is paid out until a coloured mark settles on the sea surface, indicating that the CPR has reached the desired depth between seven and 10 m. The pointed nose of the CPR has a small opening of around 1.5 cm^2, through which seawater enters a tunnel inside the CPR that leads to the filtering silk (SAHFOS 2016: 18; Reid et al. 2003: 126). The tunnel widens, so that the water pressure and the speed of flow inside the tunnel reduce significantly. A layer of silk (the filtering silk) spans across the tunnel, acts as a filter and retains a share of organisms and materials that have entered the tunnel. While the CPR is being towed, a propeller attached to the external body drives a mechanism that pulls the silk continuously across the flow of water. The silk that has crossed the tunnel is met by a second layer of silk (the covering silk), which is drawn by the same mechanism. The covering silk goes on top of the filtering silk, so that the organisms are held between the two layers.[10] The silk rolls are drawn together into a closed chamber filled with a formalin solution. The organisms cannot survive this process, but the formalin prevents the decay of their bodies.

Plankton organisms often get damaged and deformed during the sampling process. They may knock against the steel walls of the CPR or against other organisms that are already on the silk.[11] If towed through a plankton bloom, areas of the silk can actually get clogged with organisms, which affects the volume of filtered sea water (Hunt and Hosie 2006). The biggest cause of deformation is, however, the sandwiching of organisms between the two silk layers. With regard to some of the larger zooplankton species,[12] the survey's lab manager David Johns explains that "the stuff

[9] In a video of a CPR deployment, the device jumps on the sea surface for several seconds before submerging. When the CPR is hauled in, it sometimes smashes against the ship's hull strong enough for the steel body to be damaged and require refitting in the survey's workshop (DR1960).

[10] Two bands of silk are marked, cut, folded, rolled up, and placed inside the internal cassette by hand before a CPR is deployed. A metre of silk covers around one hundred nautical miles, so up two five metres of silk are rolled up for each of the two silk rolls.

[11] This cause of deformation is alleviated to some degree by the widening of the tunnel and the reduction of flow speed inside the CPR by a factor of around 1/30 (Batten et al. 2003: 196; DR2901: 2).

[12] Only zooplankton species larger than two millimetres are identified and counted the way described here. Smaller plankton, including single-celled phytoplankton, are identified with up to 625x magnification; Richardson et al. (2006: 35).

is squashed" and "it is very, very flat", when it arrives in Plymouth (DR0934: 19). The organisms thus look very different under a microscope in the survey's lab than out in the ocean or in taxonomic reference literature; Johns explains how the altered appearance by deformation affects the identification process:

> "Textbooks are obviously really useful, but it is not the same as looking down and actually seeing a physical specimen there. … They do look quite different, so you need to manipulate the organism." (DR0934: 18–19).

In order to be manipulated, turned around, and viewed from different angles, the zooplankton organisms are manually scraped off the silk and placed into a Bogorov tray under a different microscope for identification and counting.[13] Johns explains:

> "It is just so much easier to identify them. You can't do it on the silk very easily. It is so much easier, you take them off, put them into that tray, add some fluid and then you can manipulate them easily, flip them around. Because a lot of them, depending on how they are lying, they can hide their identification features, so you need to kind of manipulate them 360." (DR0533: 6)

The deformation during sampling and the way some organisms—especially those with spiny body features—are caught up in the silk requires manipulation and removal of organisms in order to create data. In this stage of the analysis, which is called the "zooplankton eyecount", all organisms larger than two millimetres are taken off the silk for identification and counting and are put back onto the silk afterwards. Data are created by counting different species or taxonomic groups and recording the result with tally marks in a hand-written notebook right next to the microscope.[14]

The organisms' altered appearance also requires the sample analysts to have specific identification skills and experience. New analysts go through a training phase, which lasts several months until they are allowed to work on samples even from the survey's most frequent sampling routes all by themselves. Samples from areas that are not sampled as frequently as the North Atlantic and the North Sea can be particularly challenging, because the encountered species and the ecology are very different. Some analysts have therefore specialised in samples from certain areas after years of practice and interacting with other analysts in the lab (DR0533: 10). Johns explains that "probably most of [the training] is informal and on-the-job stuff" (DR0934: 18), due to the specific characteristics of the CPR samples; the skills and experience are best acquired in practice and in cooperation with experienced analysts. One of the experienced sample analysts describes the interaction in the lab, by which they gain expertise:

---

[13] Manipulation and turning around of organisms is also necessary, because some species are difficult to distinguish; for example *calanus finmarchicus* and *calanus helgolandicus*, two of the most important zooplankton species in the North Atlantic and the North Sea, look very similar and are identified primarily by the shape of their fifth pair of swimming legs; Richardson et al. (2006: 47).

[14] The data in the notebooks are later entered into the digital database manually by two sample analysts together in order to avoid transcription mistakes and to notice unusual looking results that might indicate an error in identification or counting (DR0533: 2).

> "We are always looking at each other's samples all the time. It's not that a day goes past where you are not going to go a look at someone else's stuff …" (DR8112: 10)

The removal of materials from the sample and the ways of acquiring expertise and experience are examples of how data practices are shaped by material interactions at the sampling stage.

In 2011, SAHFOS published the *Fish larvae atlas of the NE Atlantic* (Edwards et al. 2011), which illustrates how deformation during sampling constrains exactly what kinds of data can possibly be created during sample analysis. The atlas covers geographical distributions of fish larvae of nine different taxa for the years 1948–2005. More than 10,000 archived silk samples have been re-analysed with new molecular methods, because fish larvae are not routinely identified in the microscopic analysis:

> Due to the size of the fish larvae and the sampling method, they can often be damaged and identification to species level is not always possible using traditional microscopic methods. (Edwards et al. 2011: 2)

As the fish larvae are often too damaged for visual identification, they are only counted and recorded in the survey's database as one taxonomic group. The database's content and the knowledge of the ocean ecosystem are thus shaped by material interactions that occur during sampling.

3.2 Silk Specifications and Quantification

Albeit having changed silk suppliers several times throughout the history of the survey, silk with identical specifications has been used for sampling since the beginning of the CPR Survey. The silk bands have a mesh size of around 270 µm and are quality controlled and prepared in a standardised way, which includes marking, stamping, folding, cutting, and putting the silk onto a roll that is going to be placed inside the CPR.[15] Smooth fabrics such as nylon and much finer mesh sizes are typically used in plankton science today. The 270 µm is indeed large compared to the size of some species that are routinely recorded, as lab manager Johns explains:

> "We had people saying that there is no way that we can see coccolithophores, they said 'no, it is going to go straight through your mesh, because they are only ten microns.' But they do stay there, so we took photos and we published some of it and say 'actually, we can see these.'" (DR0934: 6)

Coccolithophores are a group of unicellular, eukaryotic phytoplankton species, which are around a magnitude smaller than the gap between the silk threads; yet, a constant portion of those species are retained. That is because the silk has a certain

[15] Marking and stamping is required for calculating the cutting points after each tow under consideration of the ship's average speed; each sample is intended to correspond to ten nautical miles of a tow, but the length of silk pulled by the mechanisms over that distance depends on how fast the ship has sailed.

roughness and the individual threads are spinous, so that small organisms stick to them; the silk also has a leno weave, which has two twisted threads going in one direction and one thread in the other direction, whereas most nylon fabrics used for filtering are heat-fused so that the junctions are smooth. Phytoplankton can thus get caught in the tiny gaps between the twisted silk threads (Richardson et al. 2006: 61; DR0934: 6).

Some interactions between certain types of organisms and the sampling technology are in fact multi-layered, because the presence of larger organisms also affects the efficiency, at which small phytoplankton are retained.[16] Large zooplankton may have spiny body features, on which smaller organisms may get caught. As a growing amount of plankton covers the silk, the filter efficiency tends to increase:

> As more and more organisms are filtered onto the mesh the open apertures are progressively clogged and reduce the effective mesh size. So as more large organisms are retained, smaller organisms, which at the start of the sampling would have been extruded, will be retained progressively more effectively (Batten et al. 2003: 206).

In general, a significant amount of small phytoplankton still flow through the silk and return into the open ocean, while most of the large zooplankton is retained. The material processes are complex and have led to experimental investigations regarding the effects of clogging with different mesh sizes (Hays 1994; Hunt and Hosie 2006). Some gelatinous plankton species can particularly enhance clogging (Richardson et al. 2006: 61). Batten et al. (2003: 206) explain the challenge posed by such interactions between organisms of different sizes and texture and the silk:

> The effect is hard to quantify since the ambient concentrations of organisms (needed to determine the true proportion retained) will never be known for a specific patch of sea water at a specific time.

The materiality of the silk and the plankton organisms thus have implications that relate to the quantities of specific organisms on the silk, which are represented in the data created by the analysts. More specific, the data created by sample analysis hardly reflect the total numbers of plankton organisms at a specific space and time in the ocean. Richardson et al. (2006: 61) state that 'there is increasing evidence that the CPR substantially underestimates absolute numbers'. The CPR data are thus often referred to as "semi-quantitative". This characteristic of the CPR Survey, which is a result of material processes, does not mean that data are false or useless; however, the materiality shapes the way data are used by scientists:

> Notwithstanding the semi-quantitative nature of CPR sampling, there is considerable evidence that it captures a roughly consistent fraction of the in situ abundance of each taxon and thus reflects the major patterns observed in the plankton. (Richardson et al. 2006: 61)

The semi-quantitative character of the data could be viewed as a shortcoming; however, as Johns explains, the consistency of the sampling is valued higher than potential increases of precision:

[16] The distinction I make between small phytoplankton and large zooplankton is a simplification and does not reflect the spectrum of shapes and sizes of the organisms on a silk sample.

> "We want to keep that consistent time series. And there are a lot of potential sort of foibles in the dataset. But the fact that it has always been done in the same way … You get lots of people who, it's not an accuse, but who would say 'well, you under-count certain things'. Well yeah, we do, but they have been consistently under-counted for sixty years. So you can just ignore the abundance values and just look at the trend to see what is happening. So yeah, if you were starting [the survey] from scratch, you would do it completely differently." (DR0533: 4)

Other "foibles"[17] result, for example, from the analysis of phytoplankton and zooplankton smaller than two millimetres, for which each sample is sub-sampled. In case of phytoplankton, only around 1/10,000th of a silk area is looked at under the microscope. The analysts further use a number of fixed abundance categories, which are subsequently converted into estimates for the quantity of organisms of a specific taxon on a sample. Richardson et al. (2006: 63) explain that 'abundance estimates from individual plankton samples are inherently imprecise because of variable zooplankton behaviour such as diel vertical migration and local weather conditions that can concentrate or disperse fine-scale patches (Robertson 1968), as well as the "broad-brush" counting procedures.'

As CPR data do not reflect total quantities of organisms in the ocean, the data are usually not expressed in units such as organisms per cubic metre of sea water; instead, they remain expressed in the unit 'numbers per sample', which is an estimate derived from the hand-written records (Richardson et al. 2006: 62).

Batten et al. (2016) is a localised study in fisheries ecology and an example of how semi-quantitative data are used. The study uses indices calculated from CPR data to explain variability of the Prince William Sound herring's first year growth. Annual abundance anomalies for groups such as large zooplankton, small zooplankton, or diatoms were calculated and then correlated with estimates of herring growth rates calculated from scale size measurements. Figures in the study use 'organisms (zooplankton) or cells (diatoms) per Continuous Plankton Recorder (CPR) sample' as a unit (Batten et al. 2016: 428); the authors also explain the relation between the silk's mesh size and filter efficiency, and clarify what their data may represent:

> Only an undefined proportion of the phytoplankton and microzooplankton community … is enumerated by CPR sample analysis. The data shown here then do not necessarily indicate whether more or less chlorophyll or ciliates were available, but as the CPR is an internally-consistent sampler, they do indicate when relatively more, or less, of the large diatoms and hard-shelled microzooplankton were present and available as a food source. (Batten et al. 2016: 429)

The specifications of the used silk and material interactions at the sampling stage between the silk and plankton organisms thus affect how many organisms end up on the silk, the quantities subsequently recorded by analysts in their notebooks, and how the data can be used to create knowledge of the ocean ecosystem.

---

[17] Johns seemed to be searching for the right term before saying "foibles". However, the term seems very fitting, as it refers to a 'minor flaw or shortcoming', but not as a complete fault or failure. Persons or things with foibles are still valued and useful, despite minor shortcomings; <https://www.merriam-webster.com/dictionary/foible> [accessed 24 August 2017].

4 Material Integration and Continuity

The previous sections illustrate how many of the sample's material characteristics that restrict how the object can be manipulated and used originate when the CPR is in the water. By contrast, the materials themselves, the silk, the steel, and the organisms, have their respective origins in factories, in plankton life cycles, or even further back. In the course of a CPR tow, physical parts of both the sampling technology and the ocean ecosystem not only "clash" against each other; they become integrated. A variety of effects during integration—some of which are described above—lead to the formation of a novel object, the silk roll, which is later processed into individual samples.

Material integration is a constitutive phase and can be regarded as the realisation of an 'apparatus-world complex', a term used by philosopher Rom Harré (2003: 28–31), who explains that a technical device is capable of being 'integrated into a unitary entity by fusion with nature' (Harré 2003: 28); furthermore, 'the apparatus and the neighbouring part of the world in which it is embedded constitute one thing' (Harré 2003: 29).[18] The point is that the material integration realised in the CPR Survey is a constellation that results in the constitution of a new research object with properties that have been shaped during integration by material interactions.[19] Both the plankton organisms and the silk are physically transformed during the integration: the organisms are immediately deformed and the silk assumes a different colour. The silk as well as the organisms are constitutive parts of the newly formed object and a research sample in the CPR Survey could not exist without either one.

My understanding of "integration" as the constitution of a new research object resonates with Tempini's (this volume a, b) account of assembling and integrating data from various sources to create new digital datasets. There is obviously a strong contrast between a sample integrated physically from silk, ocean water, and marine organisms and digital data that have been integrated from various datasets by computational commands; however, epistemologically, both integration procedures are geared towards forming objects that are analysable and meaningful in specific epistemic contexts.

In my case, it is important that the very materials that have been integrated are preserved throughout various stages of transportation, unloading, cutting, analysis,

[18] Rheinberger's (2010: 217–218) description of an 'intersection' as a 'surface', 'plane', or 'point of contact' between a technical device and the object studied by scientists is similar to Harré's apparatus-world complex; according to Rheinberger, an interface is a 'fertile analytical constellation', which certainly resonates with the idea that new entities are "born" during sampling.

[19] While this is not describing a case of reproduction, my view of silk rolls as novel objects, from which individual samples are created, is inspired by Griesemer's (2014: 39–40) view of hybrids as individuals in biological reproduction; individuality is not an intrinsic property of certain objects, but can be understood as designating a relation between attention, abilities, and interest of the person tracking a phenomenon and properties, relations, behaviours, and activities attributed to what is being tracked. My account tracks materiality and contrasts with a view of the sample as a mere assembly of materials which could easily be disassembled to its original components.

and long-term storage. In the CPR Survey, material continuity is achieved between the silk roll's formation process out in the oceans and the object that is placed under a microscope and eventually archived in Plymouth. In his account of biological reproduction, Griesemer (2014: 26–27) emphasises the notion of 'material continuity' and material 'overlap' between parent and offspring when 'organized material parts' are transferred between the two; form or information are transferred materially and not by any kind of impression or translation to a different medium. Although being pressed severely into the silk, the plankton material usually remains sufficiently organised for the sample analyst to identify and count the organisms using specific tools and methods of manipulation.

Rheinberger (2015: 323–325) asserts preparations a materiality and durability similar to the research samples in the CPR Survey: Preparations 'participate in, are part of, the very materiality of the object under scrutiny'; their 'configuration' is expressed in physical, biological, and chemical properties (Rheinberger 2015: 323). A CPR silk sample has assumed a specific configuration that makes it analysable and the configuration is preserved by material continuity.[20] It is important, however, that "preservation" and "continuity" are not intended to imply that samples are immutable or "frozen": Due to the formalin, the organisms' green colour fades over time[21]; their spatial arrangement on the silk changes when plankton are removed and put back onto the silk during the zooplankton eyecount; and samples in the archive might get contaminated and slowly decay, impeding the ability to perform a re-analysis. Material continuity is an absence of "jumps" from one medium to another, as in the hand-written recording of plankton counts or the digitisation of hand-written notes.[22]

Material integration and material continuity frame an understanding of "materiality" that—despite being based on the physicality of objects and practices—emphasises the epistemological significance of material objects over characteristics that categorise objects ontologically. The next section discusses exactly how materiality shapes scientific practices.

[20] Rheinberger (2015: 323) further claims that 'preparations are renderings, not representations' with a 'particular indexicality' that points to themselves and not to something that is represented by the preparation. The material characteristics of the silk samples seem to point primarily to the processes involved in their formation; additionally, the bias between the number of organisms on the sample and plankton distributions in the ocean poses questions regarding the samples' potential use as representations. These issues relating to scientific representation require deeper discussion elsewhere.

[21] The survey derives a set of data from the colour of returning silk samples, as sample colour is used as an indicator of relative phytoplankton biomass in the geographical area of the tow. Due to fading of the colour, the assessment is performed when the silk roll is cut into individual samples and can only be performed once.

[22] The lack of translation to another medium is another reason why considering samples as straightup representations is problematic (see note 20); a sample is a product of continuity starting with the fusion of materials in the oceans, and not by intentionally writing or imprinting information onto a medium.

5 Scaffolding Sample Analysis and the Creation of Knowledge

A CPR sample's physical properties require specific epistemic practices that are applied to the sample or to the data that have resulted from the analysis. The examples described in this chapter are the removal of plankton organisms from the silk, on-the-job transfer of identification skills, and the consideration of relative quantities and trends instead of total quantities. Regarding the removal of large zooplankton from the silk, the scraping together of organisms, the Bogorov tray, the additional microscope, and the manipulation of organisms are artefacts and practices, which scaffold the identification and counting of the organisms. Without this step, the identification would hardly be possible, be much more difficult, or at the very least take much longer to perform. The plankton analyst faces what Caporael et al. (2014a: 15) call a 'productive resistance or challenge', which can be overcome through scaffolding. The aided identification results in a growing volume of scientific data created from an individual sample, and eventually in growth of the database and of the data's interpretive scope. Besides development and maintenance, growth, as a change of size or status without change of organisation, is a plausible function or goal of scaffolding procedures, as Caporael et al. (2014a: 15–16) remark. Similar to a scaffold that is removed from a building after construction work has finished, the additional tray is removed, the organisms are placed back onto the silk and evenly spread out. Except for an altered distribution of the larger organisms, which has never been recorded in any way before the removal of organisms, no visual characteristic of the sample indicates that the scaffolding procedure and the identification of large zooplankton have been performed.

The second example, the on-the-job training of analysts, is a scaffold that develops the skills and capacities of the laboratory staff. Frequent interactions between experienced analysts and new staff members scaffold the acquisition of identification skills, which could hardly be learned without the informal exchanges. Challenges and resistance are caused by the deformed appearances of the organisms, the specific composition of various species on samples depending on the region they are from, or any kind of unusual or surprising encounter on a sample. This type of on-the-job development of capacities and resolving of challenges is an example of what is called 'developmental agent scaffolding' by Caporael et al. (2014a: 15), which is characterised by cooperation and response between agents and their targets rather than just by application of an artefact or structure. The scaffolding in this example is anything but permanent, as people in the lab are not constantly assisting each other; it is utilised as needed, either if new analysts receive basic training, if a special expertise is going to be acquired, or if an analyst is simply in doubt about an organism's taxonomic identity.

The third example of scaffolding relates to the interpretation of the semi-quantitative data created by sample analysis. Although the distribution of organisms on a sample is not representative of the species' total quantities in the ocean, researchers are capable of creating knowledge about the oceans with the data. The use of the data is scaffolded by multiple studies carried out throughout the history

of the CPR Survey into the technical details and uncertainties introduced by material interactions such as clogging of the silk. This is how the survey has accumulated 'considerable evidence' (Richardson et al. 2006: 61) that CPRs filter each taxon consistently and that the data reflect the patterns and trends of the plankton in the ocean. SAHFOS has likewise conducted studies regarding the effects of different ship speeds: The average speed of ships has almost doubled since the 1950s and in general, not all ships tow CPRs at the same speed due to season, weather, or other restrictions (SAHFOS 2016: 19; Batten et al. 2003: 200–02).[23]

Knowledge and evidence accumulated from these studies scaffold long-term consistency of the sampling and data analysis methods; the consistency, in turn, scaffolds commensurability and comparability of data created decades apart. A wide range of knowledge claims about the ocean ecosystem, especially those based on averaged data, depend on this commensurability. Only because the methods of sampling and data creation have been maintained for multiple decades, the CPR data are as valuable and relevant for plankton science as they are today.

As Caporael et al. (2014a: 16) explain, 'maintenance seems more different from development than it really is'; in a dynamic system, 'maintenance sustains a steady state, that is, it preserves organization in the face of stress, deterioration, and change, so maintenance is a change operation' (Caporael et al. 2014a: 16). In the face of uncertainties, the inner consistency of the CPR Survey is maintained, although potential "foibles" (as the lab manager called them) may be maintained in the data as well. After decades of performing sampling and analysis the same way, the practice has become historically "entrenched" (Wimsatt 2014). However, the use CPR data still hinges on the abilities to evaluate the data's accuracy and potential bias; each study of the survey's materiality develops this ability. Along with the material interactions themselves, such scaffolds shape the data practices in my case.

Similar to other scaffolds, efforts aimed at understanding the materiality are expended on different time scales than the CPR Survey as a whole, because they are normally time-limited projects explicitly concerned with one detail or interaction. These studies are not completely invisible, as they are frequently published in scientific journals or referenced in publications using the data. In terms of scaffolding, however, this referencing seems more like a certificate that a development has happened or that a particular aspect of the survey is being maintained. The scaffolding itself, that is the actual practice aimed at development, has been removed, whereas the developed skill or capacity has been internalised.[24]

[23] The effects of the towing speed on the average depth and filter volume of the CPR are still not fully understood; experiments from 2015 showed greater depth with higher towing speeds, but earlier studies suggested a constant towing depth independent of speed (SAHFOS 2016: 19; Batten et al. 2003: 201–02). The average increase of speed from around 10 knots in the 1950s to around 20 knots today had a negative effect on the towing stability. By 1970 more and more CPRs were actually torn off and lost. As a consequence, a stronger and more flexible steel wire was introduced since 1976 (Batten et al. 2003: 199).

[24] "Internalisation" is also a characteristic of scaffolding; a capacity, a skill, or sometimes the entire scaffold may be internalised by the developed structure, so that it is not visible from the outside; the internalised scaffold (for example a new method, or new knowledge) may then become a stable platform for new scaffolding procedures (Wimsatt and Griesemer 2007: 245). In the CPR Survey,

6 Conclusion

My study of an example of long-term ecological monitoring in ocean science emphasises the importance of samples and material interactions during their formation for epistemic processes and data practices. Materials of the sampling device interact with materials of the research target in ways that require transient and dynamic scaffolding activities[25]; scientists apply specific practices and techniques to material objects in order to achieve results and progress that would not be realisable otherwise or only realisable with much more difficulty and under much higher economical costs. The continuity of methods, how scientific practice can remain unchanged in the context of historical developments, deserves particular emphasis and certainly offers opportunities for intriguing philosophical study. Without scaffolding the continuity of sampling and data practices, much of the data in my case study would hardly be usable at all to study long-term changes of the ocean ecosystem. Temporary scaffolds are necessary in order to keep an historically "entrenched" scientific method stable for decades and in order to learn about sources of uncertainties in the resulting data.

This chapter approaches the materiality of scientific objects by regarding it as the integration of physical parts from different sources into one novel entity and as the realisation of material continuity—a preservation of physical matter without any "jumps" to a different medium—throughout the epistemic process; this approach is not intended as a readily generalisable definition of the term "materiality". The aim of this chapter was to flesh out the epistemological relevance of material interactions by showing how such interactions between research technologies and research targets can shape data journeys.

References

Ankeny, Rachel A., and Sabina Leonelli. 2011. What's So Special About Model Organisms? *Studies in the History and Philosophy of Science* 42: 313–323.

Batten, Sonia D., et al. 2003. CPR sampling: The technical background, materials and methods, consistency and comparability. *Progress in Oceanography* 58: 193–215.

Batten, S.D., S. Moffitt, W.S. Pegau, and R. Campbell. 2016. Plankton Indices Explain Interannual Variability in Prince William Sound Herring First Year Growth. *Fisheries Oceanography* 25: 420–432.

Caporael, Linnda R., James R. Griesemer, and William C. Wimsatt. 2014a. Developing Scaffolds: An Introduction. In *Developing Scaffolds in Evolution, Culture, and Cognition*, ed. Linnda

any new capacity or knowledge about the survey itself may function as a platform for further development of capacities in the future.

[25] Using Pickering's (1993: 567) words, the structure of such transient and dynamic activities can be interpreted as 'activities in an evolving field of human and material agencies reciprocally engaged in the play of resistance and accommodation' or in other words, as a "mangle of practice".

R. Caporael, James R. Griesemer, and William C. Wimsatt, 1–20. Cambridge MA/London: The MIT Press.

———, eds. 2014b. *Developing Scaffolds in Evolution, Culture, and Cognition*, Vienna Series in Theoretical Biology. Cambridge, MA/London: The MIT Press.

Duarte, Carlos M., Just Cebrián, and Núria Marbá. 1992. Uncertainty of Detecting Sea Change. *Nature* 356: 190.

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

Edwards, M., et al. 2011. *Fish Larvae Atlas of the NE Atlantic. Results from the Continuous Plankton Recorder Survey 1948–2005*. Plymouth, UK: Sir Alister Hardy Foundation for Ocean Science.

Griesemer, James R. 1990. Material Models in Biology. In *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Volume 2: Symposia and Invited Papers, 79–93.

———. 2014. Reproduction and the Scaffolded Development of Hybrids. In *Developing Scaffolds in Evolution, Culture, and Cognition*, ed. Linnda R. Caporael, James R. Griesemer, and William C. Wimsatt, 23–55. Cambridge, MA/London: The MIT Press.

Harré, Rom. 2003. The Materiality of Instruments in a Metaphysics for Experiments. In *The Philosophy of Experimentation*, ed. Hans Radder, 19–38. Pittsburgh, PA: University of Pittsburgh Press.

Hays, Graeme C. 1994. Mesh Selection and Filtration Efficiency of the Continuous Plankton Recorder. *Journal of Plankton Research* (4): 403–412.

Hunt, Brian P.V., and Graham W. Hosie. 2006. Continuous Plankton Recorder Flow Rates Revisited: Clogging, Ship Speed and Flow Meter Design. *Journal of Plankton Research* 28: 847–855.

Lauro, Federico M., Svend Jacob Senstius, Jay Cullen, Russell Neches, Rachelle M. Jensen, et al. 2014. The Common Oceanographer: Crowdsourcing the Collection of Oceanographic Data. *PLoS Biol* 12: e1001947.

Leonelli, Sabina. 2016. *Data-centric biology: a philosophical study*. Chicago, IL: Chicago University Press.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

McQuatters-Gollop, Abigail, et al. 2015. The Continuous Plankton Recorder Survey: How Can Long-term Phytoplankton Datasets Contribute to the Assessment of Good Environmental Status? *Estuarine, Coastal and Shelf Science*.

Morgan, Mary S. 2003. Experiments Without Material Intervention: Model Experiments, Virtual Experiments, and Virtually Experiments. In *The Philosophy of Experimentation*, ed. Hans Radder, 216–235. Pittsburgh, PA: University of Pittsburgh Press.

Owens, Nicholas J.P. 2015. *'Director's Review', in 2014 Annual Report*. Plymouth, UK: Sir Alister Hardy Foundation for Ocean Science.

Parker, Wendy S. 2009. Does Matter Really Matter? Computer Simulations, Experiments, and Materiality. *Synthese* 169: 483–496.

Pickering, Andrew. 1993. The Mangle of Practice: Agency and Emergence in the Sociology of Science. *American Journal of Sociology* 99: 559–589.

Reid, P.C., J.M. Colebrook, J.B.L. Matthews, J. Aiken, and Continuous Plankton Recorder Team. 2003. The Continuous Plankton Recorder: Concepts and History, From Plankton Indicator to Undulating Recorders. *Progress in Oceanography* 58: 117–173.

Rheinberger, Hans-Jörg. 2010. *An Epistemology of the Concrete: Twentieth-Century Histories of Life*. Durham/London: Duke University Press.

———. 2011. Infra-Experimentality: From Traces to Data, from Data to Patterning Facts. *History of Science* 49: 337–348.

———. 2015. Preparations, Models, and Simulations. *History and Philosophy of the Life Sciences* 36: 321–334.

Richardson, A.J., A.W. Walne, A.W.G. John, T.D. Jonas, J.A. Lindley, D.W. Sims, D. Stevens, and M. Witt. 2006. Using Continuous Plankton Recorder Data. *Progress in Oceanography* 68: 27–74.

Sir Alister Hardy Foundation for Ocean Science. 2015. *2014 Annual Report*. Plymouth, UK: Sir Alister Hardy Foundation for Ocean Science.

———. 2016. *2015 Annual Report*. Plymouth, UK: Sir Alister Hardy Foundation for Ocean Science.

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.

Wimsatt, William C. 2014. Entrenchment and Scaffolding: An Architecture for a Theory of Cultural Change. In *Developing Scaffolds in Evolution, Culture, and Cognition*, ed. Linnda R. Caporael, James R. Griesemer, and William C. Wimsatt, 77–105. Cambridge, MA/London: The MIT Press.

Wimsatt, William C., and James R. Griesemer. 2007. Reproducing Entrenchments to Scaffold Culture: The Central Role of Development in Cultural Evolution. In *Integrating Evolution and Development: From Theory to Practice*, 227–323. Cambridge, MA: MIT Press.

Woolgar, Steve, and Javier Lezaun. 2013. The Wrong Bin Bag: A Turn to Ontology in Science and Technology Studies? *Social Studies of Science* 43: 321–340.

Wylie, Alison. 2016. How Archaeological Evidence Bites Back: Strategies for Putting Old Data to Work in New Ways. *Science, Technology, & Human Values* 42: 203–225.

———. this volume. Radiocarbon Dating in Archaeology: Triangulation and Traceability. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer..

Gregor Halfmann earned his Master of Science degree in Physical Oceanography at the University of Hamburg and a Master of Arts in History and Culture of Science and Technology at Technical University of Berlin. In 2014, he joined the "Epistemology of Data-Intensive Science" research project at Egenis, Centre for the Study of Life Sciences, University of Exeter, where he completed his PhD thesis in Philosophy of Science, titled "Seafarers, Silk, and Science: Oceanographic Data in the Making", which focuses on epistemological, material and social conditions of data practices in an empirical example of long-term monitoring of oceanic plankton ecology. He is now back in Germany working as a teacher.

What Data Get to Travel in High Energy Physics? The Construction of Data at the Large Hadron Collider

Koray Karaca

Abstract In present-day high-energy physics experiments, experimenters need to make various judgments in order to design automated data processing systems within the existing technical limitations. In this chapter, as a case study, I consider the automated data acquisition system used in the ATLAS experiment at the Large Hadron Collider (LHC) located at CERN, where the Higgs boson was discovered in 2012. I show that the design of this system involves both theoretical and experimental judgments each of which has different functions in the initial data journey through which usable data are constructed out of collision events detected by the ATLAS detector. I also explore what requirements the foregoing judgments impose on the LHC data in terms of usability, mobility and mutability. I argue that in present-day HEP experiments these aspects of data are distinct but related to each other due to the fact that they are subjected to some common requirements imposed by the theoretical and experimental judgments involved in the design of data acquisition systems.

1 Introduction

The introduction of computer technologies to experimental high-energy physics (HEP) experiments in the fifties and sixties resulted in the automation of data processing in HEP experiments (Galison 1997). Continuous advances in computer technologies have led to the ever-increasing automation of data processing in experimental HEP. This has made it possible to process increasingly large and complex data produced by increasingly more advanced particle detectors and colliders. As a result, experimental HEP has been progressively data intensive over the past 60 years, and this has been accompanied by important changes not only in terms of methods, techniques, and tools employed in HEP experiments (Franklin 2013; Gutsche et al. 2017), but also in terms of organizational structures (Boisot et al. 2011; Knorr-Cetina 1999) and authorship (Galison 2003) in experimental HEP collaborations.

K. Karaca (✉)
Department of Philosophy, University of Twente, Enschede, The Netherlands
e-mail: k.karaca@utwente.nl

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_3

The ATLAS and CMS experiments[1] currently running at the Large Hadron Collider (LHC) located at CERN represent the state of the art in automated data processing in HEP experiments, as the level of automation achieved in these experiments is unparalleled in previous HEP experiments. While automation enables processing unprecedently large and complex data in the foregoing LHC experiments, it greatly reduces the need for human intervention in data processing. However, automation does not diminish the role of human judgments in this process. As I will discuss in this chapter, experimenters at the LHC need to make various judgments to be able to design automated data processing systems within the existing technical limitations.[2] As a case study, I will examine the automated data acquisition system used in the ATLAS experiment. I will argue that the design of this system involves both theoretical and experimental judgments each of which has different functions in the automation of data processing in the ATLAS experiment. I will also explore what kinds of requirements the foregoing judgments impose on the LHC data in terms of usability, mobility and mutability, which are the general aspects of data in physical and biological sciences (Leonelli 2016).

In addressing the foregoing issues, I shall make use of the notion of *data journey*, which is a useful metaphor to characterize various processes that data undergo in experiments performed in physical and biological sciences (ibid.). In these experiments, data journeys start with the process of data acquisition. Some of the philosophical aspects of this process have already been discussed in the context of the LHC experiments (see, e.g., Morrison 2015; Beauchemin 2018; Karaca 2017, 2018), and also in other contexts in this volume. In a case study concerning ocean science, Gregor Halfmann (in this volume) discusses the initial stage of data acquisition where data is first produced. In a case study concerning astronomy, Götz Hoeppe (in this volume) discusses aspects of data acquisition concerning data interpretation. In this chapter, I will focus on the initial data journey in the ATLAS experiment that links the production of collision events at the LHC to the stage of data acquisition where *usable* data are constructed out of collision *events* detected by the LHC, prior to the stage of data analysis and modeling (Karaca 2018; Leonelli 2019; Boumans and Leonelli in this volume).

In scientific experimentation, data usability means the fitness of experimental data for its intended uses, namely data analysis and data modelling, which are aimed at serving the objectives of an experiment. In the context of present-day HEP experiments, the term *data* is used to refer to *collision events* produced by collider systems such as the LHC and detected by detector systems such as the ATLAS and CMS detectors. In the terminology of HEP, the term event denotes "the record of all the products from a given bunch crossing," (Ellis 2010, 6) which occurs when two beams of particles collide with each other inside the collider. In the ATLAS experiment, proton

[1] The names of these HEP experiments are derived from the ATLAS (A Toroidal LHC Apparatus) and CMS (Compact Muon Solenoid) detectors located at the LHC.

[2] The details of the design of the automated data processing systems used in the ATLAS and CMS experiments are explained in the technical design reports of these experiments, see ATLAS Collaboration 2003; CMS Collaboration 2002.

bunches, rather than individual protons, collide inside the LHC at a rate of approximately 40 million times per second. These recorded collision events, amounting to petabytes (=10^{25} bytes) of data, are then processed and finally digitally recorded on tapes in databases at CERN. I shall call the foregoing journey of the LHC data the *local data journey*, as opposed to the *global journey* that I take to refer to the journey of the LHC data concerning its dissemination to researchers located inside and outside CERN.

The plan of this chapter is as follows. In Sect. 2, I will discuss how the criteria for usable data are specified in the ATLAS experiment. Also, I will characterize the experimental strategy used to search for usable data in this experiment. In Sect. 3, I will examine the local data journey at the LHC and show how usable LHC data are constructed out of event fragments detected by the ATLAS detector. In the final section, I will argue that in the ATLAS experiment data mutability is required for data usability, and that the former is enabled by data mobility through the local data journey at the LHC. Furthermore, I will identify the judgments involved in the design of the ATLAS data acquisition system. I will argue that as a result of the requirements imposed by the foregoing judgments, usability, mutability, and mobility are related, though distinct, aspects of the LHC data during its local journey.

2 Selection Criteria and Search Strategy for Usable Data in the ATLAS Experiment

The ATLAS experiment at the LHC is a multi-purpose HEP experiment with two sets of objectives (ATLAS Collaboration 2003, Sect. 4): (1) to test the predictions of the present models of HEP concerning new particles, including the Higgs boson predicted by the Standard Model (SM)[3] of elementary particle physics and the particles, such as new heavy gauge bosons, superpartners and gravitons, predicted by the theoretical *models beyond the SM* (BSM models) that have been offered as possible extensions of the SM model, such as super-symmetric and extra-dimensional models (Ellis 2012); and (2) to search for *unforeseen* physics processes, i.e., those that have not been predicted by the present HEP models, including possible deviations from the SM at *low* energies. As I shall show in this section, the diversity of the objectives of the ATLAS experiment has a crucial bearing on what is considered usable data in this experiment, and also on the procedure through which this data is acquired.

The first set of objectives of the ATLAS experiment concerns a range of predictions concerning different kinds of heavy particles (including the SM Higgs boson) that are predicted to be produced at high energies, while its second set of objectives concerns unforeseen physics processes which might occur at both high and low ener-

[3] The SM consists of two different gauge theories; namely, the electroweak theory of the weak and electromagnetic interactions, and the theory of quantum chromo-dynamics which describes the strong interaction.

gies. This means that the collision events relevant to the first set of objectives of the ATLAS experiment are also relevant to its second set of objectives concerning the discovery of unforeseen physics processes that might occur at high energies. Therefore, the objectives of the ATLAS experiment require different, but partly overlapping, types of collision events to be acquired during the stage of data acquisition.

In the context of present-day HEP experiments, collision events that have the potential to serve the objectives of the experiment are often referred to as *interesting events*. In the case of the ATLAS experiment, the *signatures*[4] predicted by the SM for the Higgs boson are high transverse-momentum (p_T)[5] photons and leptons,[6] and the ones predicted by the BSM models for new particles beyond the SM, such as new heavy gauge bosons W' and Z' and supersymmetric particles, are high p_T single particles, namely photons and leptons, high p_T jets as well as high missing and total transverse energy (E_T).[7] The aforementioned high p_T and E_T types of signatures might be produced at the LHC as a result of the decay processes involving the Higgs boson and the aforementioned particles predicted by the BSM models. The same types of signatures might also be produced at the LHC as a result of some *unforeseen* physics processes occurring at high energies (i.e. approximately above 10 GeV). This means that the collision events containing high p_T and E_T types of signatures are relevant to both sets of objectives of the ATLAS experiment, thus making them *interesting* for the process of data selection.[8] For this reason, in the ATLAS experiment, the selection of the interesting events relevant to the predictions of the SM and BSM models, as well as to the discovery of unforeseen processes at high energies, is performed by using selection criteria that consist of *only* the aforementioned high p_T and E_T types of signatures. These selection criteria are often referred to as *inclusive triggers*, in the sense that they constitute the main set of selection criteria in the trigger menu used in the ATLAS experiment.

As the above discussion indicates, the range of interesting events in the ATLAS experiment includes a wide variety of high p_T and E_T types of signatures across a wide range of p_T and E_T values, i.e., approximately from 10 *GeV* to 1 *TeV*. The technical limitations in terms of data storage capacity and data process time make it necessary to apply data selection criteria to collisions events themselves in *real-time*, i.e., during the course of particle collisions at the collider (ATLAS Collaboration

[4] The term *signature* is used in experimental HEP to denote stable sub-atomic particles or energies into which unstable sub-atomic particles decay as a result of a physical process.

[5] Transverse-momentum is the component of the momentum of a particle that is transverse to the proton-proton collision axis, and transverse-energy is obtained from energy measurements in the calorimeter detector.

[6] A lepton is a spin ½ particle that interacts through electromagnetic and weak interactions, but not through strong interaction. In the SM, leptons include electron, muon and tau, and their respective neutrinos.

[7] In this context, the term *high* refers to the *pT* and *ET* values that are approximately of the order of 10 *GeV* for particles, and 100 *GeV* for jets.

[8] The foregoing types of signatures also differ among each other, as the predictions to which they are relevant, namely those by the SM and the BSM models, are different from each other (for details, see ATLAS Collaboration 2003, Sect. 4; Karaca 2017).

2012). Moreover, due to the aforementioned technological limitations, only a minute fraction of the interesting events could be selected for further evaluation at the stage of data analysis. This necessitates, for the fulfillment of the objectives of the ATLAS experiment, that the trigger menu (i.e. the full list of data selection criteria) be sensitive enough to select the range of types of interesting events that will serve the entire range of objectives of the ATLAS experiment. If the trigger menu were not appropriate to this end, then the data selection procedure would be biased against certain types of interesting events. As a result, the ATLAS experiment would fail to achieve some of its objectives, as the fulfillment of a particular objective of the ATLAS experiment requires the acquisition of certain types of interesting events.

A major challenge in the ATLAS experiment is to perform data selection in an unbiased manner with respect to the various objectives of the experiment. This challenge has been addressed through a particular data selection strategy that aims at increasing the sensitivity of the trigger menu, and thus of the selection procedure. To this end, the foregoing selection strategy requires the trigger menu to be sufficiently diversified in terms of types of selection signatures that are appropriate for the various objectives of the experiment. Since the ATLAS experiment is largely aimed to test the SM's prediction of the Higgs boson and the predictions of the BSM models, the adopted strategy in the first place requires the trigger menu to be sufficiently diversified in terms selection signatures composed of only high p_T and E_T types of signatures relevant to the aforementioned predictions. This aims at extending the range of the relevant LHC data that could be acquired through the trigger menu.

In the ATLAS experiment, unforeseen physics processes might also occur at low energies, i.e., approximately below 10 *GeV*. Inclusive triggers are not appropriate for the search for novel p_T and E_T processes at low energies, as these selection criteria consist of only *high* p_T and E_T types of signatures. Therefore, the selection strategy adopted in the ATLAS experiment also requires the trigger menu to be sufficiently diversified in terms low p_T and E_T types of selection signatures. These selection signatures are referred as to *prescaled triggers* and determined by prescaling inclusive triggers with lower p_T and E_T thresholds (<10 *GeV*) (for details, see ATLAS Collaboration 2003, Sect. 4.4.2). In this context, *prescaling* means that the amount of events that a trigger could accept is suppressed by what is called a *prescale factor* in order for the selection process not to be *swamped* by the events containing vastly abundant low p_T and E_T types of signatures, so that the aforementioned first set of objectives of the ATLAS experiment is not endangered. Prescaled triggers are necessary for the trigger menu, and thus of the selection procedure, to be sensitive enough to the search for novel p_T and E_T processes at low energies. Since the events containing low p_T and E_T types of signatures have the potential to be of use for some SM studies of strong interactions (see, e.g., ATLAS Collaboration 2016) as well as to provide support for new physics searches at low energies, prescaled triggers are especially aimed at further extending the range of the LHC data relevant to the second set of the objectives of the ATLAS experiment.[9]

[9] Note that these events are also used to determine trigger efficiencies and detector performance.

3 Local Data Journey at the LHC

In the ATLAS experiment, the trigger menu is applied to collision events at three different levels through the use of what are called *trigger systems* (Ellis 2010).[10] These are automated systems designed and used to select the desired events from the collision events. The first stage of the data selection process is carried out by the level-1 trigger system that provides a crude selection of the interesting events in real-time. In the ATLAS experiment, the initial event rate of the proton-proton collisions is ~40 MHz, corresponding to approximately 40,000,000 collision events per second. The first level of the data selection process is performed by the level-1 trigger system, whose technical features allow for an event-acceptance rate of 75-100 kHz. The second and third levels of the data selection process are respectively carried out by the level-2 and level-3 trigger systems, which are jointly called the *High-level Trigger and Data Acquisition System* (HLT/DAQ). Unlike the level-1 trigger system, which is hardware-based, the HLT/DAQ system is software-based, meaning that the level-1 and level-2 selection processes are performed directly by the specialized software algorithms according to the trigger menu. The level-2 and level-3 trigger systems have much smaller event-acceptance rates, which are respectively around ~2 kHz and ~200 Hz, and thereby provide finer selections of the desired events.[11] Therefore, in the ATLAS experiment, the initial event rate is gradually lowered from 40 MHz down to around 200 Hz at the end of the level-3 selection process, meaning that the interesting events are selected from the collision events at a ratio of approximately 200/40,000,000, i.e., 5 in every 1 million events.

The first stage of the data acquisition process is carried out by the level-1 trigger system that performs a *crude* selection of potentially interesting events from the collision events detected by the calorimeter and muon detectors, which are the components of the ATLAS detector system.[12] The level-1 trigger system produces a trigger decision within 2.5 μs and thereby reduces the LHC event-rate frequency of 40 MHz down to the range of 75–100 kHz. In addition to the calorimeter and muon detectors, the tracking detectors are also used in the ATLAS experiment.[13] Since the event rate is so high and thus the trigger decision time is so short, it is technologically impossible for the tracking detectors to determine particle tracks quickly enough for the level-1 event selection. Only the *hit points* produced by particles inside the tracking detectors could be recorded. These space points are later assem-

[10] The treatment in this section is based on the ATLAS Technical Design Report (ATLAS Collaboration 2003), which is a technical document that contains the design information concerning the principal components and functions of the ATLAS data acquisition system.

[11] Note that the aforementioned event-acceptance rates are valid only for the early data-taking run (Run-1) and have changed significantly during Run-1 and also during Run-2.

[12] ATLAS is a detector system that consists of different individual detectors, including the inner detector and the calorimeter and muon detectors.

[13] In HEP experiments, the tracking detectors are used to determine particle tracks as well as to measure the momenta of electrically charged particles by means of the curvatures of their tracks in a magnetic field.

bled by software algorithms in order to determine particle tracks. As a result, the data from the tracking detectors are not used directly by the level-1 trigger system for event selection. Moreover, due to the shortness of the level-1 trigger-decision time, even though the hit points are recorded, they are not completely read out from the tracking detectors during the level-1 selection. This means that the information (i.e., in terms of location in the detector, and p_T or E_T for each particle or jet contained, or associated missing E_T) necessary to fully specify a selected event is fragmented across the individual detectors of the ATLAS detector system, and that all pieces of this fragmented information are not assembled yet. Therefore, the full description of the event is not yet known, and as a result, the level-1 event selection is performed without *full granularity,* i.e., without the availability of data from all the channels of the individual detectors.

As shown in Fig. 1, the level-2 event selection begins when the sub-unit called Level-2 Supervisor sends (arrow 1)[14] the results of the level-1 selection to the sub-unit called Level-2 Processing Unit (arrow 2). Unlike the level-1 trigger system, the level-2 trigger system uses the RoI data[15] processed by the sub-unit called Read-out System (ROS) from all the sub-detectors of the ATLAS detector with full granularity. The event fragments, which are temporarily stored in the ROS, are accepted to the level-2 selection in small amounts. This way of performing event selection is called the *seeding mechanism* (ATLAS Collaboration 2003, Sect. 9.5.3.1). The ROS sends (arrows 2.1 and 2.2) to Level2Processing a subset of the event-fragments data, namely, the information regarding the locations (in the detector), momenta, and energies of the events selected at the level-1 selection. LVL2Processing sends (arrow 3.1) the information regarding the events accepted by the level-2 trigger system back to the ROS. LVL2Processing also sends (arrow 3.2) this information to LVL2Supervisor. LVL2Supervisor forwards (arrow 4) the same information to the sub-unit called Event Builder, which receives from the ROS the event-fragments data for the events selected by LVL2Processsing. Event Builder (arrow 5.1) requests from the ROS the event-fragments data for the events selected by the LVL2Processsing unit. Upon this, ROS (arrow 5.2) sends the event fragments to the Event Builder. The component called Sub-Farm Input (SFI) of the Event Builder assembles the event fragments associated with each selected event into a single record. At this stage, the full description of each selected event is available. The events that have been built are then passed (arrow 6) to the sub-unit called Event Filter Processor (EFP), through which the level-3 event selection, which is also called "event filter" (EF) selection, is carried out by specialized software algorithms (arrow 7).[16] The events that have passed the level-3 selection are then sent (arrow 8) to the sub-unit called Sub-Farm Output (SFO) for permanent storage and offline data analysis.

[14] Arrows refer to Fig. 1.

[15] The regions in the ATLAS detector that contain signals for interesting events are called *regions of interest* (RoIs). The RoIs and the energy information associated with the signals detected in the RoIs are together called the RoI data.

[16] Note that in Fig. 1, the correct arrow numbers for the messages "EFSelection" and "SendEvent" should be "7" and "8" respectively.

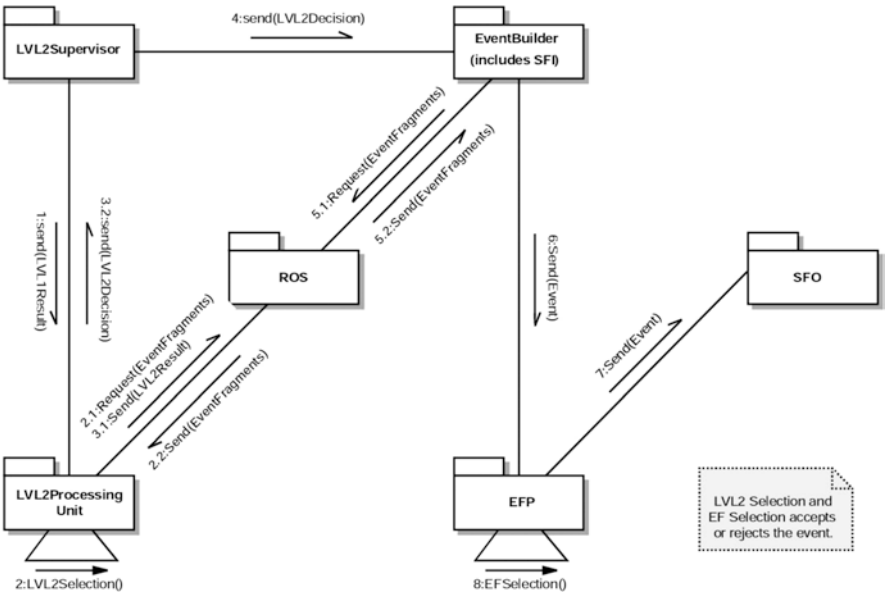


Fig. 1 The relationships between the different components of the HLT/DAQ system in the ATLAS experiment. (Source: Fig. 9–2 in ATLAS Collaboration 2003)

The details of the level-2 and level-3 selection processes are not shown in Fig. 1. These selection processes are carried out by the Event Selection Software (ESS) system, which is a software component of the HLT system (ATLAS Collaboration 2003, Sect. 9).[17] The level-2 selection of an event is carried out in a series of *steps* each of which consists of two stages. In the first stage, the event is partially reconstructed, meaning that the trigger elements (TEs)[18] associated with the event are refined and reconstructed by the *reconstruction algorithms* according to what is called the *sequence table* of the step. Each sequence in this table consists of an input TE and a reconstruction algorithm that is to be executed to refine and reconstruct an input TE into an output TE.[19] In the second stage, the event partially reconstructed undergoes a selection process based on what is called the *menu table* of the step that contains a list of the selection signatures required for this step.

The Step Handler initiates the first stage of the level-2 selection by executing the Step Sequencer to access the list of the *active* input TEs associated with an event

[17] For future reference, note that the following units to be mentioned in what follows, namely, Step Handler, Step Sequencer, Step Decision, Step Controller and Result Builder, are the software components of the ESS system that steers the HLT selection process.

[18] A TE denotes one specific signature identified by the level-1 trigger system, e.g.,"*e25i*". A TE is said to be *active* if it has previously satisfied a selection signature at the level-1 selection, or at the previous step of the level-2 selection, if the step under consideration is not the first step of the level-2 selection.

[19] Reconstruction algorithms are a class of HLT algorithms that act on the RoI data with full granularity from all sub-detectors to find new features associated with input TEs, such as a track or an isolation requirement.

selected by the level-1 trigger system. The Step Sequencer next compares the list of the active TEs with the required TEs given in the sequence table of the step. For all matching TEs, the Step Sequencer executes the reconstruction algorithms to refine and reconstruct the input TEs into the *output* TEs according to the sequence table of the step. The Step Sequencer also creates the list of the output TEs for the implementation of the seeding mechanism discussed earlier. The Step Sequencer also marks each output TE as "seeded by input TE" depending on from which input TE it has been previously created. Then, it passes each output TE to the relevant *hypothesis algorithms*—another class of HLT algorithms—that decide whether the TE is valid, depending on whether its reconstructed features are consistent with its physics interpretation. For example, if a track or an isolation requirement associated with a TE is found by a reconstruction algorithm, then the relevant hypothesis algorithm determines whether this track or isolation requirement matches the physics interpretation of the TE. The hypothesis algorithms *activate* the validated TEs and discard the invalidated TEs by deactivating them.

The Step Handler initiates the second stage of the level-2 selection by calling the Step Decision to access the list of the active output TEs, i.e., the TEs validated by the hypothesis algorithms in the first stage of the level-1 selection. The Step Decision compares the list of the active output TEs with the required selection signatures given in the menu table of the step. For the TE combinations that match the selection signatures in the menu table, the Step Decision creates a list of the satisfied signatures that consist of those matching TE combinations. The event is accepted for the next step by the Step Decision, if the TE combinations it contains satisfy at least one signature given in the menu table of the step; otherwise it is rejected and thus not considered for the level-3 selection. The Step Decision sends the information regarding the decision about the event to the Step Handler that will initiate the next step configured with a different sequence table and a menu table. The level-2 selection of an event ends at the step where it is rejected, or it continues until all required steps are completed, indicating that the event is finally accepted for the level-3 selection.

If an event is accepted at the level-2 selection, the Step Controller executes the Result Builder to provide the information necessary to *seed* the level-3 selection. This includes all satisfied signatures and the associated TE combinations, as well as the level-1 RoI data. The Result Builder assembles all these data-fragments, and the results are subsequently used for the seeding of the level-3 selection. The level-3 selection is implemented and coordinated by the Step Handler in the similar way as the level-2 selection is carried out as described above. But, the level-3 selection differs from the level-2 selection in that the TEs are now the active TEs of the level-2 selection, and that more sophisticated HLT algorithms are used to achieve a much finer event selection. As has been mentioned previously, the events that have passed the level-3 selection are stored in the Sub-Farm Output for data analysis. This marks the end of the local journey of the LHC data.

The collision events that have been rejected by the level-1 and level-2 trigger systems are removed from the data selection system. However, all the data selection operations carried out by the ATLAS data acquisition system are recorded by the system called Online Bookkeeper that produces logs stored in the form of logbook data (ATLAS Collaboration 2003, Sect. 10.4.1.2). Therefore, the ATLAS data

acquisition system is traceable in the sense that the decision regarding the acceptance or rejection of an event (already selected by the level-1 trigger system) by the level-1 and level-2 system systems can be reassessed by using the logbook data.

The LHC data is disseminated to the researchers located outside CERN through its global journey implemented by the ATLAS Distributed Data Management system (ADDM) where the acquired collision events are digitally written to *datafiles* aggregated into what are called *datasets* (for details, see Branco et al. 2008). The latter are disseminated through its four-tier hierarchical structure.[20] Tier-0 is the CERN Data Center where datasets are created, stored and distributed to Tier-1 which consist of (currently) 13 computer centers located in the following countries: Canada, Germany, Spain, France, Italy, Nordic countries, Netherlands, Republic of Korea, Russian Federation, Taipei, UK, and US. Tier-1 temporarily store datasets and distribute them to Tier-2 which consists of computer centers located typically at universities and similar scientific institutions. There are currently 150 Tier-2 sites around the world. Researchers located outside CERN can access data sets (for the purpose of data analysis) through Tier-3 which consists of local computer clusters located at universities and similar research centers or even through individual personal computers.

4 Conclusions

The technical limitations at CERN in terms of data storage capacity and data process time do not allow applying the trigger menu to the detected events without subjecting them to the construction and selection processes that make up the local data journey in the ATLAS experiment. Since the requirements for data usability are specified by the selection criteria in the trigger menu, data mobility is necessary for data usability and constitutes an essential aspect of the ATLAS data acquisition process. During the local data journey, collision events detected by the ATLAS detector system are constructed out of the fragments of proton-proton collision events that are produced by the LHC and detected by the ATLAS detector system. The first part of the local journey is a construction process in the sense that event fragments are assembled by the level-1 and level-2 triggers into *full* events. This part of the local journey is at the same time a selection process, because both events and event fragments that do not satisfy the selection criteria are filtered out and discarded from further consideration. The second part of the local data journey, which is carried out by the level-2 trigger, is solely a selection process that filters out the events constructed in the first part that do not satisfy the selection criteria. The third level of the local journey is also solely a selection process that further refines event selections made in previous levels. The above considerations show that during the local journey, events are mutable in the sense that their contents—namely, their constituent signatures—are transformed into full events by the construction and selection processes according to the selection criteria in the trigger menu. Therefore, in the context of the ATLAS experiment, data mutability in the sense of changeabil-

[20] For more information, see the URL: https://home.cern/about/computing/grid-system-tiers

ity of event content is a consequence of data mobility, which is in turn a necessary condition to apply selection criteria and thereby ensure data usability.

The above discussion indicates that the trigger menu used in the ATLAS data acquisition process should also be regarded as the set of event construction criteria, as it serves to construct events out of event fragments. The determination of the trigger menu is partly based on the theoretical judgment that the selection criteria considered relevant to the testing of the predictions of the SM and BSM models should consist of only types of signatures predicted by these models. The determination of the trigger menu also requires a judgment in the form of a data selection strategy, namely that the trigger menu should be sufficiently diversified in terms of types of signatures that are relevant to the intended objectives of the ATLAS experiment. Since the ATLAS experiment also aims at discovering unforeseen phenomena that are not accounted for by the SM and BSM models, the foregoing selection strategy also requires the trigger menu to include selection criteria that are not necessarily based on the predictions of these models. This enables using the same trigger menu to acquire data sets relevant to the entirety of the intended objectives of the ATLAS experiment. The judgment on which the data selection strategy is based is experimental, as it does not follow from the predictions of the SM and BSM models that not dictate how the trigger menu should be diversified in terms of signatures. Therefore, the foregoing theoretical and experimental judgments jointly contribute to the determination of the trigger menu and thereby impose requirements on what counts as usable data in the ATLAS experiment.

The implementation of the above-mentioned experimental strategy in the ATLAS experiment requires taking account of the technical limitations at CERN in terms of data storage capacity and data process time. This is turn leads to the judgment that the trigger menu should to be applied to collision events in real time, i.e. while proton collisions are taking place inside the ATLAS detector. This is a technical judgment based on the consideration that the amount of events produced by the LHC is so large that the foregoing technical limitations make it impracticable to apply the trigger menu after events are recorded. It is also experimental in the sense that unlike the experimental judgment concerning the trigger menu, it dictates which specific experimental procedures to use to apply the trigger menu to collision events. It thereby imposes certain technical requirements on the design of the ATLAS data acquisition system. The main technical requirement is the three-level arrangement of the trigger systems in the way it is described in the previous section. There are also more specific requirements concerning the details of the event construction and selection processes. An important technical detail is the use of the seeding mechanism according to which events fragments are accepted to the level-2 trigger in small amounts. If event fragments were accepted at once, this would considerably diminish the level-2 trigger decision time and thus render the level-2 selection process ineffective. The factors such as data processing capacity of each trigger and the amount of events produced by the LHC are also considered in specifying the details of the ATLAS data acquisition system. These technical requirements, together with the ones imposed by the experimental judgments, can be seen as the requirements imposed on the mobility and mutability of the LHC data during its local journey. While the requirements on mobility specify the ways in which events are made to

travel during the construction and selection processes, the requirements on mutability specify the ways in which the contents of events transform during these processes.

In the philosophical literature, the necessity of data mobility and data mutability for data usability has been studied and stressed in relation to data dissemination (see, e.g. Morgan 2010; Leonelli 2015). The present case-study shows that data usability is an essential concern in present-day HEP experiments already in the stage of data acquisition. In this context, in order for the experiment to achieve its intended objectives, it is necessary that the issue of data usability be dealt with before data are disseminated for analysis and interpretation. As the case of the ATLAS experiment illustrates, data mobility and data mutability are necessary conditions to deal with the issue of data usability encountered in data acquisition stage. Thus, in present-day HEP experiments, data does not come ready-made from the detector but rather is constructed to be usable for the purposes of the experiment. As a result of this construction process, data is both mobile and mutable from the outset and prior to its dissemination. Therefore, usability, mobility and mutability are related, though distinct, aspects of data in the context of present-day HEP experiments. What makes these aspects of data related to each other is the fact that they are subjected to some common requirements imposed by theoretical, experimental and technical judgments involved in the design of data acquisition systems.

Acknowledgments I would like to thank the editors of this volume and two anonymous reviewers for their valuable comments and suggestions on earlier drafts of this chapter.

References

ATLAS Collaboration. 2003. *Technical Design Report: ATLAS High-Level Trigger, Data-Acquisition and Controls*. CERN-LHCC-2003-022.

———. 2012. Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC. *Physics Letters B* 716: 1–29.

———. 2016. Measurement of $D^{*\pm}$, $D^{\pm}$ and $D_s^{\pm}$ Meson Production Cross Sections in pp Collisions at $\sqrt{s} = 7$ TeV with the ATLAS Detector. *Nuclear Physics B* 907: 717–740.

Beauchemin, Pierre-Hugues. 2018. Autopsy of Measurements with the ATLAS Detector at the LHC. *Synthese* 194: 275–312.

Boisot, Max, Markus Nordberg, Saïd Yami, and Bertrand Nicquevert. 2011. *Collisions and Collaboration: The Organization of Learning in the ATLAS Experiment at the LHC*. Oxford: Oxford University Press.

Branco, Miguel, et al. 2008. Managing ATLAS Data on a Petabyte-Scale with DQ2. *Journal of Physics Conference Series* 119: 062017.

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.

CMS Collaboration. 2002. *CMS The TriDAS project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger*. CERN-LHCC-2002-026.

Ellis, Nick. 2010. *Trigger and Data Acquisition*. Lecture given at the 5th CERN-Latin-American School of High-Energy Physics, Recinto Quirama, Colombia, 15–28 Mar 2009. CERN Yellow Report CERN-2010-001, 1–32. http://lanl.arxiv.org/abs/1010.2942

Ellis, John. 2012. Outstanding Questions: Physics Beyond the Standard Model. *Philosophical Transactions of the Royal Society A* 370: 818–830.

Franklin, Allan. 2013. *Shifting Standards: Experiments in Particle Physics in the Twentieth Century*. Pittsburgh: University of Pittsburgh Press.

Galison, Peter. 1997. *Image and Logic*. Chicago: University of Chicago Press.

———. 2003. The Collective Author. In *Scientific Authorship: Credit and Intellectual Property in Science*, ed. Peter Galison and Mario Biagioli, 325–353. New York/Oxford: Routledge.

Gutsche, Oliver, et al. 2017. Big data in HEP: A Comprehensive Use Case Study. *Journal of Physics: Conference Series* 898: 072012.

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.

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. 2017. A Case Study in Experimental Exploration: Exploratory Data Selection at the Large Hadron Collider. *Synthese* 194: 333–354.

———. 2018. Lessons from the Large Hadron Collider for Model-Based Experimentation: The Concept of a Model of Data Acquisition and the Scope of the Hierarchy of Models. *Synthese* 195: 5431–5452.

Knorr-Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Leonelli, Sabina. 2015. What Counts as Scientific Data? A Relational Framework. *Philosophy of Science* 82: 810–821.

———. 2016. *Data-Centric Biology: A Philosophical study*. Chicago: University of Chicago Press.

———. 2019. What Distinguishes Data from Models? *European Journal for Philosophy of Science* 9: 22.

Morgan, Mary S. 2010. Travelling Facts. In *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*, ed. Peter Howlett and Mary S. Morgan, 3–42. Cambridge: Cambridge University Press.

Morrison, Margaret. 2015. *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.

Koray Karaca is Assistant Professor in the Department of Philosophy at the University of Twente in the Netherlands. He received his PhD (2010) in History and Philosophy of Science from Indiana University, Bloomington, USA, and his PhD (2005) in Theoretical Physics from the Middle East Technical University, Ankara, Turkey. Before coming to the University of Twente in September 2015, he worked as a Postdoctoral Researcher between 2010 and 2015 in the interdisciplinary research collaboration "The Epistemology of the Large Hadron Collider", which is funded by the German Science Foundation and based at the University of Wuppertal, Germany. In the academic year of 2009–2010, he was Visiting Assistant Professor in the Department of Philosophy at the University of South Florida. His research interests include the philosophy and history of modern physics, philosophy of scientific experimentation in the context of high-energy physics experiments and philosophy of modelling, computer simulation and machine learning. His publications appeared in journals such as *The British Journal for the Philosophy of Science*, *Synthese*, *Studies in History and Philosophy of Modern Physics*, *Science in Context* and *Perspectives on Science*.

Tracing Data Journeys Through Medical Case Reports: Conceptualizing Case Reports Not as "Anecdotes" but Productive Epistemic Constructs, or Why Zebras Can Be Useful

Rachel A. Ankeny

Abstract Medical case reports provide an important example of data journeying: they are used to collect data and make them available for re-use to others in the field including clinicians, biomedical researchers, and health policymakers. In this paper, I explore how data journey in case reports, with particular focus on the earliest stages of the process, namely from creation and publication of case reports to the initial re-uses of them and data within them. I investigate key themes relating to case reporting and re-use, including factors which seem to smooth the path along which the data captured by a case report journey via broader citation patterns and detailed qualitative analysis of highly re-used case reports. This analysis reveals some of the key factors associated with the case reports whose data have greater amounts of journeying including publication in a general medical journal; that the data have broader implications and evidential value for topical or even urgent issues for instance in public health; and use in the case report of multiple research methods or concepts from diverse subfields. These findings along with standardization of case reporting are shown to have epistemological implications, particularly for how we understand the journeying of data.

1 Introduction

Data never stand on their own: they are gathered and become accessible via different forms of "packaging" (Ankeny 2010; Leonelli 2010, 2016) and travel over space and time. These journeys associated with their use and re-use in various contexts shape how they are understood, interpreted, and subsequently utilized. The issues associated with curation, imposing ontologies, and establishing metadata via online databases are well recognized, including the resulting epistemological limitations

R. A. Ankeny (✉)
Departments of History and Philosophy, University of Adelaide, Adelaide, SA, Australia
e-mail: rachel.ankeny@adelaide.edu.au

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_4

(Leonelli 2016). Standardization of data is a critical part of such processes and is extremely complex even where the data in question are relatively simple (such as genomic sequencing data in organism-based databases, see Leonelli and Ankeny 2012), let alone in fields where data are highly heterogeneous (e.g., in this volume see the chapters by Halfmann, Parker, Ramsden, and Wylie).

Clinical research is a domain of scientific practice where data often are extremely complex and collected in highly variable and non-standardized ways. The complexities associated with the data collected typically arise not because of the content of the data but because of our (high) level of interest in the details and the mixture of subjective and objective types of information in play whenever the main focus is on humans and particularly patients. Various types of data can be more easily standardized than others, for instance those collected in randomized controlled trials (RCTs), which can be easily aggregated using meta-analysis or similar. However other types of clinical data are much more diverse in terms of quantity, quality, provenance, means of production, attached metadata, and so on.

Medical case reports are a particularly striking example: among other purposes, they are used to collect data and make them available for re-use by others including clinicians, biomedical researchers, and health policymakers (for other uses, see e.g. Ankeny 2010, 2014, 2017a). Case reports are an ideal focus for exploring of how data "journey" at their earliest stages. They do not tend to cover great distances in any literal sense, but instead move from one context to another and thus allow exploration and development of understanding via application in new domains.

In this paper, I explore how data journey in case reports with particular focus on the earliest stages of the process, namely from creation and publication of case reports to the initial re-uses of them and data within them. Following presentation of background on medical case reports, I investigate key themes relating to case reporting and re-use, including factors that seem to smooth the path of the journey along which the data travel, via broader citation patterns and detailed qualitative analysis of highly re-used case reports. This analysis reveals some of the key factors associated with the valuing of data captured by case reports by those in the broader biomedical and health communities, as well as allowing reflections on how and when case reports are most useful and how standardization of case reporting might support the journeying of data.

As in the historical sciences, much of what is contained in medical case reports is contingent. In both fields, narratives are particularly useful ways of accounting for contingent outcomes by providing detailed data relating to them since narratives allow capture of rich descriptions that permit the envisioning of alternative possibilities or relationships.[1] Medical case reporting also involves processes of toggling back and forth between individual instances (observations on a specific patient) and the generalizations that might follow from them, if only implicitly.[2] Thus this

[1] See Beatty 2016, 2017 on issues relating to contingency as well as what makes something narrative-worthy.

[2] On similar processes in natural history, see Terrall 2017.

account also has relevance for the historical sciences and other sciences which depend closely on contingent and local observational data.

2 Background: What Are Medical Case Reports?

Medical case reports have been utilized for centuries to record and disseminate unusual presentations of illness that cannot be readily identified or that do not easily map onto recognized clinical conditions. Using a detailed narrative format,[3] they outline the diagnosis, treatment, and outcomes typically of a single patient (or a small series of patients) with a focus on practice-based observations and clinical care, rather than the results of RCTs or other experimental methodologies. One goal of case reporting is to capture data on specific instances of phenomena including many details that may not be immediately relevant, but may prove to be: the data do not have an immediate or definite purpose or target, but are collected because of their potential and future evidential value, which often is not clear when the case report is written or published. Thus these data (and the case report itself) are made available for re-use over time as subsequent instances of similar illnesses arise or as the data within the case report becomes relevant for another purpose, and so can be systematically combined into larger datasets and hence journey beyond their original domain.

Unlike RCTs or similar, the data typically contained in case reports are highly non-standardized, and include a mixture of quantitative and qualitative information. Accordingly, they are treated as one of the lowest types of evidence in the hierarchy associated with the evidence-based medicine (EBM) movement (Nissen and Wynn 2012). Some critics even have argued that highlighting the rare and unusual (termed by them "anecdotal") is dangerous, because they can lead clinician-readers to mistaken interpretations about what they are seeing in seeing in their patients and what is likely (Hoffman 1999) or that they rely on specious claims made by clinicians who wish to get published but without doing the required research (McGee 2006). It also has been documented that case reports do not receive nearly as many citations as meta-analyses or randomized controlled trial (Patsopoulos et al. 2005), and are read far less often (Leopold 2015). Hence some journals have limited the number of case reports that they publish, imposed much more detailed and stringent guidelines, or even stopped publishing them altogether. From their point of view, to borrow a phrase, "the plural of anecdote is not data" (Leopold 2015, 3074).

Advocates of case reports defend their use for particular types of purposes (e.g., Godlee 1998; Vandenbroucke 1999, 2001; Wright and Kouroukis 2000; Carey 2006; Smith 2008; Smalheiser et al. 2015; Rison et al. 2017): first, they can serve as the basis of hypotheses and direct future clinical research especially about the efficacy

[3] On narratives in case reports, see especially Hurwitz 2017; on narrative in medicine, see Gygax and Locher 2015; Hurwitz and Bates 2016.

of interventions, side effects of certain treatments or drugs, and aspects of clinical practice relating to individualized treatment (see Ankeny 2014, 2017a). Case reports have proven useful for identifying adverse and beneficial effects and recognizing new or rare diseases or unusual manifestations of common diseases; an oft-cited example of the success of case reporting is the recognition of the relationship between use of the drug thalidomide by pregnant women and congenital abnormalities in newborns (McBride 1961; Lenz 1962). Case reports can serve as a type of evidence even in EBM when used in the appropriate manner (Jenicek 2001) and can also be useful in clinical education (Cabán-Martinez and García-Beltrán 2012), particularly given the dominance of problem-based learning approaches in medicine. Finally, there is some evidence that case reports can make significant contributions to medical research planning (Albrecht et al. 2005).

Case reports account for a rapidly growing number of medical publications and an increasing number of dedicated journals in recent years, with at least 160 case reports journals from 78 publishers documented as of mid-2015 (for a useful summary, see Akers 2016, Table 1 available online), with observers commenting that there has been a "renaissance of the case reporting literature" (Smalheiser et al. 2015, 171). More generally in the field of medicine considered as a whole, the number of MEDLINE-listed case reports is said to substantially exceed the number of published clinical studies (Kiene et al. 2013). The newer journals tend to be open access and range from having a focus on general medical issues to accounts of case reports in more specialized subfields. Unfortunately, predatory publishing practices are particularly rampant among case report journals (Akers 2016), with nearly 50% of publishers engaging in questionable publishing practices. In addition, few have impact factors, in part because of the infrequency with which case reports are cited, but nearly half of the journals (as of mid-2015) are indexed in PubMed (Akers 2016), making them accessible to clinicians and researchers and for analysis of the types performed in the current paper.

Unlike other parts of medical training and publication (e.g., differential diagnostic processes or mortality and morbidity reporting, see Bosk 1979), the processes of recording this type of historical data generally have not been made consistent or standardized. Thus case reports have been viewed by many within the field as insufficiently rigorous for aggregation for data analysis which would be rigorous enough to inform research design and allow data to journey to new domains to permit comparisons across diverse contexts including different sociocultural settings. Traditional approaches to gathering data via case reports make it difficult to locate and re-use relevant data despite considerable technological improvements related to the rise of open access and internet-based systems.

Out of recognition of many of these limitations, consensus-based international guidelines have been developed, called the "CAse REport" or CARE guidelines (Gagnier et al. 2013), to increase the completeness in the presentation of published case reports, create more comparability between the data contained in case reports particularly with regard to potential therapeutic interventions and outcomes, and generate more transparency for patients and practitioners, and in turn to inform

clinical practice guidelines. When adopted by a journal, these types of guidelines have been argued to be associated with an increase in the completeness of the information published (e.g., Turner et al. 2012) and hence can be viewed as critical pragmatic constructs.

The CARE guidelines are a 13-item checklist outlining basic reporting requirements for published case reports, provided in a structured manner. The key goal is to increase completeness and transparency in published case reports. The authors stress that they "attempted to strike a balance between adequate detail and the concise writing that is one of the appealing characteristics of a case report" (Gagnier et al. 2013, 4).[4] As discussed elsewhere (Ankeny 2017b), these guidelines are extremely revealing with regard to the underlying epistemology of case reporting particularly in the current era of dedicated journal outlets which have considerable investment in establishing case reports as a valid form of evidence. For the purposes of this paper, I do not analyze them in any detail particularly because their promulgation has been quite recent but do use some of the issues highlighted in the guidelines in my analysis of re-use patterns.

3 Detecting Patterns and Themes in Case Reporting and Re-use

3.1 Broader Patterns of Re-use

One of the main potential benefits of publishing case reports (and providing the necessary infrastructures to make them more accessible) is so they can be re-used by others who come across similar phenomena particularly in clinical settings, or so that the observational data can be used as the basis for initiating various types of research. Hence it is useful to look at the broader patterns of re-use to get a sense of the uptake of medical case reports.

More generally, it must be noted that citation analysis may severely underestimate the impact of clinical-oriented research in certain fields particularly in comparison to basic research (e.g., Van Eck et al. 2013) and case reports specifically are cited at a negligible rate compared to other types of publications (Patsopoulos et al. 2005). However for the purposes of this paper, a focus on published literature is appropriate because I am primarily interested in explicit re-uses of data captured in

[4]This structure also aims to capture "useful" information including that required by the U.S. Department Health and Human Services to demonstrate so-called "meaningful use" of certified electronic health records, which in turn is required by some private insurers for physicians to for health providers to quality for certain types of performance incentives. Although intriguing to consider the epistemological impacts of these social and financial incentives, an analysis of the interplay between these requirements and the content of case reporting guidelines is beyond the scope of this paper.

case reports; note that this approach of necessity will fail to capture negative instances, that is, where a case report was accessed and utilized but found not to be relevant to the current problem or phenomenon under examination, in similar ways to the type of publication bias that has been well recognized with regard to negative results (e.g., Kicinski et al. 2015). Article usage statistics (available via many journals and databases) might well provide more accurate quantitative information that could be compared across case reports but would fail to allow any assessment of whether or how a case report is being re-used.

Tracking re-use of medical case reports is plagued with technological difficulties, particularly when assessing case reports via citations across all types of journals and medical subfields: for instance although PubMed[5] indexes nearly half of the journals that publish case reports, excludes most that are likely to be predatory journals (Akers 2016), and provides a "case reports" filter, it does not allow analysis of articles by number of citations (similar limitations occur with Embase, another major medical database). An additional issue is that there are inaccuracies in the tagging of publications as "case reports" (see note 6 below for a rough estimate of the rate of inaccuracy). Even tracing case report patterns by journal by focusing on the dedicated journals is complicated by the fact that several major case report journals have changed name over time and full datasets are thus not readily available.

Hence I used two strategies to analyze case reporting and re-use over the past 25 years: (1) a broader strategy allowing general patterns of re-use (using citations as a proxy) to be visualized; and (2) a more specific strategy focused on highly cited case reports. The temporal window of 1997–2017 was selected to permit inclusion of both the newer journals focused on case reports as well as more traditional journals which publish case reports; it also allows medium- and longer-term re-use to be tracked, since as the analysis reveals, re-use often only occurs over considerable periods of time.

For the first broader search, Web of Science was utilized using a case report focused strategy for medically related fields[6] to extract data for the years 1997–2017, which generated a total of 108,348 case reports. Just over 30% of these reports have no citations to date, and just over 17% have between one and ten citations since time of publication. A second analysis used the Medline subset within Web of Science,

[5] PubMed is a free search engine which primarily allows access to the MEDLINE database of references and abstracts on life sciences and biomedical topics, which in turn is managed by the United States National Library of Medicine at the National Institutes of Health as part of the Entrez system of information retrieval.

[6] This search was performed on 14 November 2018 by utilizing Web of Science Core Collection to search for all items tagged as topic = "case report∗" for the years 1997–2017, which generated a list of all items tagged as case reports from more generalized journals. This set was then supplemented by inclusion of all publications in case report-focused journals for the same time period (identified by explicit inclusion of "case report" or similar in the journal titles, and drawing on the list published in Akers 2016). These sets were combined and then narrowed to include only those with topic = "human∗" or "patient∗", and by excluding publications coded to non-medical categories. The publications were then run through Clarivate InCites to obtain rates of citation.

to perform a search for case reports for the years 1997–2017.[7] The results were 102,195 articles, of which only 98 (slightly more than .09%) of the publications verified to be case reports[8] were highly cited in their respective field as of March/April 2018 (i.e., they received enough citations to place them in the top 1% of their academic field based on a highly cited threshold for the field and publication year); of these publications, only four were published in the past 2 years and received enough citations to place them in the top 0.1% of papers in their respective academic field. Thus these findings echo previous analyses of the relative neglect of uptake of case reports, but do permit us to focus on those that may have resulted in important instances of re-use.

The highly cited case reports do share certain characteristics: first, they tend to appear in highly popular, general medical journals (e.g., *The New England Journal of Medicine),* which have extremely large readerships. They also cover one of three main topics broadly defined, namely non-randomized and non-controlled trials of experimental drugs or therapies on individuals or very small groups of patients, often on a compassionate or emergency basis; epidemiological or other features of emerging or novel diseases that are typically infectious in nature; and characterization of underlying mutated genetic sequences of disease-related phenotypes or processes at other levels (such as tumors). Less frequent topics include adverse effects of or reactions to therapies of various types; reporting of new illegal drug use and effects; and longer-term outcomes of novel surgical procedures, particularly organ and other transplants. Despite all of these publications being considered to be highly cited, there is no particularly robust correlation between year of publication and the number of citations, and the range in the number of citations is large, from nearly 1500 for a 2011 paper on using modified T-cells to treat leukemia, to 15 for a 2017 paper published in a more narrow subfield, toxicology, focused on episodes of intoxication via a new synthetic opioid.

Although these broader trends give us hints about how data can journey to new domains via case reports, more qualitative analysis helps to reveal precisely what travels from early stage case reports and what roles such data journeying serves. Hence in the following sections, a series of highly cited case reports are analyzed to provide insights into the valuing of data captured by case reports and what factors are associated with re-use. I have opportunistically selected two case reports to explore which have particularly interesting patterns of data re-use but have attempted to represent two of the main types of case reporting captured in the quantitative analysis above.

[7] This search was performed by selecting "case report" in the document type field for the years 1997–2017, then limiting to core clinical journals and to humans (17 July 2018).

[8] The original set that was automatically generated on 17 July 2018 included 118 articles, of which 20 (17%) were determined not to be case reports based on manual review of abstracts; some appeared to be review articles that had been mistagged whereas others were very large observational studies that strictly speaking would not typically be considered to be serial case reports but which some journals nonetheless place in their "case report" sections.

3.2 Case Reports on Infectious Diseases

One key role played by case reports is to draw attention to emerging or novel infectious disease processes: in recent years, occurrences of Zika, Middle East Respiratory Syndrome (MERS), Ebola, and Influenza A have been described via case reporting, with attention to a range of aspects of the phenomena under study. These case reports often contain important data that then can journey rapidly from their location of creation and reproduce faithfully, as long as certain features are in place.

For instance, a case report (Gao et al. 2013) of three observed human fatalities related to infection with a new form of the avian influenza A virus (H7N9) in Shanghai, China was among the most highly cited (1247 times) in the data set above, as the initial publication relating to what subsequently became a pandemic. Previously the transmission of H7 viruses to mammals had been rarely reported in Asia, human infection with the N9 subtype had not been documented anywhere, and these types of infections had rarely been fatal or as severe as in the patients who presented for care in Shanghai. The case report summarizes the typical information about the patients, including demographic and epidemiological characteristics, particularly those associated with pre-existing conditions likely to have depressed their immune systems as well as potential contact with chickens; the complications, treatment, and clinical outcomes of the patients; and detailed analysis of the characteristics of the virus isolated from the patients.

In conclusion, the authors (many of whom have numerous previous publications on different forms of epidemic influenza particularly in China) make an urgent call to others in the medical field: "We are concerned by the sudden emergence of these infections and the potential threat to the human population. An understanding of the source and mode of transmission of these infections, further surveillance, and appropriate counter measures are urgently required" (Gao et al. 2013, 1896). Among the key points discussed is whether this novel version of the virus occurred within these human hosts or was directly transmitted by birds, with the latter said to be the preferred explanation, particularly based on genetic sequencing and other forms of analysis. However a critical point made in the case report is that influenza surveillance of birds, swine, and humans is limited in China and nearby countries, which makes it very difficult to provide an answer to this question.

With regard to the processes associated with data journeying, a few critical points are notable. First, the initial journeying of the data from the clinical setting to the printed case report (and hence to them becoming available publicly on a global basis) occurred over a highly compressed time period[9]: the patients were seen

[9] Case reports typically have longer gestation times between clinical observation, laboratory analysis, and other processes, and actual publication, even when focused on similar public health related issues: see for instance Colson et al. 2010 on a small case controlled study within a single family on the transmission of hepatitis E via figatellu, a traditional pig liver sausage widely eaten in France and commonly consumed raw, where initial observations and data collection occurred in 2007–9 but which was not published until 2010; nonetheless this case report also is among the most highly cited in its field.

between mid-February and the end of March 2013, and the case report was published online in mid-April 2013.[10] Subsequently when published in print in mid-May, it was accompanied by a high-profile editorial by researchers at the US Centers for Disease Control which lauded the authors of the case report for the speed with which the virus was identified and whole genome sequences of it made available, particularly given the global public health issues raised (Uyeki and Cox 2013, which echoes an earlier editorial in *Nature* in April, Anonymous 2013), which was important because of the lack of transparency that sometimes had occurred in the context of past epidemics in China (e.g., with reference to SARS, see Knobler et al. 2004). These factors underscore that the speed with which data from a case report journeys and the extent to which it travels (i.e., how often it is picked up by others reporting research and whether it reaches a global audience) is directly related to a number of factors including the perceived usefulness of the original case report in terms of the data contained within it and the potential threat posed by the condition(s) described, both of which are common in infectious disease related case reports.

Second, data within case reports are more likely to be re-used if they relate to multiple research methods or fields. For instance, the editorial cited above underscored many other critical points raised by the case report, namely that some of the sequence data suggested that this virus was likely to result in asymptomatic or mild avian disease, and thus had the potential to generate a silent widespread epizootic epidemic in China and neighboring countries. Many of the subsequent publications citing the original case report explore these types of issues (e.g., Xu et al. 2013). In addition, in the 6–12 months after the original case report, various members of the research team published more detailed reports (sometimes as research letters, presumably in order to get them published quickly given the urgency of what was quickly becoming a public health crisis)[11] in high-profile outlets, such as on the biological features of the virus, epidemiological surveillance, and tracing the genesis of the infection via various types of birds (e.g., Lam et al. 2013) which helped to widen the exposure of the original publication particularly in fields beyond infectious disease. Hence various types of data originally contained within the original case report journeyed without necessarily being closely connected to the initial case report. Examples include numerous publications related to technology development such as new methods for real-time detection of infection (e.g., Zhu et al. 2013).

[10] Although beyond the scope of this paper, it is worth noting that formal mechanisms such as infectious disease reporting and more informal mechanisms such as media coverage can help data in a case study to journey. According to the journal *Nature* (Anonymous 2013), China reported the H7N9 outbreak to the World Health Organization (WHO) on 31 March 2013, and simultaneously published the genomic sequences of viruses from the three human cases on the database of the Global Initiative on Sharing Avian Influenza Data (GISAID). It also shared all of the sequences with the WHO, and live virus with the WHO and other laboratories. In addition, the Chinese media reported new cases on a daily basis and discussed H7N9 fairly openly, with Chinese President Xi Jinping publicly calling for an effective response, noting that the government should ensure release of accurate information about the outbreaks.

[11] Self-citations are common among citations to previously published case reports, and are difficult to systematically eliminate from larger datasets when mapping patterns of re-use.

Finally, the "call to arms" for more surveillance and reporting in the case report (and associated publications such as the accompanying editorial) resulted in numerous publications about additional instances of the disease, as well as having clear public policy implications, which also appears to be a mark of a case report from which data are likely to journey. Thus potential wider relevance along with "actionability" of data (see Ramsden in this volume) is often associated with wider patterns of journeying. For instance following the case report and in part based on its findings, H7N9 influenza was established as a notifiable infectious disease in Taiwan which experienced a spike of cases amongst travelers returning from China soon after the initial outbreak in China (TCDC 2013). As underscored in a paper citing the original case report, one of the lessons to be learned from this case report is more generic, and relates to the importance of this type of data having a way to journey outward, particularly given certain tendencies reinforced in medical training: "Instead of recognizing that billions of people worldwide are exposed to important and emerging infectious diseases, our training has relegated this topic mostly to 'tropical medicine' or public health or labelled the threat as a 'zebra' item" (McFee 2013), referring to the medical training adage that "if you hear hoofbeats, think horses, not zebras" (see Hunter 1996; Wright and Kouroukis 2000). Given increased globalization together with the emergence of various serious health threats, some "zebras" are now critically important, and there is a critical need for pandemic preparedness. Thus these sorts of public health emergencies require not only rapid data collection and analysis, but also data sharing and feedback (Uyeki and Cox 2013; see also Lurie et al. 2013) via "data journeying" particularly in conceptual terms. Case reporting provides a clear mechanism for these processes to occur, especially where detailed data are provided in case reports that are useful for epidemiological tracking and related processes (Anonymous 2013).

3.3 Case Reporting of Adverse Effects

Another key category of case reporting relates to adverse or unexpected effects particularly of commonly utilized treatments or drugs. Consider a highly cited case report detailing two fatalities and one life-threatening incident in young children related to consumption of codeine for pain relief after adenotonsillectomy for obstructive sleep apnea syndrome (Kelly et al. 2012). The Canadian team proposed that where the surgery has not resolved the sleep apnea, morphine is particularly dangerous as it may further worsen the respiratory condition, can be fatal in cases where children have a certain genetic allele that can lead to a toxic accumulation of morphine exceeding therapeutic levels, and is of particular concern in individuals of North African descent where the mutation is more common (occurring in 30% of the population).

Some members of the team (together with the chief coroner for the province) had previously published a letter in 2009 focused on a single case similar to the 2012

series which documented the death of an otherwise healthy 2 year old with functional duplication of a particular genetic allele known to be associated with increased rates of conversion of codeine to morphine and which may have contributed to respiratory depression and death, in concert with other factors (Ciszkowski et al. 2009). In this letter, the authors declare that "given the polymorphic nature of codeine metabolism and the fact that adenotonsillectomy does not reverse all cases of obstructive sleep apnea, codeine cannot be considered a safe outpatient analgesic for young children after adenotonsillectomy."

Tracing the citations to the 2012 case report reveals several key themes: first, the uptake of the 2009 letter (and the data contained in it) was much more limited, based on citation patterns, than the case report which appeared later, despite both appearing in very high-profile medical journals (*The New England Journal of Medicine* and *Pediatrics* respectively). However there are several reasons which seem to be correlated with this difference, notably that the 2012 case report was in fact peer-reviewed and detailed multiple instances of the observed phenomenon. Description of multiple occurrences of a phenomenon appears to result in the case report and the data contained in it being valued more highly, likely because it is viewed by readers as providing more or more robust evidence especially because other underlying factors can be ruled out; even if three cases may seem to many to still be anecdotal, in this example multiple cases appear to have resulted in more re-use of data and of the case report itself, at least in the form of citations.

An additional trigger which contributes to wider recognition and re-use of case reports is whether the observed adverse effects come to be formally certified, such as in recognition by regulatory authorities or professional organizations. In the current case, during late 2011, the Patient Safety and Quality Improvement Committee of the American Academy of Otolaryngology–Head and Neck Surgery (AAO-HNS) had become concerned about adverse events, particularly respiratory depression, after adenotonsillectomy and conducted a nationwide, anonymous survey of otolaryngologists about such events (Racoosin et al. 2013). By August 2012 following an evaluation of the safety of use of codeine in children including a comprehensive review of the literature and case reports submitted to the US Food and Drug Administration (FDA)'s Adverse Event Reporting System, the FDA issued a press release and drug safety communication warning of the risk of respiratory depression and/or death following the use of codeine after tonsillectomy. Its review found 13 cases, including 10 deaths and 3 cases of life-threatening respiratory depression associated with codeine use during the period 1969 and 1 May 2012 (including the original case reports). The issuing of the FDA advisory is correlated with a sharp increase in citations to the 2012 report, likely simply out of increased awareness of these issues, with many of the publications exploring implications of these findings for codeine use in children in this or other types of care settings.

In addition, the scale of the potential for adverse effects clearly contributes to the re-use of case reports. Although the complication in the case at hand is likely rare, it has the potential to affect a significant number of children given the huge number

of adenotonsillectomies performed per year, about half a million annually (Cheng and Sobol 2013). Case reports are more likely to be viewed as oddities or mere anecdotes if they seem to have very small-scale effects, in which case they are obviously not particularly ripe for re-use of the data contained within them.

A final factor about whether data contained in case reports about adverse effects are subsequently re-used seems to be related to whether they align with other broader epistemological understandings or trends in patient care, public health, or other types of medical practices (again here compare the chapters by Cambrosio et al. and Ramsden with particular attention to the idea of actionability of data especially in clinical research practices). There are at least two potential ways in which these issues are likely have been in play in this example: first, as noted in a Perspectives piece published in *The New England Journal of Medicine* following the FDA warning, increased awareness of what they term "the value of both personalized medicine and the reporting of rare adverse outcomes" (Racoosin et al. 2013, 2155) has resulted in more attention to and publicity about such adverse effects. In other words, the genomic turn of the early 2000s has resulted in greater awareness of genetic diversity including mechanisms relating to drug reactions, and greater abilities to provide alternative clinical treatments. These claims are substantiated in the types of articles citing this case report, many of which make reference to the need for more precise methods to determine optimal approaches to pain control, particularly with young children post-adenotonsillectomy, and some of which position these claims explicitly within the emerging field of pharmacogenetics (e.g., Lee et al. 2014; Smith et al. 2018).

But a second likely trigger of the patterns of re-use observed relates to the increasing awareness of the so-called "opioid epidemic" in the 2010s, especially in the United States.[12] Due to increases in opioid-related addiction, overdoses, and deaths, opioid use came to be viewed as a public health crisis in this period, in part related to illegal drug use but also in concert with over-prescription of legal pain medications including oxycodone which is chemically and otherwise similar to codeine. Thus in the re-use of data from the original case report, we find it cited simply as evidence of the potential dangers of codeine use in articles more broadly exploring the potential benefits and dangers of prescribing it not only for children (e.g., Carter et al. 2013: Martin et al. 2014) but in certain groups likely to be more at risk such as immigrants (e.g., Ray et al. 2014, which in fact observed no increased risk in these groups despite language and genetic differences). Further, due to subsequent changes in the way the FDA classified ("scheduled") hydrocodone combination products in 2013, several of the publications (e.g., Fleming and Wanat 2014) emphasize the potential dangers of codeine-based products for pain management, in part out of recognition that there would be a tendency to increase use of these products as these remain accessible at levels requiring less approval

[12] Even using the terminology of "epidemic" in this context raises a range of historical, political, and sociological issues, but this issue is not a main focus in this paper; for discussion, see for instance Green et al. 2002; Martin and Martin-Granel 2006.

processes. Hence as this case report shows, re-use of data can become quite loose and its journeying more akin to wandering where some part of the case report proves to have much broader relevance, and particularly where there are practice and public health implications.

4 Conclusions: Implications for Understanding How Data Journey

What makes data more likely to journey beyond their original case reports? It is clear that a few factors can be identified; though these are neither necessary or sufficient, they do provide some marks that assist us with understanding the potential epistemological value of case reporting and the data contained within them. First, case reports that have implications well beyond their immediate domain are likely to be published in general medical journals which allows them to be read much more widely, and hence to much more easily be conceptualized as having broader relevance. Second, the data contained in case reports tend to journey when they have content with broader implications well beyond the case report at hand, and particularly when the data have evidential value for topical or even urgent issues, particularly those arising in public health. Any potential for wider applicability may well not be explicitly detailed in the original case report, but can be spurred on by additional factors, such as relevance for policy, uptake and endorsement by professional organizations or governmental authorities, description in other contexts such as framing editorials accompanying the case report, and so on.

Third, use of multiple research methods or concepts from diverse subfields within medicine can expedite the journeying of data within a case report into a range of types of journals and allow the data to journey well beyond their original context. Thus larger teams of authors are often common in the most highly cited case reports, likely in part because diverse expertise is necessary for case reports that bring together different types of data, but this pattern in turn seems to support greater potential for the data to journey more widely. Finally, data from case reports tend to journey where there is alignment with broader epistemological understandings or agendas within medicine: for instance the turn toward genomics in the 2000s resulted in journeying of data associated with numerous case reports related to unusual phenotypic disease patterns or adverse effects to other contexts, notably to publications detailing more fundamental biomedical research to determine the genetic basis for these patterns or effects.

What can be said about the efficiency of the journeying of data from case reports? The empirical data and qualitative analysis presented above reveal that the speed with which data from a case report journey and the extent to which they travel is correlated not only with the perceived usefulness of the original case report in terms of the original data contained within it (as would be expected) but also by the potential

threat posed by the condition(s) described: so infectious disease-related case reports often are urgently reported and data from them picked up elsewhere. In addition, as occurred in the case report on the adverse effects of codeine in young tonsillectomy patients, data associated with case reports where broader implications subsequently come to be recognized (e.g., for postoperative pain control or even pain control in general in this example) have their journeying expedited by their application in these broader contexts.

These issues related to journeying take us back to the various efforts to standardize case reporting: why bother limiting data captured by case reports to certain categories when our technological infrastructures in fact might permit us to "write down everything," and in principle create more potential for journeying? One part of the answer clearly relates to the requirement that case reports be useable by medical practitioners who are both the authors of the guidelines and many of the likely users (and re-users) of the case reports and the data contained in them: not all data that might be captured and packaged in a case report are of equal relevance, which can be seen in the factors more closely associated with journeying outlined above. Thus new efforts at highly structured guidelines about what must be included impose a certain rigor to what is thought to be essential for understanding a case report and for re-using the data contained in them to identify similar cases or other domains where the data might have relevance. Though in some sense it is technically possible to include absolutely all data (or many more pieces of data than currently contained in case reports), to do so would undermine the structures (narrative and otherwise) that form the basis for what the case report is a case *of,* and hence place limits on the abilities of practitioners to re-use it.

In addition, these guidelines have certain merits beyond mere standardization for ease of re-use of case reports and the data within them: at a deeper level, they constitute a line of attack on traditional assumptions regarding what types of data are valued and under what circumstances. Case reporting in a standardized manner reinforces the value of data derived from individual case reports and helps to establish methods for consistent re-use. These types of guidelines also underscore how data can serve evidence in these sorts of observational settings that previously have been assumed to be unable to be systematized in any significant ways, particularly as compared to RCTs and other experimental methodologies. As the authors note, what is most critical is that case reporting be made more precise, complete, and transparent (Gagnier et al. 2013), which no doubt is correct. However as this paper has shown, there are deeper epistemic issues underlying the re-use of case reports and the journeying of the data within them, and these guidelines have the potential to allow both creators and users to be reflective about both the potential (and limitations) of case reporting, particularly in the context of re-use.

Exploring the effective journeying of data contained in case reports together with efforts to standardize the presentation of data are important parts of developing deeper understandings of appropriate, effective, and rigorous ways of using observation-based methodologies in the biomedical sciences and other fields that rely on such approaches, given that these have been largely neglected, for instance in medicine due to the rise of EBM and related approaches in which data are relatively easy to systematize (cf. Tempini and Teira in this volume on

the difficulties of circulating data in other settings). As the guideline authors state, "When it becomes clear how new data contributes to evidence, the stewardship needed to produce high-quality data will become more rewarding and our attitude toward 'observation' will shift…This will transform how we think about 'evidence' and revolutionize its creation, diffusion, and use—opening new opportunity landscapes" (Gagnier et al. 2013, 5). How these types of data journey faithfully and efficiently in a variety of contexts and hence come to be valued as a form of evidence warrants further exploration.

Acknowledgements I wish to acknowledge Vikki Langton, Liaison Librarian for the Faculty of Health and Medical Sciences, University Library, University of Adelaide, as well as Anthony Dona at Clarivate, for their support and expertise with regard to the search strategies employed to obtain the empirical information in this paper. Two honours students whose theses I had the pleasure of supervising in History at the University of Adelaide, Patrick Reynolds (on discourses surrounding the SARS epidemic in China) and Hugh Scobie (on the sociohistorical context of the "opioid epidemic" in the United States), helped to inform my approaches to some of the cases examined in this paper. I also am grateful to the participants in the PSA 2016 symposium and the Data project workshop at which earlier versions of this paper were presented, particularly Sabina Leonelli, James McAllister, and Mary Morgan, for insightful comments that have helped considerably to shape the final version of this paper.

References

Akers, Katherine G. 2016. New Journals for Publishing Medical Case Reports. *Journal of the Medical Library Association* 104: 146–149.

Albrecht, Joerg, Alexander Meves, and Michael Bigby. 2005. Case Reports and Case Series from *Lancet* Had Significant Impact on Medical Literature. *Journal of Clinical Epidemiology* 58: 1227–1232.

Ankeny, Rachel A. 2010. Using Cases to Establish Novel Diagnoses: Creating Generic Facts by Making Particular Facts Travel Together. In *How Well Do Facts Travel?* ed. Peter Howlett and Mary S. Morgan, 252–272. Cambridge: Cambridge University Press.

———. 2014. The Overlooked Role of Cases in Causal Attribution in Medicine. *Philosophy of Science* 81: 999–1016.

———. 2017a. The case study in medicine. In *The Routledge Companion to Philosophy of Medicine*, ed. Miriam Solomon, Jeremy R. Simon, and Harold Kincaid, 310–318. New York: Routledge.

———. 2017b. The Role of Patient Perspectives in Clinical Case Reporting. In *Knowing and Acting in Medicine*, ed. Robyn Bluhm, 97–112. New York: Rowman & Littlefield.

Anonymous. 2013. The Fight Against Bird Flu. *Nature* 496: 397.

Beatty, John. 2016. What Are Narratives Good For? *Studies in History and Philosophy of Biological and Biomedical Sciences* 58: 33–40.

———. 2017. Narrative Possibility and Narrative Explanation. *Studies in History and Philosophy of Science* 62: 31–41.

Bosk, Charles. 1979. *Forgive and Remember: Managing Medical Failure*. Chicago: University of Chicago Press.

Cabán-Martinez, Alberto, and Wilfredo F. García-Beltrán. 2012. Advancing Medicine One Research Note at a Time: The Educational Value in Clinical Case Reports. *BMC Research Notes* 6: 293.

Carey, John C. 2006. Significance of Case Reports in the Advancement of Medical Scientific Knowledge. *American Journal of Medical Genetics* 140A: 2131–2134.

Carter, Bernie, Daniel B. Hawcutt, and Janine Arnott. 2013. The Restrictions to the Use of Codeine and Dilemmas About Safe Alternatives. *Journal of Child Health Care* 17: 335–337.

Cheng, Jeffrey, and Steven Sobol. 2013. Despite Warnings, Some Graduating Otolaryngology Residents Planning to use Codeine for Posttonsillectomy Pain Control (Letter). *Otolaryngology–Head and Neck Surgery* 148: 356–357.

Ciszkowski, Catherine, et al. 2009. Codeine, Ultrarapid-Metabolism Genotype, and Postoperative Death (Letter). *The New England Journal of Medicine* 361: 827–828.

Colson, Philippe, et al. 2010. Pig Liver Sausage as a Source of Hepatitis E Virus Transmission to Humans. *The Journal of Infectious Diseases* 202: 825–834.

FDA (Food and Drug Association), US Department of Health and Human Services. 2012. FDA Drug Safety Communication: Codeine Use in Certain Children After Tonsillectomy and/or Adenoidectomy May Lead to Rare, but Life Threatening Adverse Events or Death. http://www.fda.gov/Drugs/DrugSafety/ucm313631.htm. Accessed 20 July 2018.

Fleming, Marc L., and Matthew A. Wanat. 2014. To Prescribe Codeine or Not to Prescribe Codeine? *Journal of Pain & Palliative Care Pharmacotherapy* 28: 251–254.

Gagnier, Joel J. et al. or the CARE Group. 2013. The CARE Guidelines: Consensus-Based Clinical Case Reporting Guideline Development. *Journal of Medical Case Reports* 7: 223, https://doi.org/10.1186/1752-1947-7-223 (published in numerous journals simultaneously as an open-access publication).

Gao, Rongbao, et al. 2013. Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus. *The New England Journal of Medicine* 368: 1888–1897.

Godlee, Fiona. 1998. Applying Research Evidence to Individual Patients. *British Medical Journal* 30: 1621–1622.

Green, Manfred S., et al. 2002. When Is an Epidemic an Epidemic? *Israel Medical Association Journal* 4: 3–6.

Gygax, Franziska, and Miriam A. Locher, eds. 2015. *Narrative Matters in Medical Contexts across Disciplines*. Amsterdam: John Benjamins Publishing Company.

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.

Hoffman, J.R. 1999. Rethinking Case Reports: Highlighting the Extremely Unusual Can Do More Harm than Good. *The Western Journal of Medicine* 170: 253–254.

Hunter, Kathryn M. 1996. Don't Think Zebras': Uncertainty, Interpretation, and the Place of Paradox in Clinical Education. *Theoretical Medicine* 17: 225–241.

Hurwitz, Brian. 2017. Narrative Constructs in Modern Clinical Case Reporting. *Studies in History and Philosophy of Science* 62: 65–73.

Hurwitz, Brian, and Victoria Bates. 2016. The roots and ramifications of narrative in modern medicine (Chapter 32). In *The Edinburgh Companion to the Critical Medical Humanities*, ed. Anne Whitehead and Angela Woods. Edinburgh: Edinburgh University Press.

Jenicek, Milos. 2001. *Clinical Case Reporting in Evidence-Based Medicine*, 2nd ed. London: Arnold Publishers

Kelly, Lauren E., et al. 2012. More Codeine Fatalities After Tonsillectomy in North American Children (Case Study). *Pediatrics* 129: 1343–1347.

Kicinski, Michal, David A. Springate, and Evangelos Kontopantelis. 2015. Publication Bias in Meta-Analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine* 34: 2781–2793.

Kiene, Helmut, Harald J. Hamre, and Gunver S. Kienle. 2013. In Support of Clinical Case Reports: A System of Causality Assessment. *Global Advances in Health and Medicine* 2: 64–75.

Knobler, Stacey, et al., eds. 2004. *Learning from SARS: Preparing for the Next Disease Outbreak: Workshop Summary*, Institute of Medicine (US) Forum on Microbial Threats. Washington, DC: National Academies Press.

Lam, Tommy Tsan-Yuk, et al. 2013. The Genesis and Source of the H7N9 Influenza Viruses Causing Human Infections in China (Letter). *Nature* 502: 241–244.

Lee, Judith W., et al. 2014. The Emerging Era of Pharmacogenomics: Current Successes, Future Potential, and Challenges. *Clinical Genetics* 86: 21–28.

Lenz, Widukind. 1962. Thalidomide and Congenital Abnormalities. *Lancet* 1: 45.

Leonelli, Sabina. 2010. Packaging Data for Re-use: Databases in Model Organism Biology. In *How Well Do Facts Travel?* ed. Peter Howlett and Mary S. Morgan, 325–348. Cambridge: Cambridge University Press.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

Leonelli, Sabina, and Rachel A. Ankeny. 2012. Re-thinking Organisms: The Impact of Databases on Model Organism Biology. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 29–36.

Leopold, Seth S. 2015. Case Closed—Discontinuing Case Reports in *Clinical Orthopaedics and Related Research®*. *Clinical Orthopaedics and Related Research* 473: 3074–3075.

Lurie, Nicole, et al. 2013. Research as a Part of Public Health Emergency Response. *The New England Journal of Medicine* 368: 1251–1255.

Martin, Paul M.V., and Estelle Martin-Granel. 2006. 2,500-Year Evolution of the Term Epidemic. *Emerging Infectious Diseases* 12: 976–980.

Martin, D.P., et al. 2014. The Safety of Prescribing Opioids in Pediatrics. *Expert Opinion on Drug Safety* 13: 93–101.

McBride, William G. 1961. Thalidomide and Congenital Abnormalities. *Lancet* 2: 1358.

McFee, Robin B. 2013. Global Infectious Diseases: The New Norm for the United States? *Disease-a-Month* 59: 426–433.

McGee, Glenn. 2006. The Plural of Anecdote Is Not Ambien: Using Case Reports to Get Rewards. *The Scientist* 20 (10): 30.

Nissen, Trygve, and Rolf Wynn. 2012. The Recent History of the Clinical Case Report: A Narrative Review. *JRSM Open* 3: 1–5.

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.

Patsopoulos, Nikolaos A., Apostolos A. Analatos, and John P.A. Ioannidis. 2005. Relative Citation Impact of Various Study Designs in the Health Sciences. *Journal of the American Medical Association* 293: 2362–2366.

Racoosin, Judith A., et al. 2013. New Evidence About an Old Drug: Risk with Codeine After Adenotonsillectomy. *The New England Journal of Medicine* 368: 2155–2157.

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.

Ray, Joel G., et al. 2014. Risk of Overdose and Death Following Codeine Prescription Among Immigrants. *Journal of Epidemiology and Community Health* 68: 1057–1063.

Rison, Richard A., Jennifer Kelly Shepphird, and Michael R. Kidd. 2017. How to Choose the Best Journal for Your Case Report. *Journal of Medical Case Reports* 11: 198.

Smalheiser, Neil R., Weixiang Shao, and Philip S. Yu. 2015. Nuggets: Findings Shared in Multiple Clinical Case Reports. *Journal of the Medical Library Association* 103: 171–176.

Smith, Richard. 2008. Why Do We Need *Cases Journal? Cases Journal* 1: 1.

Smith, D. Max, et al. 2018. Clinical Application of Pharmacogenetics in Pain Management. *Personalized Medicine* 15: 117–126.

TCDC (Taiwan Centers for Disease Control). 2013. https://www.cdc.gov.tw/english/info.aspx?treeid=bc2d4e89b154059b&nowtreeid=ee0a2987cfba3222&tid=49B0A3D6814EEACE. Accessed 20 July 2018.

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.

Terrall, Mary. 2017. Narrative and Natural History in the Eighteenth Century. *Studies in History and Philosophy of Science* 62: 51–64.
Turner, Lucy, et al. 2012. Does Use of the CONSORT Statement Impact the Completeness of Reporting of Randomised Controlled Trials Published in Medical Journals? A Cochrane Review. *Systematic Reviews* 1: 60.
Uyeki, Timothy M., and Nancy J. Cox. 2013. Global Concerns Regarding Novel Influenza A (H7N9) Virus Infections. *The New England Journal of Medicine* 368: 1862–1864.
Van Eck, Nees Jan, et al. 2013. Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLoS One* 8 (4): e62395.
Vandenbroucke, Jan P. 1999. Case Reports in an Evidence-Based World. *Journal of the Royal Society of Medicine* 92: 159–163.
———. 2001. In Defense of Case Reports and Case Series. *Annals of Internal Medicine* 134: 330–334.
Wright, Scott M., and Chrisostomos Kouroukis. 2000. Capturing Zebras: What to Do With a Reportable Case. *Canadian Medical Association Journal* 163: 429–431.
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.
Xu, Wei, et al. 2013. PA-356R Is a Unique Signature of the Avian Influenza A (H7N9) Viruses with Bird-to-Human Transmissibility: Potential Implication for Animal Surveillances. *Journal of Infection* 67: 490–494.
Zhu, Zheng, et al. 2013. Development and Evaluation of a SYBR Green-Based Real Time RT-PCR Assay for Detection of the Emerging Avian Influenza A (H7N9) Virus. *PLoS One* 8 (11): e80028.

Rachel A. Ankeny is Professor of History and Philosophy and Deputy Dean Research in the Faculty of Arts at the University of Adelaide and Honorary Visiting Professor in the College of Social Sciences and International Studies (Philosophy) at the University of Exeter. Her research crosses several fields including history and philosophy of the biological and biomedical sciences, science policy and bioethics and food studies. In the history and philosophy of science, her research focuses on the roles of models and case-based reasoning in science, model organisms, the philosophy of medicine and the history of contemporary life sciences. Her major ongoing projects include the Australian Research Council-funded Discovery Project, "Organisms and Us: How Living Things Help Us to Understand Our World", which is a historical and philosophical exploration of the changing roles and understandings of research with organisms in the twentieth-century and early twenty-first-century science.

Part II
Clustering: Data Ordering and Visualization

From Dirty Data to Tidy Facts: Clustering Practices in Plant Phenomics and Business Cycle Analysis

Marcel Boumans and Sabina Leonelli

Abstract This chapter considers and compares the ways in which two types of data, economic observations and phenotypic data in plant science, are prepared for use as evidence for claims about phenomena such as business cycles and gene-environment interactions. We focus on what we call "cleaning by clustering" procedures, and investigate the principles underpinning this kind of cleaning. These cases illustrate the epistemic significance of preparing data for use as evidence in both the social and natural sciences. At the same time, the comparison points to differences and similarities between data cleaning practices, which are grounded in the characteristics of the objects of interests as well as the conceptual commitments, community standards and research tools used by economics and plant science towards producing and validating claims.

1 Introduction: Preparing Big Data for Analysis

Big data cannot be interpreted without extensive and laborious preparation, including various stages of processing and ordering to make it possible for data to be disseminated and subjected to analysis. Several chapters in this volume – including Halfmann's on sampling in oceanography, Karaca on data acquisition in particle physics and Hoeppe on sharing observations in astronomy - stress the decisive impact that such preparation practices have on the subsequent journeys of data and the use of data as evidence for claims about phenomena. In this chapter we discuss the epistemological significance of yet another practice of data preparation: *data cleaning*, that is the efforts involved in formatting, manipulating and visualising data so that they are sufficiently tractable to be amenable for analysis.

M. Boumans (✉)
Utrecht University School of Economics, Utrecht University, Utrecht, The Netherlands
e-mail: m.j.boumans@uu.nl

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_5

The cleaning strategies that we aim to discuss are not focused on scrubbing and scraping dirt away, but rather on tidying up, sorting and ordering. In everyday life as in data practices, tidying up can be done in a variety of different ways depending on existing habits and future requirements. In what follows, we focus on two strategies for tidying up data which both rely, in different ways, on the clustering of objects into groups. The first strategy is to get rid of smudges and flecks by arranging objects so that unruly bits are less visible, and the eye is drawn to the more orderly – cleaner - parts of the ensemble. We exemplify this strategy through the analysis of data cleaning practices in economics, and specifically in relation to business cycle analysis, where data consist of observations of journalists, business annals, and social and economic statistical time-series. The second strategy is to put everything in boxes and store them some place out of sight, placing labels on each box to be able to retrieve its contents when needed (the more boxes and objects one has, of course, the more complex the labels will need to be).[1] We exemplify this strategy through the analysis of data cleaning practices in biology, and specifically the handling of phenomic data about plants, where data include images and measurements documenting the morphology, physiology and behaviour of organisms and their environments.

We compare a case from the natural sciences (biology) with one from social sciences (economics) in some detail to exemplify the complexity of the research practices involved, which mirrors the complexity of the phenomena under study in both areas. While the conceptual commitments, community standards and research tools used by economics and biology are starkly different, in both cases data cleaning and subsequent analysis involve bringing together voluminous datasets of diverse types and formats, generated by a broad range of heterogeneous sources. The projected value of these data as evidence for scientific claims grows with aggregation: the more data analysts are able to link together and consider as a single body of evidence, the more sophisticated and reliable the resulting insights are expected to be.

The chapter is organised as follows. In the first section, we examine the work required to create meaningful clusters from these forms of big data, and the extent to which data cleaning transforms datasets. In section two we draw on Mary Douglas's seminal analysis of dirt and impurity, in which she argued that cleaning is not about removal but about ordering, to identify a common strategy used by researchers in both cases, which we call *cleaning by clustering*. After discussing this general approach, we note how the specific mechanisms and tools used to enact this strategy differ considerably in the two domains of practice. In economics, cleaning by clustering is largely a question of exercising visual judgement grounded on principles similar to the Gestalt principles, thus arranging data in ways that are *aesthetically appealing and intuitively intelligible* to the analyst. This strategy goes a long way towards facilitating data mining, for instance through the construction of

[1] This approach to cleaning is heavily built on the strategies of packaging, curating and labelling explored by Leonelli (2011, 2016). Contrary to data packaging in her previous studies, however, tidying up is not primarily aimed at making data portable across contexts, but rather at making it possible for data to be analysed and interpreted.

data models that highlight meaningful correlations and direct analysts towards specific interpretations. By the same token, this form of clustering is difficult to undo, leading to a situation where the aesthetic criteria employed to arrange the data are traded off with the ways in which the data could be used as evidence. In plant phenomics, cleaning by clustering is instead guided by the attempt to define a "landscape" for the re-purposing of data: a set of conditions, in other words, through which researchers may be able to re-use data for new goals.[2] The priority in this case is not achieving visual intelligibility alone, but rather the creation of data visualisation and retrieval tools that enable users to *disaggregate data clusters when needed to confront new research questions*. This enables researchers to trace the origin of the relevant data journeys, and evaluate the reliability and appropriateness of every step of "cleaning" in light of novel situations of inquiry within which data may be re-purposed. We are particularly interested in identifying the principles that guide data cleaning activities in these cases, and the conceptual, material and social circumstances within which these principles are grounded and through which they originate. To this aim, in section three we explore the relation between data cleaning practices and how data are subsequently moved and used. Comparing our two cases points to significant differences between data practices, which are grounded in the nature of the objects of interest as well as in the conceptual commitments, community standards and research tools used by economics and plant science towards producing and validating claims. It also points to the difficulties experienced by data analysts in providing general principles of cleanliness with regard to research data, as exemplified by the recent debate around "tidy data" in computational data science, which we discuss in our closing section.

2 Cleaning Data: Empirical Cases from Plant Science and Economics

Our starting point is a close look at two cases of "data cleaning" taken from economics and plant science, respectively. The cases exemplify some of the most sophisticated forms of data processing in each field, aiming to encompass very different types and formats of data coming from a wide variety of sources, which can only be considered as a single body of evidence thanks to laborious processing. The economic case, concerning the generation of quantitative facts about the business cycle at the National Bureau of Economic Research in the 1940s, was selected for two reasons. On the one hand, this post-war research at the NBER is exemplary for many current practices of data preparation in economics, and on the other hand this practice was described so explicitly and in such great detail in a publication, *Measuring Business Cycle* (1946), that it enables and ensures insight and under-

[2] The landscape may include data collection strategies, repositories and visualisation tools enabling researchers to retrieve, compare and analyse data coming from a variety of sources.

standing of this specific clustering practice. The plant science case, concerning the processing of phenotypic data in plant phenomics, constitutes one of the most discussed examples of complex data processing in contemporary biology, with several ongoing debates documenting the rationale and strategies used to make data usable for further analysis. Below, we focus on the discussions surrounding the identification of essential data and related standards ("minimal information") for this kind of research.

2.1 Empirical Case: Measuring Business Cycles

Founded in 1920, the National Bureau of Economic Research (NBER) is a private, non-profit, non-partisan organization dedicated to conducting economic research and to disseminating research findings among academics, public policy makers, and business professionals.[3] The object of the NBER is "to ascertain and to present to the public important economic facts and their interpretation in a scientific and impartial manner" (Burns and Mitchell 1946, p. v). Wesley C. Mitchell, the first director of the NBER till 1945, was well-known for his contributions to the empirical analysis of business cycles.[4] The NBER is not a statistical office or bureau that aims at collecting economic and social data, but instead aims to analyse existing economic and social statistics, in this case to "measure business conditions." These statistics were data of various aspects of economic and business life and came from various different sources. An 11 page long appendix of *Measuring Business Cycle* (1946) list these statistics such as of industrial production, freight, sales, milk used in factory production, transit rides, railway passengers miles, wholesale prices, total income payments, employment, bank debits, electric power production, payrolls, business failures, from organisations such as Federal Reserve, Interstate Commerce Commission, Bureau of Foreign and Domestic Commerce, Railroad Companies, Bureau of Labor Statistics, Chicago Board of Trade, and Bureau of Foreign and Domestic Commerce.

The book *Measuring Business Cycles* (1946) was the result of 20 years of empirical business studies at the Bureau under the supervision of Mitchell. The aim was to identify and establish facts about the business cycles, which could be used to test existing business cycle theories. Burns and Mitchell stated that theoretical work on business cycles was "often highly suggestive; yet rest so much upon simplifying assumptions and is so imperfectly tested for conformity to experience that, for our purposes, the conclusions must serve mainly as hypotheses" (p. 4). At the same time, they observed that "satisfactory tests cannot be made unless hypotheses have been framed with an eye to testing, and unless, observations upon many economic

[3] See the NBER website, http://www.nber.org

[4] See Morgan 1990, pp. 44–56, for a more detailed background of the NBER and Mitchell's approach.

activities have been made in a uniform manner" (p. 4). Although theories were seen as "incomplete in coverage" and "highly suggestive," they were not "put aside" but used "as hypotheses concerning what activities and what relations among them are worth studying. In that way they will be of inestimable value in his factual inquiries" (p. 10). Hence the point of departure for data analysis was not a theory of the business cycle but a very general definition covering commonly accepted characteristics of the business cycles:

> Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own. (Burns and Mitchell 1946, p. 3)

This working definition was supposed to list the observable characteristics of a "distinct species of economic phenomena" (p. 3), that is the business cycle. This definition focused on what should be measured, such as the average duration of the cycle. To achieve this aim, all kinds of questions raised by this definition had first to be answered.[5]

To understand which principles of clustering were used in this case of business cycle measurement, we need to have a closer look at the four implicit assumptions made within this definition. The first assumption is that the cyclical turns of different processes are concentrated around certain points in time. The second assumption is that the business cycle is not a periodic but a recurrent process, a "regularity" that is different from "seasonal variations, random change, and secular trends" (p. 6). Another assumption of the definition is that business cycles run in a continuous round, "no intervals are admitted between one phase and its successor, or between the end of one cycle and the beginning of the next" (p. 7). And the last assumption is the duration of the cycle, somewhere between 1 year and 10 or 12 years.

The main problem for analysts is that business indexes and time series do not show "cyclical patterns" that are "sweeping smoothly upward from depressions to a single peak of prosperity and the declining steadily to a new trough" (p. 7), and so a business cycle has to be identified from an irregular process, where the movements are interrupted by others in the opposite direction, and where one may see double or triple peaks and troughs. What therefore is needed are criteria to identify the characteristics of the business cycles, such as "what reversals in direction mark the end of a cyclical phase" (p. 8). Crucial to our analysis is the fact that such criteria cannot be derived from any (business cycle) theory,[6] but rather they relate to aesthetic

[5] Such as, for instance: How large or small does a nation have to be to have a business cycle, or is it an international phenomenon? How far back in time can business cycles be traced? What is the most appropriate level of aggregation? Which economic activities should be included?

[6] See Bogen and Woodward (1988) for a similar, more general claim about the incompleteness of theories in this respect.

judgements based on visual displays of the data. In other words, certain smooth and simple shapes turn out to be used as tools to process and visualise the data. The approach is based on pattern recognition, described by Burns and Mitchell (1946, p. 8n) as "the source of all true knowledge", but nevertheless it is required to be as objective as possible. Indeed, these criteria are presented as "a 'brake' on an investigator's pattern sense which […] may lead to mischievous fictions" (p. 8n).

Burns and Mitchell emphasized that the cyclical pattern can be seen "only by the eye of the mind" (p. 12). "What we literally observe is not a congeries of economic activities rising and falling in unison, but changes in readings taken from many recording instruments of varying reliability" (p. 14). To "see" the business cycle "in the mind's eye," these recordings have "to be decomposed for our purposes; then one set of components must be put together in a new fashion" (p. 14).

> We conceive business cycles to consist of roughly synchronous movements in many activities. To determine whether this thought symbol represents experience or fantasy, our measures of the cyclical behavior characteristics of many activities must be assembled into the end products of which our definition is the blueprint. In statistical jargon, time-series analysis must be followed by a time-series synthesis. (Burns and Mitchell 1946, p. 17)

The idea is the decomposition of the time series into cyclical, secular, seasonal and random movements, but the "isolation of cyclical fluctuations" was considered to be a "highly uncertain operation" (p. 37), particularly if it is done in a "mechanical manner". The components cannot be segregated without considerable testing and experimenting by skilled technicians. "There is always danger that the statistical operations performed on the original data may lead an investigator to bury real problems and worry about false ones" (p. 38).[7]

Most of the analysis was in the determination of cyclical timing. It had become clear that the data needed to be adjusted for – i.e., cleaned from – seasonal variations "to be more useful in explaining business cycles than would measures made from highly fabricated data" (p. 43). We therefore briefly focus on this aspect of the business cycle analysis, to show how much it was a combination of "hunch and judgment" (p. 44) and mechanical methods, which results were evaluated based on their visual displays.

Two methods were used, one consisted in taking averages of the original figures for each months, which were adjusted for secular trend; and the other entailed taking a 12-month moving average of the original figures, placing each average in the seventh month of shifting 12-month intervals. The rationale for both methods are the assumptions that "random components of a series [will] cancel one another" and that "the process of averaging will tend also to make the cyclical component of a series sum to zero" (p. 47).

When the data was adjusted for seasonal variations, the next problem was the dating of cyclical fluctuations. Therefore the data was plotted upon a semi-logarith-

[7] See Boumans 2015 for a more detailed account of measurement, which sees measurement as a considered balance between mechanical objectivity and expert judgement.

mic chart (typically about 7 feet long) such that the whole record was studied in this graphic form. As far as possible the scales were kept uniform.

The basic criterion for distinguishing the three types of movements, that is the cyclical, secular and erratic movements, was their duration. Secular trends were conceived as drifts that persist in a given direction for a few decades. Erratic movements, the "saw-tooth contour" (p. 57) were supposed to cover no longer than a few months. But even with this basic criterion, the judgments were often difficult:

> When specific cycles are made doubtful by random movements, we smooth the data by moving averages and base judgments upon the curve of moving averages. When the secular trend rises sharply, we allow brief and mild declines to count as contractions of specific cycles. Similarly, when the secular trend falls sharply, brief and mild rises are counted a specific-cycle expansions. (Burns and Mitchell 1946, p. 57)

Once the cycles had been distinguished the NBER researchers proceeded with the dating of the turning points. The idea is to take the highest and lowest points of the plotted curves as the dates of the cyclical turns. But often it is not clear to decide which points these are, for example when erratic movements are prominent in the vicinity of a cyclical turn. Then all kinds of checks or averages have to be considered to arrive at a determination.

> Our methods of determining specific cycles make no pretensions to elegance. Since no fast line separates erratic or episodic movements from specific cycles, or erratic turns from cyclical turns, there is ample opportunity for vagaries of judgment. At times our rules fail to yield a clear-cut decision. At times the members of our statistical staff disagree in their efforts to apply the rules to a given series. Our experience indicates that this difficulty cannot be removed by multiplying rules. (Burns and Mitchell 1946, p. 64)

The judgment is instead based on a consensus of three persons who have worked independently on marking off the cycle. Once arrived at this consensus, the whole process is audited by an "experienced member of the staff" (p. 64) (Fig. 1).

2.2 Empirical Case: Processing and Interoperability Requirements for Imaging Data in Plant Phenomics

Plant phenotyping involves analysing plant trait data with the aim to study development and gene-environment interactions. It emerged in the 1960s with an initial emphasis on quantitative analysis, which was later broadened to imaging data obtained via high-throughput experiments performed in fields, glasshouses, and/or laboratories. Such imaging data, and the accompanying observations about the conditions under which the images were obtained, now constitute the most coveted type of data in this field, with increasingly sophisticated tools being developed for their visualisation and automated analysis. This shift of emphasis on complex data formats proceeded in parallel to the broadening of the term "phenotyping" to include any type of morphological variability within organisms, thus encompassing not only the immediately visible features of organisms, but also (1) features of tissues,

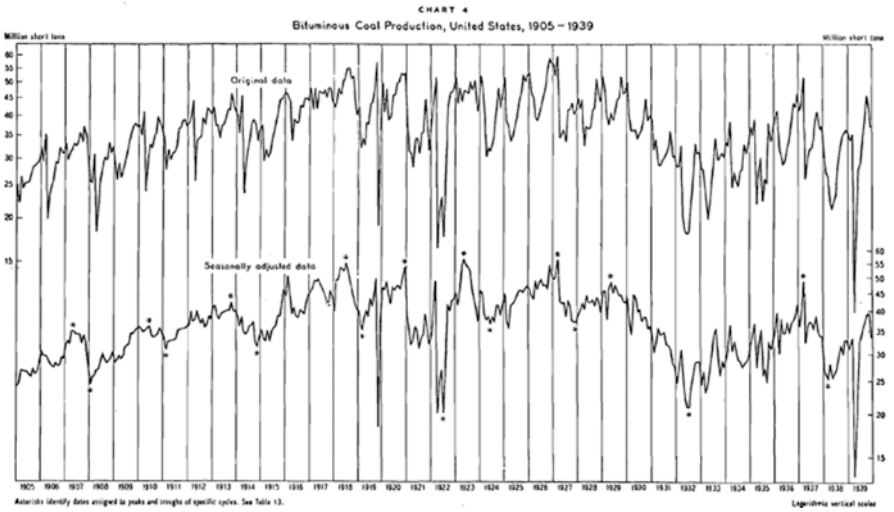


Fig. 1 Example chart of a time series in its original shape and after it has been adjusted for seasonal variation. The adjustment is supposed to facilitate dating of turning points, indicated by the asterisks. (Source: Burns and Mitchell 1946, p. 60, Chart 4)

proteins, metabolic pathways and other aspects only accessible through intervention and specialised imaging techniques; and (2) the ways in which such features vary across environments that range from laboratories to glasshouses, field trials and the "wild" – which involves collecting data on the soil, climate, other organisms and microbiome with which plants interact. In the words of prominent contributors to the field, phenotyping – also called "phenomics" – "broadened its focus from the initial characterization of single-plant traits in controlled conditions towards 'real-life' applications of robust field techniques in plant plots and canopies" (Walter et al. 2015). Importantly for our analysis, this shift in the conceptualisation of phenotypic traits made them much less obviously identifiable as concrete descriptors. Collecting data about the size of a leaf or the structure of a metabolic pathway is not simply a matter of observation, but rather is informed by a rich conceptual apparatus defining what counts as leaf surface and metabolism. Thus, just as much as business cycles are no pure theoretical constructs, phenotypes are no 'brute facts' about the world: in both cases, empirical and theoretical considerations remain firmly intertwined, and affect researchers' approach to data processing and interpretation.

A key component of contemporary phenomics, and the reason why it is regarded as generating knowledge that can underpin and guide agricultural production, is a holistic characterisation of plant performance, which involves the employment of several investigative methods and the generation and analysis of a wide variety of data types. These include, for instance, multispectral and thermographic imaging of plant growth, which is often carried out within so-called "smart glasshouses" in an automated fashion (by robots or conveyor belts that transport the plants to various imaging chambers, multiple times per day, over an extended period of time).

Photographs and measurements are produced that document how plants develop, how their leaves and roots change, and how they respond to external stimuli.

Cleaning such images for analysis involves judgements around the quality and resolution of the photograph, the lighting and background conditions, the position in which plants have been captured and the extent and clarity to which relevant leaves and roots show in the picture. The quantity of images generated through any one experiment makes it hard for researchers to do such work manually, and yet it is hard to fully automate due to the large amount of know-how and theoretical commitments involved in judging image quality – encompassing familiarity with the plants and their full life-cycle, expectations around how plants may respond to environmental conditions, existing conceptualisations of plant development and growth, and assumptions around which environmental and morphological elements need to be valued and prioritised over others.

Another popular type of phenomic data is acquired through top-view imaging of the plant canopy in the field, which can be performed by humans in helicopters, robots or remote-controlled drones. These photographs can be analysed to measure leaf greenness, via tools such as the Normalised Difference Vegetation Index, or plant biomass and growth in the area under scrutiny. Again, while some basic parameters can be established for what counts as a "bad image" and which elements of each image may be classified as "noise", cleaning such images involves expert assessment based on detailed knowledge of the characteristics and patterns of growth of the plants at hand. An example (Fig. 2) is an imaging study of soy-bean fields to determine patterns of growth, in which researchers prepare images for further analysis (in their own words, "classify" the images) through models that are manually trained at every step to respond to the traits of interest in the beans (Xavier et al. 2017).

Given the sensitivity of phenomic studies to local conditions and the conceptual preferences and know-how of specific researchers, consensus around how to clean data is hard to achieve. Nevertheless, such consensus is highly valued and sought for, as it enables researchers to compare results obtained across species, field types and environmental conditions. One attempt towards establishing general standards for data collection and processing is the Minimal Information About Plant Phenotypic Experiments, or MIAPPE. MIAPPE is part of a broader set of "minimal information about data" movement now recognized and coordinated by the FAIR sharing international initiative for reusable data curation.[8] This is an attempt to standardize the practices and variables required to tidy up data formatting and analysis enough to make data searchable, visualisable and retrievable through digital means. The idea of "minimal" information is meant to foster an evaluation of which contextual information is most important to data interpretation, resulting in as small a set

[8] See https://fairsharing.org/collection/MIBBI. Among the first incarnations of the movement, and now highly successful standards in their own right, were the Minimal Information About a Microarray Experiment, or MIAME (Rogers and Cambrosio 2007) and the Minimal Information for Biological and Biomedical Investigations, or MIBBI (http://www.nature.com/nbt/journal/v26/n8/full/nbt.1411.html)

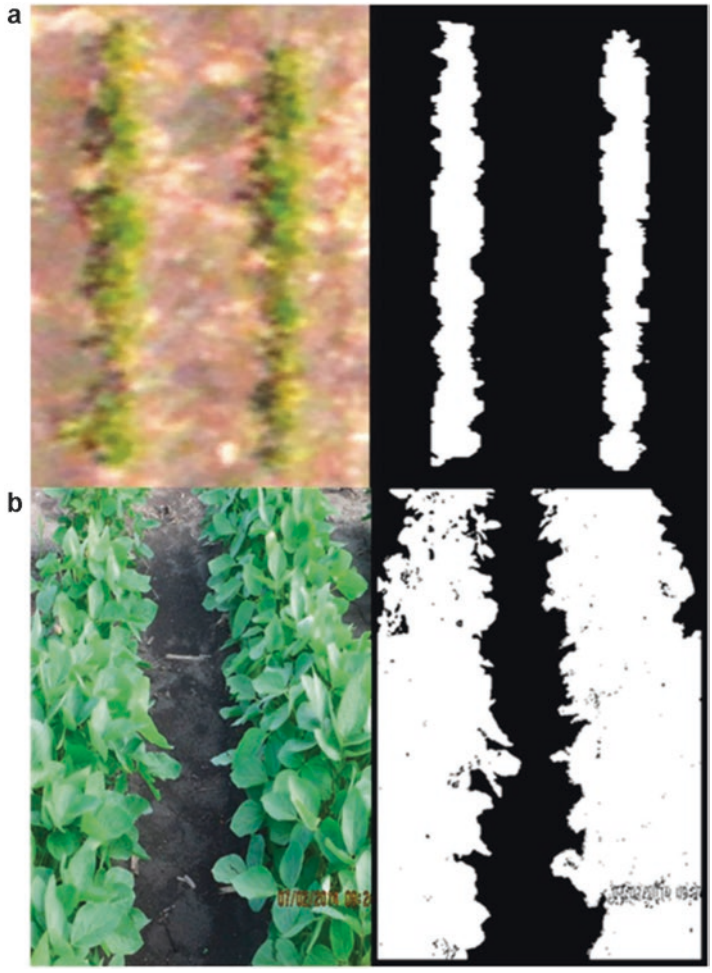


Fig. 2 Example imagery of a single plot of soy-bean canopy, used to calculate a percentage canopy coverage on a given sampling date. (**a**, **b**) From aerial (above; **a**) or ground (below; **b**) platforms, with raw (left) and classified (right) imagery. (Source: http://www.genetics.org/content/206/2/1081)

as possible of metadata that researchers view as essential to phenotypic data reuse. Somewhat paradoxically, within MIAPPE this aspiration towards minimal information is accompanied by the wish to lose as little information as possible about the original format of the data, the circumstances under which they were generated, and the ways in which they were processed since. This is because the specificity of the provenance and formatting of data in each case is regarded as highly valuable by the plant scientists using such data for their own research, a requirement that researchers and engineers involved in the development of MIAPPE take seriously: "We had to allow for differences that occur between particular types of plant experiments, e.g. performed in different growth facilities. This is reflected in a varying set of attributes recommended in MIAPPE" (Ćwiek-Kupczyńska et al. 2016). Indeed, the

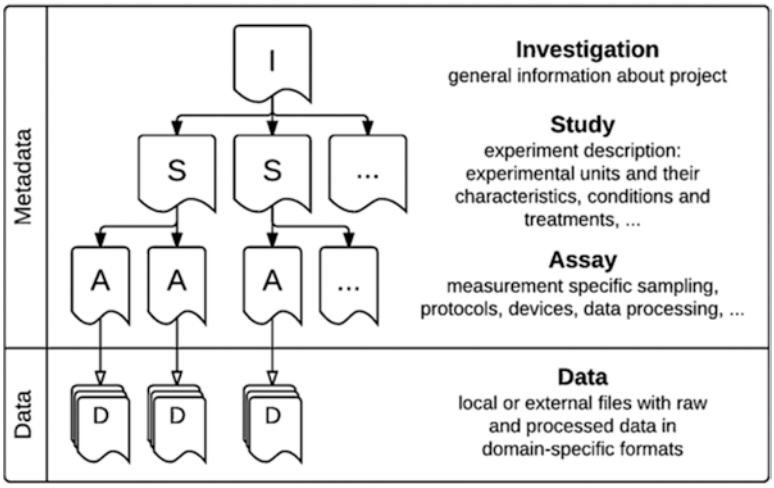
list of attributes to be reported to MIAPPE involves over 80 items, which can extend to over a hundred depending on the field conditions. The basic categories are themselves relatively broad, encompassing general metadata, timing and location, biosources, environment, treatments, experimental design, observed variables and as much information as possible on sample collection, processing and management – a far cry from the minimalism that the MIAPPE criteria were expected to exemplify.

It is useful to consider a couple of the simplest examples from this list. Take for instance the item "location and timing of an experiment". Here MIAPPE developers note that "depending on the nature of the study and scientific objectives, different initial time points might be crucial—sowing date or transfer date, treatment application time, etc. The duration of particular stages is also important." (Ćwiek-Kupczyńska et al. 2016, p. 3). Thus, even a relatively straightforward measure such as the time of the experiment turns out to be a complex and context-dependent issue, for which it is hard to establish any hard and fast boundaries to ensure comparability across different experiments.[9] Another example is item "biosource" – that is, the identification of the plant material at hand. Here MIAPPE recommends using at least two attributes, one consisting of the species name as in standard taxonomic classifications, and the other consisting of the "infraspecific" name, pointing to the specific variant, accession or line in question. Complications arise due to the types and history of the plant materials at hand. While the taxonomy of plant species is, though controversial, subject to international standards, the identification and classification of sub-species variants is highly decentralised and context-dependent, with no overarching agreement around classification and often not even a clear awareness of the differences between local systems. For example the varieties of the plant *Manihot esculenta*, whose root cassava is a key crop in West Africa and South America, are often defined by the different ways in which local breeders value specific traits (like the humidity and colour of the root) when processing the plant for food production. Aware of this fact, the authors point to the importance of referencing any "public collection of names", and/or a specific experimental station or genebank in which the variant may be stored and or the seeds may have been sourced, and to which they can be physically traced. There are international identification systems for crops of commercial interest, such as the FAO/Bioversity Multi-Crop Passport Descriptors, but these do not cover all possible variants. The ways in which data about specific attributes are structured in MIAPPE conform to the ISA-Tab standards for data ordering, which is widely adopted in biology and looks as follows (Table 1).

This table aims to impose a clear conceptual ordering of the data, resulting in their presentation in a format and structure that is amenable to computational analysis. At the same time, the application of the ISA-Tab standard to the specific case of phenotyping is complex, as demonstrated by challenges encountered in developing the so-called "ISA-Tab Phenotyping Configuration". This consists of a standard Investigation file, a Phenotyping Assay file describing phenotypic procedures and observed variables (according to the dozens of attributes identified by MIAPPE,

[9] See Leonelli (2018) for an analysis of data time and its significance particularly within experiments.

Table 1 The structure of an ISA-Tab dataset



Source: Ćwiek-Kupczyńska et al. (2016)

such as location and biosources), and three versions of a Study file: one called "basic study" and consisting of a default general description of all plant experiments, which needs to be extended by added recommended MIAPPE attributes as applicable to the specific case[10]; and two extensions called "field" and "greenhouse" studies, featuring specific attributes for growth facilities and environmental information (Ćwiek-Kupczyńska et al. 2016, p. 8) (Table 2).

Notably, despite the drive towards comparability, MIAPPE emphasizes the need to capture any data format in use within the relevant scientific communities, rather than attempting to impose overarching standards on the ways in which data are produced: "in our implementation of MIAPPE, we do not restrict the format of the raw data in any way; it can be any custom, platform- or device- specific format, including texts, images, binary data, etc." (Ćwiek-Kupczyńska et al. 2016, p. 11). At the same time, MIAPPE requires that information about data provenance (metadata) is reported in ways that are comprehensive and retrievable by later data users. The most stringent MIAPPE instructions concern how to organize and display such metadata:

> If there is no description, the Derived Data File should be a standard, plain tab-separated sample-by-variable matrix. Its first column should contain (in the simplest situation) values from the Assay Name column in the Assay file, and the rest of the columns provide values for all variables. The names of those columns should correspond to the values in the Variable

[10] In practice, it can be also used when very little is known about the origin of observations, e.g. for simple, external or legacy phenotypic datasets that should be formatted as ISA-Tab, without the ambition to satisfy the MIAPPE recommendations.

Table 2 Illustration of what the basic ISA-TAB fields correspond to when implemented by plant scientists in the field and in the greenhouse, respectively

Basic	Field	Greenhouse
Source name	Source name	Source name
Characteristics[organism]	Characteristics[organism]	Characteristics[organism]
Characteristics[Infraspecific name]	Characteristics[Infraspecific name]	Characteristics[Infraspecific name]
Characteristics[seed origin]	Characteristics[seed origin]	Characteristics[seed origin]
Characteristics[study start]	Characteristics[study start]	Characteristics[study start]
Characteristics[study duration]	Characteristics[study duration	Characteristics[study duration
Characteristics[growth facility]	Characteristics[growth facility]	Characteristics[growth facility]
Characteristics[geographic location]	Characteristics[geographic location]	Characteristics[geographic location]
	Protocol REF[rooting]	Protocol REF[rooting]
	Parameter value[rooting medium]	Parameter value[rooting medium]
		Parameter value[container type]
		Parameter value[container volume]
	Parameter value[plot size]	Parameter value[container dimension]
	Unit	Unit
	Parameter value[sowing density]	Parameter value[number of plants per container]
	Parameter value[pH]	Parameter value[pH]
	Protocol REF[aerial conditions]	Protocol REF[aerial conditions]
	Parameter value[air humidity]	Parameter value[air humidity]
	Parameter value[daily photon flux]	Parameter value[daily photon flux]
	Parameter value[length of light period]	Parameter value[length of light period]
	Parameter value[day temperature]	Parameter value[day temperature]
	Parameter value[night temperature]	Parameter value[night temperature]
	Protocol REF[nutrition]	Protocol REF[nutrition]
	Parameter value[N before fertilisation]	Parameter value[N before fertilisation]
	Parameter value[type of fertiliser]	Parameter value[type of fertiliser]
	Parameter value[amount of fertiliser]	Parameter value[amount of fertiliser]

(continued)

Table 2 (continued)

Basic	Field	Greenhouse
	Protocol REF[watering]	Protocol REF[watering]
	Parameter value[irrigation type]	Parameter value[irrigation type]
	Parameter value[volume]	Parameter value[volume]
	Parameter value[frequency]	Parameter value[frequency]
	Protocol REF[sampling]	Protocol REF[sampling]
	Parameter value[experimental unit]	Parameter value[experimental unit]
Sample name	Sample name	Sample name

Source: Ćwiek-Kupczyńska et al. (2016)

> ID column in the Trait Definition File […]. So, a default derived data format is an "Assay Name × Variable" matrix of observations, that can be quantitative or qualitative. An extension of the above rule governing the format of the Derived Data File is possible by using values from another "data node" column (e.g. Source Name, Sample Name, Extract Name, etc.) as unique identifiers of the rows in the table with the associated observations. (Ćwiek-Kupczyńska et al. 2016, p. 12)

This is because such ordering is what enables researchers to initiate comparisons:

> we can provide separate data files with measurements taken for different observational units, e.g., morphological traits like "height" and "number of leaves" can be assigned to the whole plant, whereas physiological traits can be restricted to samples taken from particular leaf of a plant. Also conveying data aggregated over "data nodes" is possible in this way. (Ćwiek-Kupczyńska et al. 2016, p. 12)

Despite the attention placed by MIAPPE developers on the variability and contextuality of data and related preparation procedures, applying MIAPPE criteria to the processing of data in the field remains a big challenge. As a concrete example, we take the data processing performed at a leading station for the collection of phenomics data in the UK. The North Wyke Farm Platform is a research facility built around a working farm in Devon, in which researchers can study the interactions between climate, soil, animals, plants and microbiota in as close a setting as possible to real farming. The whole area is full of sensors and measurement devices, which collect data at regular intervals (15 minutes) about a variety of aspects of the farm: temperature, soil composition, humidity and rainfall, etc. The sensors are calibrated and checked in 15 huts ("monitoring cabins") positioned around the fields, and the data produced is sent wirelessly to the central computing facility based in the manor house, where researchers proceed to prepare the data, cluster them and store/disseminate them through a database. There are also three meteorological stations that move around the fields. An important activity besides collecting numerical measurements is the collection of samples (of soil, air, water, insects and plants),

which are acquired manually (e.g. manual sampling device for soil), prepared and stored in fridges at various temperatures).[11]

Researchers interviewed[12] in North Wyke have stressed that the data collected by the Farm Platform are not yet being interpreted: this will only be possible when enough longitudinal data are collected over the course of the next few years.[13] This makes the task of data cleaning ever more important, since the researchers' main task at the moment is to make sure that the data collected is reliable and clustered and displayed in ways that will facilitate further analysis, and prove informative for interested farmers. Cleaning the data means first of all making them comparable and consistent with other datasets generated within the Farm, an arduous task given the variety of measurements taken and images collected. Equally important is to make sure that such data would be comparable and consistent with other phenomics data from outside North Wyke. While researchers attempt to follow criteria similar to those formulated by MIAPPE, the variability in the interpretation of the attributes and values is a serious threat to automated mining and comparison among the data. Researchers aim to enable analysis in the future, but caution against any automated search. They also emphasize how the power of this evidence is in the meta-data, the information that enables researchers to contextualize the findings and evaluate their significance in relation to findings from other locations enacting different epistemic cultures and methods.

3 Cleaning by Clustering: The Principles Underpinning Data Cleaning Practices

Renowned anthropologist Mary Douglas provided an important argument for understanding the process of cleaning as being not about removal, but about ordering. According to Douglas (2002), dirt is essentially disorder: "There is no such thing as absolute dirt: it exists in the eye of the beholder. […] Dirt offends against order. Eliminating it is not a negative movement, but a positive effort to organize the envi-

[11] The facility attracts researchers from different communities and disciplines seeking to develop sustainable agriculture and ruminant production systems http://www.nature.com/news/agriculture-steps-to-sustainable-livestock-1.14796. It is the only currently functioning facility of its kind world-wide, and the Global Farm Platform http://www.globalfarmplatform.org/ was born to attempt to export this model and initiate similar sites elsewhere.

[12] Interviews were carried out by Leonelli in January 2016. A subset of the interviews, which interviewees consented to release in an open access format, is available here: https://zenodo.org/communities/datastudies/?page=1&size=20

[13] North Wyke researchers are also conducting short-term studies in which the data are used as evidence for claims about phenomena. Examples include research on replacing nitrogen as fertilizer, the use of plants to manage soil and water during floods, shifts in soil biota as land use changes, and the modelling of grassland production systems. At the same time, researchers only take up research that will not "distort" on-going, long-term data collection by forcing them to "clean" data with too narrow a set of epistemic goals in mind.

ronment" (p. 2). In chasing dirt when tidying we are "positively re-ordering our environment, making it conform to an idea […] it is a creative moment, an attempt to relate form to function, to make unity of experience" (p. 3). Douglas emphasizes that the identification of dirt should not be considered as a unique, isolated event. "Where there is dirt there is system. Dirt is the by-product of a systematic ordering and classification of matter, in so far as ordering involves rejecting inappropriate elements" (p. 44). Cleaning is the reaction which condemns any object or idea likely to confuse or contradict cherished classifications, thus "reducing dissonance" (Douglas 2002, p. 340). Thus cleaning is part of the epistemological activity of systematization, such as ordering and classification. Douglas distinguishes two phases to such systematization practices:

> In the course of any imposing of order, the attitude to rejecting bits and pieces of dirt goes through two stages. First they are recognisably out of place, a threat to good order, and so are regarded as objectionable and vigorously brushed away. At this stage they have some identity: they can be seen to be unwanted bits of whatever it was they came from, hair or food or wrappings. This is the stage at which they are dangerous; their half-identity still clings to them and the clarity of the scene in which they obtrude is impaired by their presence. But a long process of pulverizing, dissolving and rotting awaits any physical things that have been recognized as dirt. In the end, all identity is gone. The origin of the various bits and pieces is lost and they have entered into the mass of common rubbish. It is unpleasant to poke about in the refuse to try to recover anything, for this revives identity. So long as identity is absent, rubbish is not dangerous. It does not even create ambiguous perceptions since it clearly belongs in a defined place, a rubbish heap of one kind or another. (Douglas 2002, pp. 197-8)

The stage of total disintegration is the stage in which dirt has become undifferentiated. Then a cycle has been completed, resulting in an order that is either continuous with what was there before the cleaning or created by the process of cleaning itself.

Drawing on Douglas's analysis, we argue that in both of our cases researchers adopt the same broad strategy for data cleaning: they *clean by clustering*. Cleaning is a way to impose order and intelligibility on a dataset, by identifying categories and typologies for classification, models and algorithms through which data can be filtered and selected, and/or tools through which data can be displayed and organised so as to enable further analysis and interpretation.

The specific mechanisms and tools used to enact this strategy, however, differ considerably across our cases, revealing a divergence in the heuristic principles used to guide and motivate the cleaning strategies, and the extent to which whatever is neutralized from a given stage of data cleaning is regarded as "unwanted bits" with "some half-identity clinging to them", or as dirt where "identity is absent".

In our economics case, clustering involves looking for cyclical patterns through visual judgement. To understand the heuristic behind this cleaning procedure, it is useful to discuss briefly Gestalt theory first. Gestalt psychologists study perceptual organization: "how all the bits and pieces of visual information are structured into larger units of perceived objects and their interrelations" (Palmer 1999, p. 255). A "naïve realist" explanation of this organization could be that this organization simply reflects the structure of the external world. A problem with this explanation is that the visual system does not have direct access to how the environment is structured, it has only access to the image projected onto the retina, the "array of light that falls on the

retinal mosaic" (p. 257). This optic array allows for an infinite variety of possible organizations. The question therefore is how the visual system picks out one of them. To answer this question Max Wertheimer, one of the founders of Gestalt psychology, studied the stimulus factors that affect perceptual grouping: "how various elements in a complex display are perceived as 'going together' in one's perceptual experience" (Palmer 1999, p. 257). The theoretical approach of the Gestalt psychologists is that perceptual organization is grounded in the wish to maximize simplicity, or equivalently, minimize complexity. They called this hypothesis the principle of Prägnanz, today also called the minimum principle. It states that the percept will be as good as the prevailing conditions allow. The term "good" refer to the degree of figural simplicity or regularity, and the prevailing conditions refer to the structure of the current stimulus image (Palmer 1999, p. 289). The Gestalt psychologists saw symmetry as a global property with which figural goodness could be analysed.

The organising Gestalt in the case of the NBER business cycle analysis was a cyclical pattern, such as the Fig. 3. By taking averages, whether weighted or not (which is an act of clustering), one aimed at reducing the noise in the observations as much as possible. Because it is not possible to tidy up by a kind of physical intervention on some physical material, the tidying up is not done by removal but by clustering in such a way that the cluster itself is "cleaner" than the individual data. The principle of Prägnanz that was implicitly applied and was the underlying goal of the procedures is an as simple as possible shaped cycle with clear peaks and troughs.

In the economic case, the original data end up as what Douglas classified as undifferentiated dirt – that is, as objects that are forever disconnected from their original source.

> [T]hese symbols are derived by extensive technical operations from symbolic records kept for practical ends, or combinations of such records. We are, in truth, transmuting actual experience in the workaday world into something new and strange […]. (Burns and Mitchell 1946, p. 17)

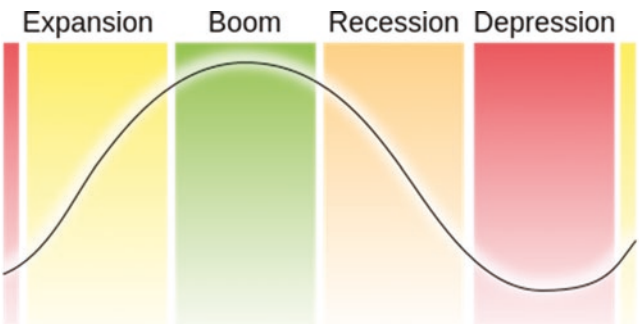


Fig. 3 Example of a "typical" business cycle pattern (Source: https://seekingalpha.com/article/2716385-investing-in-business-cycles)

In other words, the process of cleaning by clustering in this case transforms a large quantity of objects that were previously identified as data into objects that have new evidential value, but are no longer available or retrievable as sources of information about the contexts from which they were inferred.[14] At the same time, it is important to note that the resulting records do not completely fail to provide an identity to the discarded objects. Keeping some traces of the original time series is relevant if only to verify that results are not artificial products of spurious cyclical patterns. The visualisations of original times and the adjusted one should show sufficient similarity. "A common method of judging the goodness of [an] adjustment is to see whether the adjusted figures show similar movements in successive years" (Burns and Mitchell 1946, p. 54).

In plant phenomics, clustering instead involves defining a "landscape" for the potential re-contextualisation of data. The starting assumption is that phenomics data, in all their richness, variability and multiplicity of features, may be used for all sorts of research goals, ranging from studies of irrigation systems to investigations of plant growth and nutrition (as in the case of North Wyke data). Therefore the priority for researchers is not the visual intelligibility of a particular way of arranging data, but rather the creation of categorisations that facilitate the *disaggregation* of data clusters when needed by the inquiry at hand. In other words, researchers want to retain the ability to trace the origin of the relevant data journeys, and evaluate the adequacy of every step of data cleaning towards producing reliable evidence for new research questions. Key heuristic principles here are: *accuracy*, in the sense of being as faithful as possible to the specific characteristics of the research objects at hand; and *traceability* of data sources, in the sense of making sure that prospective data analysts have what they need to assess the quality of the data and, if needed, process them differently (which typically includes as extensive an access as possible to metadata).

This approach is hard to compare to the application of Gestalt principles, because those are focusing on visual appearance and presentation, while phenomics practices of cleaning by clustering focus on interpretability and the potential to disaggregate existing data clusters. Nevertheless, like the economics case, this is in striking opposition to common sense interpretations of the metaphors of "cleaning" and "dirt" that focus on the removal of blatantly unwanted items. Both in biology and economics "dirt" may (and often does) contain useful information, which needs to be ordered so as to be retrievable depending on the interests of the prospective analyst. The original datasets and related metadata never fully become undifferentiated dirt as in Douglas's analysis. Rather, researchers attempt to "cling on to their half-identity", in Douglas's terms, thus leaving open the option for these objects to be re-identified as data and fully reinstated as significant sources of evidence for a claim. The main difference between the two fields is that economic data have lost more of the identity of their original data than is the case in phenomics. While in plant phenomics accuracy and traceability are leading, in economics accuracy has to be balanced with Prägnanz, and traceability is not required.

[14] This interpretation assumes a relational account of data epistemology, as outlined in Leonelli (2016) and in the introduction to this volume.

4 Comparing Heuristics across Research Communities in Natural and Social Sciences

Economic data are processed in ways that make them much more computationally tractable than phenomics data due to their numerical format. Economic data are thus better amenable to aggregation and analysis in comparison to many other data types, which potentially expands their scope for linkage and aggregation with other datasets but also limits the power of investigators to contextualise and situate the data in relation to their origin. In this case, cleaning by clustering is a cumulative process, in which the bulk of "raw" data is replaced by a smaller set of business-cycle "facts" through the exercise of visual principles.[15] As a result, analysts working at later stages of these data journeys are left mostly with data models that conform to specific criteria and are best used to address a narrow set of questions, in conformity with the principles and assumptions made while preparing them for analysis. The original "raw" data are no longer accessible, having been "cleaned out" in the data visualisations.

By contrast, phenomics data remain more difficult to analyse through computational tools, and can only be compared and linked with other datasets by employing case-by-case adjustments. They are so heterogeneous, and their ordering into clusters so pluralistic and open to multiple interpretations, that additional processing is needed every time researchers re-use them for a specific project. When considering data on biosource as discussed in section two, for instance, researchers need to double-check what assumptions have been made about the taxonomy of plant varieties when ordering plant traits into groups. At the same time, the richness of data formats and of the information that they carry make them useful evidence for a large variety of inquiries, and makes it easier to interrogate their reliability and quality in relation to different research conditions and aims. Phenomics data can potentially be used to answer many research questions. Cleaning by clustering in this case is not a cumulative process: it is crucial for researchers to lose as few data and metadata as possible, as one never knows what will turn out to be important later.

It has been frequently observed that big data aggregation is often accompanied by loss of contextual information (metadata).[16] While in both of our cases the role and ordering of contextual information plays a key role in the process of cleaning by clustering, the principles associated to handling such contextual information are considerably different. In economics, metadata become increasingly less relevant: the principles guiding data ordering and clustering are those of Prägnanz. In plant phenomics, metadata never cease to be relevant, as the principles guiding ordering and clustering are those of accuracy and traceability.

[15] Facts about phenomena, in the sense of Bogen and Woodward 1988.

[16] Lawrence Busch (2014, also discussed in Mittlestand and Floridi 2016) lists several reasons for this, including: Lossiness (lose aspects of the phenomena studied); Drift (phenomena change over time, but data representing them do not); Distancing (distance from phenomenon facilitates identification of patterns); Layering (reducing phenomena to set of variables, e.g. in Tidy data); Errors; Standards; Disproportionality; Amplification/reduction; Narratives.

Assumptions made about the nature of the phenomena at hand (respectively, plant morphology and business cycles) may seem to have a significant impact on the type of techniques and principles enacted by researchers. For instance, the proponents of MIAPPE explicitly note that

> we are fully aware that MIAPPE suggests a description of the experiment that is rather extended in comparison to current practices. Hence, although we think that all of the attributes in Table 1 are needed to adequately describe each dataset, we accept that, in practice, the full complement of information may not be possible to collect, or might be unavailable to the person building the dataset. Therefore, we have selected and marked those descriptors deemed absolutely essential. (Ćwiek-Kupczyńska et al. 2016, 7)

Remarkably, their "absolutely essential" list of traits still comprises 35 attributes, a skinnier list than the original list of over 80 attributes (ranging from 70 to over a hundred depending on growth conditions and type of environment/soil), but still daunting in its richness.

We do not think that these differences should be viewed simply as a measure of the difference between studying plants and studying economic conditions. Both types of phenomena are highly complex in their own ways, and arguably economic behaviour is even more difficult to reduce to a simple set of variables. A more plausible explanation lies in the methods and commitments characterizing the two fields of inquiry. Economics, business cycle analysis in particular, is a highly generalist field but it is not holistic: research focuses on analysing the business cycle as an isolated phenomenon. By contrast, plant phenomics favours a holistic approach, emphasising the complexity of the interrelated processes through which plant morphology is constituted (see Fig. 4 and also Leonelli 2016, ch. 6).

Furthermore, plant phenomics has no pretension to achieve a "complete representation" (or complete knowledge) of the plant systems it analyses, precisely because of their daunting complexity and the fact that so little is as yet known about them. Thus, any model proposed in plant science to analyse a phenomenon will be limited in scope, and need to be complemented by several others to provide a more comprehensive picture of the phenomena for specific investigative goals. Related to this, mathematical and statistical modelling – while of course strongly present in this work – are not always the primary or main tool of analysis; and their role is not always one of data validation, they are also employed as tools to order and display the data at hand in ways that may help analysis (Leonelli 2019).

5 Conclusions

Our analysis points to the difficulties experienced by analysts in providing general principles of cleanliness with regard to research data. This is nicely exemplified when considering the ongoing debate around the identification and application of overarching "tidy data principles" in contemporary data science, which seeks to outline criteria for "cleaning" and structuring data so as to make them amenable to computational analysis (Wickham 2014). Within this framework, data processing is

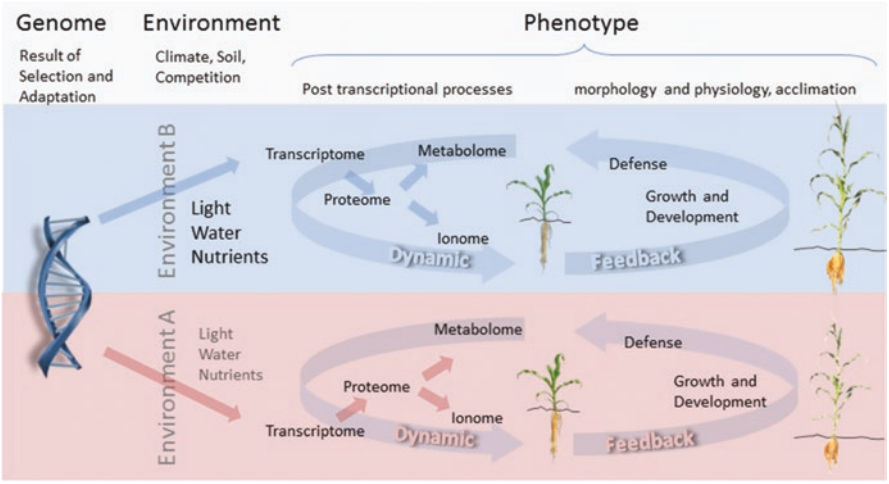


Fig. 4 Representation of the conceptual landscape for phenomics, taken from a seminal review paper from Walter et al. (2015)

conceptualised as consisting of four stages: (1) import data; (2) tidy data; (3) transform/visualise/model data; (4) communicate data. Tidy datasets are defined as providing "a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)" (Wickham 2014, 2), thus helping to prepare data for visualisation and modelling. This literature does not shy away from data diversity, and recognises that data "tidiness" comes in a variety of different flavours depending on the field and goals of inquiry, the statistical and computational tools available (which are referred to as "tidy tools", p. 20), and the cognitive preferences of investigators. The starting point for this work is to acknowledge that determining what are observations and what are variables is relatively easy in the case of specific datasets, but that such a distinction is hard to define in general terms, also because of the diversity often characterising data sources and levels of abstraction. At the same time, an attempt is made to discuss tools through which "messy data" can be "tidied up", so as to be ready for computational analysis. An example is the activity of "melting", which consists of stacking datasets by turning columns of numbers into rows. Another is "string splitting", which involves splitting the columns of any given data table into different variables. Furthermore, a series of "tidy tools" are presented, such as data aggregation, filtering, visualisation and statistical modelling, whose common aim is to "take untidy datasets as input and return tidy datasets as outputs" (p. 12). All these strategies for cleanliness are meant to "make analysis easier by easing the transitions between manipulation, visualisation and modelling" (p. 15).

This approach to data cleaning aligns nicely with the strategy that we have called "cleaning by clustering". At the same time, our reading of Douglas's work on dirt provides a conceptual framework and rationale for this approach. It makes it clear that cleanliness is not a matter of removing unnecessary items, "noise" or "mess"

from somehow predefined "meaningful datasets", thus assuming that (1) there is a "best way" to order data regardless of the research aims of specific investigations; and (2) what researchers should consider as reliable and veritable data need to be uncovered and separated from "meaningless noise". By contrast, we propose to view data cleanliness as a process of ordering data into clusters, which runs in parallel with situated attempts to assign meaning to data in relation to specific research questions and goals. Thus cleaning can take a variety of different forms – and result in very different ideas of "what counts as data" – depending on the assumptions, commitments and circumstances of the research projects at hand. Moreover, our cases have shown that the above mentioned four stages of data analysis are actually four aspects of one process of data interpretation which cannot be separated from each other.

References

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

Boumans, Marcel. 2015. *Science Outside the Laboratory*. Oxford: Oxford University Press.

Burns, Arthur F., and Wesley C. Mitchell. 1946. *Measuring Business Cycles*. New York: National Bureau of Economic Research.

Busch, Lawrence. 2014. Big Data, Big Questions | A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large Scale Data Sets. *International Journal of Communication* 8 (0): 18. https://doi.org/10.1007/SpringerReference_22340.

Ćwiek-Kupczyńska, Hanna, Thomas Altmann, Daniel Arend, Elizabeth Arnaud, Dijun Chen, Guillaume Cornut, Fabio Fiorani, et al. 2016. Measures for Interoperability of Phenotypic Data: Minimum Information Requirements and Formatting. *Plant Methods* 12 (1): Bio Med Central: 44. https://doi.org/10.1186/s13007-016-0144-4.

Douglas, Mary. 2002[1966]. *Purity and Danger. An Analysis of the Concept of Pollution and Taboo*. London/New York: Routledge.

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.

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.

Leonelli, Sabina. 2011. Packaging Small Facts for Re-Use: Databases in Model Organism Biology. In *How Well Do Facts Travel?* ed. P. Howlett and M.S. Morgan, 325–348. Cambridge: Cambridge University Press.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

———. 2018. The Time of Data: Time-Scales of Data Use in the Life Sciences. *Philosophy of Science* 85 (5): 741–754.

———. 2019. What Distinguishes Data from Models? *European Journal for the Philosophy of Science* 9: 22.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Morgan, Mary S. 1990. *The History of Econometric Ideas*. Cambridge, MA: Cambridge University Press.

Palmer, Stephen E. 1999. *Vision Science*. Cambridge, MA: MIT Press.
Rogers, Susan, and Alberto Cambrosio. 2007. Making a New Technology Work: The Standardization and Regulation of Microarrays. *Journal of Biology* 80: 165–178.
Walter, Achim, Frank Liebisch, and Andreas Hund. 2015. Plant Phenotyping: From Bean Weighing to Image Analysis. *Plant Methods* 11 (1): 14. https://doi.org/10.1186/s13007-015-0056-8.
Wickham, Hadley. 2014. Tidy Data. *Journal of Statistical Software* 59 (10). https://doi.org/10.18637/jss.v059.i10.
Xavier, Alencar, Benjamin Hall, Anthony A. Hearst, Keith A. Cherkauer, and Katy M. Rainey. 2017. Genetic Architecture of Phenomic-Enabled Canopy Coverage in *Glycine Max*. *Genetics* 206 (2): 1081–1089. https://doi.org/10.1534/genetics.116.198713.

Marcel Boumans is Pierson Professor of History of Economics at Utrecht University. His main research focus is on understanding empirical research practices in social science from a combined historical and philosophy perspective. He is particularly interested in the practices of measurement and modelling and the role of mathematics in social science. Because models are not complete as sources of knowledge for sciences outside the laboratory, additional expert judgements are needed. This is the topic of his most recent monograph *Science Outside the Laboratory* (OUP, 2015). His current research project "Vision and Visualisation" focuses on exploring how expert judgments (views) are made and how they could be validated, particularly in those research practices where visualizations are made or used. The first outcome of this project is "Graph-Based Inductive Reasoning" published in *Studies in History and Philosophy of Science* (2016).

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.

The Datum in Context: Measuring Frameworks, Data Series and the Journeys of Individual Datums

Mary S. Morgan

Abstract Studying a social whole such as a city, an economy, or a society, requires the construction of 'group data sets' where the group is made up of a number of individual data series, each one in turn made up of a string of individual data points or datums. This group set forms the most important context for considering the travels of any single numerical datum. The purpose of this paper is to explore and explain how it is that different kinds of group data sets, where the data are collected and aligned according to different measuring principles and to represent different subject matters, affect the travels of any datum point in the group. Using examples from social science, the paper examines how the relations of the data points within the whole set determine the possibilities for any single individual datum to travel within and out of its set, and how the integrity and fruitfulness of data or datum journeys will be dependent on those bit-whole relations that characterize the group data set.

1 Introduction

The natural world is full of examples of clouds of individuals travelling in groups, groups significant enough that we have given them special labels that suggest their different group behaviour in terms of individuals: swarms of midges, murmurings of starlings, armies of ants, packs of wolves. To the amateur naturalist: ants line up, wolves practice hierarchy and strategy, starlings free-wheel according to some unaccountable design, while midges just swarm. The specialist animal behaviour expert will have more exact descriptions than these folk terms, but the point to focus on is how the whole is understood as a large set of small elements which cohere in very different forms and behave in different ways to make up the whole.

We can see a similar variety in the bit-whole relations of data that are taken to represent complex group behaviour in the social world. Studying a social whole

M. S. Morgan (✉)
Department of Economic History, London School of Economics and Political Science, London, UK
e-mail: m.morgan@lse.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_6

such as a city, an economy, or a society, requires the construction of 'group data sets' where the group is made up of a number of individual data series, each one in turn made up of a string of individual data points or datums.[1] Any individual datum (or bit) has relations not just with the other data points in their series, but also with those of the group (or whole) data set. For example, the data on population growth of a society consist of individuals, who can be counted in a simple aggregate whole, but for social science purposes will more likely be found in data series divided by occupational classes, or age cohorts, or regional spaces. The bit-whole relations will depend upon the kind of group data involved, for there is variety in bit-whole relations just as in those naturalists' examples suggested above. No doubt these varied kind of datum-to-'group data set' relations can be found in other fields of science with complex wholes such as ecology, physiology, and so forth; it is not necessarily a special feature of social science data. What is important is that different kinds of data sets in the sciences have different bit-whole properties, and that these turn out to be very important for the possibilities and fruitfulness of individual datum journeys. So, while the datum and its travels take centre stage in this paper, it does so always in relation to its 'companions' not just in the individual data set, but in the group data set, which should be conceived as its primary context. This focus on the datum-group data relations sits in contrast to many other studies in this volume, and to earlier studies of travelling data, which focus on other kinds of 'companions' and other background and foreground contexts which affect the journeys of data.[2]

The purpose of this paper is to explore and explain – for quantitative data – how it is that different 'kinds' of numerical data form an important context for a single numerical datum. I take kinds of numerical data to indicate numerical data collected and aligned according to different principles into group data sets. The most important principles that I consider are those that stem from the kinds of measuring systems involved in the construction of the group data set. The subject matter of the data set is also important of course, but this is not the primary focus of my discussion here.

For an example, consider the measuring system based on statistical thinking. This involves the notion of an underlying statistical population, and modes of sampling in collecting data (random, systematic, representative, stratified, etc). The relations between individual data points within each statistical data series will depend primarily on what kind of population is involved and whether the datums come from, for example: a sample from a controlled trial in medicine, a time series in economics, a survey in sociology, or the demographic census of population. They

[1] It is important in this paper to signal the collection of individual data points in a way that maintains their individuality: as 'datums', a jarring term that enables me to insist on this important distinction to the collective plural 'data' where individual distinctions are not relevant.

[2] See the notion of 'travelling companions' for the successful journeys of data (to use the language of the How Well Do Facts Travel? project – see Morgan 2011a, and the other essays in Howlett and Morgan 2011). Sabina Leonelli's (2011) contribution to that project volume, and her subsequent book (2016), on the curation of plant research, provides an important parallel for the ideas of this paper. In her case, the information on both background and labelling are essential elements that travel with the data. Here the focus is on the other data points in the data set as companions.

will each have different bit-whole relations that depend on the statistical framework and the subject matter. For example, the data points in a rain-fall data series are clearly related in the time sequence and cannot be randomly re-ordered in the data array without losing some really important information from the data set, whereas data taken from a controlled field trial can likely be ordered and re-ordered in the array without breaking any internal relationships between the data points. Broader subject matters hold further power. Ted Porter (1986) and Ian Hacking (1990), in their writing about the history of statistical thinking, have exemplified how such subject matters meant that astronomers' personal errors of measurement were first formulated according to a 'law of error', and then show how such law-like distributions were reformulated as human social character deviations, and thence reinterpreted into natural biological variation in what became known as the 'normal' curve. Meanwhile, the behaviour of populations of human individuals became the analogy for the kinetic theory of gases and evolutionary theorizing using biometrics. The data from all these domains share notions of statistical populations and distributions, but their subject interpretations and usage differ.

Following further the original example, the data of a national population when measured by a census of population are both statistically 'governed' (by the nature of such population distributions, and principles of taking good samples) and 'governed' by the socio-economic characteristics of the nation (such as occupational class, or age aspect, or regional characteristics) that are to be measured. So, we can understand whole (or group) data sets as involving the following elements: individual datums (or bits) that are assembled into data series, which are then packed into subject category boxes, which taken together form parts of a whole data set. The category boxes depend on the purpose and framing of the whole data set, so the same data series may appear in many different whole data sets. But how those boxes or parts fit together depends on the principles of measurement of the whole that are being followed. There is rarely a simple aggregation at any point. In the population example, a simple aggregation (from samples to population, and over time and space) will tell us the total number of individuals at a given date, but this has little use. Most analysis will want to know the categories and how they fit together in the whole. Then, what can be extracted from the whole to travel with validity depends on both the base principles of measuring the bits, the categories and how they divide the world, and the conceptual nature of the whole.[3]

This point may be clarified by contrast with another data set dependent on the statistical notions of population. The data of sampled biological populations in worldwide genetics or genomics data sets depend on the hereditary properties of specimens and evolved relationships of sample subjects as well as on theoretical assumptions and empirical practices of sampling and specimen collection.[4] These

[3] And given this, it is no surprise that any data that travel have to be carefully resituated in a way that protects their integrity in any new site, as other papers in this volume make clear (see Leonelli's introduction).

[4] I thank Jim Griesemer for this parallel example from his field (see his chapter in this volume that exemplifies the point).

two different fields of science both use the term ‘population’ and rely on statistical principles of collection and ordering. The data journeys that occur in these fields have multiple valences, and their data journeys surely differ. Even so, the datums from these different fields may well share similar characteristics of detachability, and so their journeys might have more in common with each other than with the journeys of data from sets in the same subject field but constructed according to very different principles of observation and measurement.

Both principles of measurement and subject field relations have considerable impact on the way that data are conceived and used, and so on their possibilities for travel as empirical objects, as ‘theoretical’ stand ins, as stand-alone values, and the like. Whether, and under what conditions, an individual datum point can travel within the data set, or independently beyond it, and whether such data travels are associated with integrity and fruitfulness[5] in travelling will depend in part on the nature of those internal relations of measurement principles and subject matter that characterise the data set. This creates a presumption that data journeys will be affected by the characteristics of the whole, as much as of the parts and of the relations between those parts.

The importance of this framing, and emphasis, on the principles that lie behind whole data set measurement is demonstrated in this paper in a comparison of two sets of numbers that economists and social scientists use when they aim to get a grip on a national socio-economy. These two data sets are assembled according to two very different kinds of measuring and aggregating principles. One set uses accounting principles: everything must be counted once and nothing twice, columns must add, and bottom lines must balance. Using these accounting principles produces a group (or whole) data set that includes many individual data series, each of which has a place in the accounting system: – a system set up to measure national economic activity both within certain categories and as a whole. The other group data consists of a set of ‘indicators’: numbers that are not conceived as direct measurements of the concepts they relate to (such as the business cycle, or the health status, of a country), but are understood to be indicators for characteristics relevant for those concepts (such as, respectively, industrial production or infant mortality). These two kinds of group data sets were first developed in the mid twentieth century to draw together many different data series in attempts to count, measure, or capture the whole economic activity of the nation state: they were the social scientists’ ‘big data’ projects of their time. They were, and are, produced according to very different principles – accounting vs indicators – and so exhibit very different bit-whole relations within the group data set. Both provide aggregates in some sense, but according to different principles. My analysis will show how their bit-whole relations are critical for determining the very different possibilities for using individual datum points within the data set, and will explore the kinds of reasoning and analysis that goes on when data are taken out of the whole for use.

[5] See Morgan 2011a for the importance of ‘integrity’ and ‘fruitfulness’ in data journeys.

2 Data Sets and Their Kinds

Scientific discussions typically refer to data not to a datum, because scientists rarely deal with an individual datum which is not also part of a bigger set. Often, the term 'set' refers to a data series (a string of data collected under the relevant same conditions) but here the arguments relate to a group of such series – referred to here as a 'group, or whole, data set'. Typically (as suggested above) the data points – the datums (see note 2 again) – within such a group set are held together by two sorts of relationship. One comes with the theoretical and interpretative constraints of the scientific subject field in which they live. The other – more important for the argument of this paper – comes with the means and principles of measurement that underlie their collection and their colligation into the group set. At the level of the group, these measuring principles generate different kinds of relations between the individual datums and between the series in the group. Conceived as measurements, numerical data are not all the same kind of thing.

I use the term *kind* of data to point to the facts that there are different kinds of 'measuring instruments' involved in producing numerical data, a term of usage in this context due to Marcel Boumans.[6] The measuring instruments used in social sciences look rather different from the thermometers, Geiger counters, and so forth, that might be first thought of when considering scientific measuring instruments. In the social field, they are mostly various kinds of counting systems that rely on observation posts spread out across the country in government offices, banks, companies and families who all report aspects of their lives (usually for completely other purposes). The raw data collected from these observation points are numerical, and combined in different ways, according to the frameworks or principles and techniques of the measuring instruments (consisting, as Boumans argues, of models, formulae, rules, conventions, etc) used to turn such raw numbers into measurements of the economy and society.

The following analogy may communicate the point. One can think of there being families of measuring instruments rather like there are families of musical instruments in an orchestra: woodwind, percussion, brass, strings etc. Each family of instruments produces sounds according to a common principle or recipe and set of techniques; but within each family, individual instruments occur with slightly different characteristics: violins and cellos use one principle (using taught strings) for making music, but do so with different objects and range; the percussion family has their own different strategy (of hitting objects), with individual instruments of more variety of range. Within an orchestra, all play together, but still, the family voices can be separately recognised as characterised by the principles of the instrument of the relevant group. The analogy here is that in socio-economics we have different families or *kinds* of measuring instruments, all producing numbers as measurements.

[6] Marcel Boumans, in a series of papers (but especially his 2001 and 2005a and 2005b book), developed the idea of using of this term 'measuring instruments' to analyse the formulae that create numbers for the phenomena of economics.

Some of these numbers are produced using principles of statistical thinking (populations and samples); some use accounting principles (of aggregating and balancing); some use principles of tracking (indicators that track characteristics of the phenomena); and some use principles of splicing with weights to make aggregates (in the form of index numbers).[7] Thus, for the social scientists, statistical processes produce data of a different *kind* than those produced by accounting principles, which are in turn of a different *kind* than those producing indicator data, and another *kind* than those producing index numbers. These different *kinds* of data come from using four different *kinds* of 'measuring instruments', each using different principles and strategies to recognise, collect, code, assemble, and organise the information from raw observations into numbers (see Morgan 2001, 2007). Just as the instruments in the different orchestral sections produce sounds according to different principles, these different measuring instruments produce numbers of different kinds using four different principles of measurement. So when I refer to *kinds* of measurements in this account, I am pointing back to these principled-based measuring instruments that produce such kinds of data at the group level.

That specificity of the *kind* of data in question has implications for the possibilities for data travels, not just because of the different nature of those data kinds, but also because the internal relations between data points that are carried within any data series or group data set derive from their principles of construction and usage. These four different kinds of measuring instruments will produce data sets where the relationships of individual data points to their group data sets, that is of bits to wholes, have different formats. Any one datum will come from a group data set which is collected, and aligned, according to the principles of a specific kind of measuring instrument, and that datum has to be used and interpreted with that relevant set of background principles of the measuring system always in mind. This family sharing in the principles of a measurement instrument used in constructing a data set may matter as much, possibly more, than the scientific subject field for the nature of any data journeys. Thus, for socio-economic data that come from different measuring instruments, and so produce different kinds of group data sets, the very different internal relations will be critical for understanding the different possibilities for data journeys, and what happens to datums when they travel.

Conceived as measurements, the group data set produced using any one of these four socio-economic measuring instruments is expected to have some kind of a representing relation to the phenomena of interest that scientists want to investigate. These are likely to differ according to the kind of data involved. The formal 'representational theory of measurement' investigated this question seriously for a number of characteristic measuring systems (see Suppes 1998). That approach can be contrasted with the pragmatic approach of Finkelstein (1982) for whom 'measurement' always involves some form of observation. The materials here suggest that both notions are more valuable when they can be taken together. First, socio-economic

[7] A 'population-samples' example was discussed in Sect. 1 above, other are discussed later in this paper; and see Morgan 2001 for further discussion of each kind of numerical data.

numbers are often not direct measures of such phenomena by active scientists, but more often 'observations' taken for other purposes and abstracted from their original economic contexts in life. Second, for data to capture complex socio-economic phenomena, just as for complex environmental processes (such as in ecology), a single datum will rarely do so, which is why the nature of the group data set and its construction is so important. While at the level of the individual data series, social scientists habitually use different kinds of data sets produced by different measuring instruments to represent the things in their world, that does not immediately tell us what matters about the differences in these forms of representation for their group data set, nor for their data journeys (either as a set or individually). So, I use the term representing here in a pragmatic way, generic but informal, and will explore in what follows, how – for a kind of data (ie from a kind of measuring instrument) – datum and data journeys will be affected by the characteristics of the whole, as much as of the parts, and of the relations between them.

3 Economic Data: Perspectives on the World

There are two very expansive sets of data used by economists and social scientists to look at, and into, the economy/society as a whole unit. They both operate by assembling data at the national level, and they do so in standardised forms to enable comparisons across nations. They both provide a numerical account of the economy or society showing not just the whole, but also the bits of the economy/society in relation to the whole. They do so by using two (of the above four) different kinds of measuring instruments which offer very different kinds of perspectives and so create different kinds of data. One kind offers a broad view and one a deep view, and so parallel in numerical form the kinds of visual perspectival accounts that Svetlana Alpers (1984) examined in her contrast between the broad cityscapes of the Dutch painters and the deep distant landscape view provided by the Italian painters of the early modern period. Both groups of artists provided pictures of the whole, and both enabled you to see the elements in the landscapes as bits in the whole in relation to each other. These are paralleled in these social science measurement systems in that one kind looks broadly to pick up the full range and diversity of phenomena, the other looks more deeply to reveal the interrelations between a smaller range of phenomena that are taken the characterise the economy as a whole.

These two different kinds of data set examined in the rest of this paper provide the materials to consider the dependency of datum travels on the measuring structures or instruments they come from. One kind of data set, the one that looks deeply, is the national income accounts (NIA). It announces the nature of it its internal relations in its name:- an individual datum is tightly ordered in the whole by the accounting principles of the measuring instrument. The other kind, socio-economic indicators, are much less individually constrained and together they look across a wide range of the phenomena of the whole, capturing all the individual elements separately that make up a picture of socio-economic development.

I need to be careful here: for we are really talking about two master data sets – whole or group data sets – one assembled according to accounting rules, the other according to the indicator format. But inside each group data set, there are many series of data, each one consisting of data that have been collected, coded, assembled and manipulated to represent a particular element of the economy or society. These data series are not raw but highly wrought and polished. Any one set of numbers in the NIA data set, or any one indicator series in the overall database of indicators, may be constructed according to any of the measuring instruments: some may come from accounting processes, others by statistical methods from surveys or censuses, others are simple numerical counts. Regardless of the numerical provenance of the individual series, it is the relation of each of these individual and separate series to each other and to the whole that are formulated according to those group-level (accounting or indicator) measurement frameworks.

Both kinds of measuring instruments are generative, in the sense that they generate whole data sets designed to represent in some direct or indirect way some conceptualised phenomena. The middle level stuff of the social sciences represented in the separate data series is not stuff that can be found raw (with whatever practical difficulties); it is stuff that must be fashioned to fit, more or less indirectly, their conceptualised phenomena. Thus 'national income' and socio-economic development' are both highly abstract: no one can 'see' national income, or socio-economic development in any direct way *through* a microscope. But social scientists do 'see' (ie generate) *with* their microscopes, data on something they conceptualise as development, or national income. We could even label the NIA a 'national-level analytical-accounting macroscope'. The point here is not to subvert Ian Hacking's (1983) seminal point about seeing *with* rather than *through* our measuring instruments, but rather to extend it for thinking about measurement at the macro scale and in the social sciences where measuring instruments are not physical but organisational and technical.

3.1 Accounting

National Income Accounting (NIA) began in the late 1930s as a project to count all economic activity of the national economy for each year. It was developed into a usable system by the end of the 1940s, its development hastened by the needs of various national governments to organise the 'war economy', a period which stretched the limits of productive capacity and in which governments needed to plan the economy. Such accounting became equally important in peace times as the new post-war international economic arrangements and agencies required such measurements as part of their regulatory agendas. In such an accounting, a national income data set, constructed for each country (or possibly sub-region) separately, provides an accounting picture of the whole national economy and its salient parts, where all the parts are related to each other in an accounting framework. That framework provides the rules of what to count, how to count, how to check that everything is

Table 1 The simplest table of national accounting

I	II	III
Net national income	Net national output	Net national expenditure
1. Rents 2. Profits 3. Interest 4. Salaries 5. Wages	7. Net output of agriculture 8. Net output of mining 9. Net output of manufacture 10. Net output of distribution 11. Net output of transport 12. Net output of other services	14. Expenditure on goods and services for current consumption 15. Net investment
6. Total net national income	13. Total net national output	16. Total net national expenditure

Source: Adapted from Deane 1948, p. 9

counted, and uses balance checks between the wholes to ensure that everything (within its framework) is taken into account.

It is a three-dimensional account – the aggregate economy is measured according to all incomes (Column I), all things produced (Column II), and all expenditures (Column III). It appears in one of its earliest and simplest forms in Table 1 showing the three columns or dimensions each with its associated categories (adapted from Deane 1948, and see Morgan 2011b). Everything that has to be counted has to be placed in the right place (column and row), so every individual data series has to be categorised, that is, national accounting operates under a system of categorization rules for the individual series. (And these accounts can be broken down into finer sub-categories and equivalent numbers.) The bottom line categories for each column: 6, 13, and 16 form an identity based on the principles of the accounting. When the table is filled in with the relevant numbers, the three numbers for these categories should be equal because they constitute three different ways to count what economists consider to be equivalent in monetary terms. If the different columns of the system do not balance, the implication is that there is something missing somewhere. That is, 'the bottom line' of accounting must *balance* as a matter of principle.

The national income accounts operate not only to measure aspects of the aggregate economy as depicted in the data set, but as a standardised set of measurements that can be reasoned with and are essential in helping governments make policy. Those reasonings are primarily driven by the functional or behavioural economic connections between the elements in the accounts, but any reasoning will have to be reflected in the accounting numbers and consistent with the accounting principles. This is just the same as using accounting for a firm or company. A firm's accounts are both a representation of the company's health, and a functional space for thinking about changing the performance of the company. So, if a *company* invests more, it expects to grow in overall product in successive years as a functional relationship; such changes will of course be reflected in the accounting relations. But less obviously, they are also constrained by the accounting relations: if there is no profit, there is no money to invest and so it must come from other change in the company's activities. These relationships and constraints are all revealed in the accounting numbers. Similarly for the aggregate numbers of the NIA: the numbers represent the economic situation for the national state for a year

(the usual accounting period) and so function in two different, but coherent, ways: as subject categories with accounting rules, and as subject categories with economic relations. So, again, if a *nation* invests more, it expects to grow in overall product in successive years. Fruitful uses of the data can be found even when the individual datum elements are mutable, and surprisingly this is precisely because of these strong internal relations.

This may all seem obscure, so an example that demonstrates these characteristics of a travelling datum in this context may clarify. The example comes from Wolfgang's Stolper's attempt to make a plan for the Nigerian economy in the early 1960s at a time when it had just gained independence (see Morgan 2008). His planning asked each individual region to submit their specific plans for investment to the federal government so that all their plans could be put together. Each datum point supplied by the regions had to be found a place in the national aggregate plan, but the construction of the measurement system meant that to do so, it had to fit with all the other current and future pieces of information in the NIA system that represented the Nigerian economy of the day. So, for example, a region that wanted to build more schools could come along with their costed project to do so. Such a project would require more trained teachers (and so more college places in the education system), and more construction (entailing the building industry, with labour and resources), all requiring changes in Column II, row 12 (see Table 1). Both more teachers and more school buildings would necessitate more government expenditure in Column III, row 14 or 15. If this part of the plan went ahead, those activities would generate more incomes in Column I, row 4 or 5, and so consumption in the system as a whole: Column III, row 14. This last reaction is described by the economic relation, known as the 'multiplier effect', that can be traced through the categories and data set of the NIA. The individual datum elements for each numbered category can be 'taken out' of the accounts by the government planner, altered to show this change, travel and be re-situated in other contexts (such as in a local budget for a school building), and be replaced in the national accounts by a new number. But the usefulness and fruitfulness of such datum journeys are most evident when each travels as a member of the national (NIA) data set into a context where both the internal accounting principles, and the subject matter economic relations of that NIA data set are made use of.

As an accounting system, there are very strong requirements of consistency, but the processes for re-balancing the bottom lines are driven by the economic relations within and between the columns. If, for example – as a result of the new school bid – some other government funded activity (asked for by some other region perhaps) would have to be curtailed to make this schools investment possible, this in turn would reduce the multiplier effects – that is, there would be balancing effects across the accounting columns and rows. Any planning number that is taken out and replaced with another such is likely to alter the whole table, as depicted in Fig. 1. These numbers are expressed in, and represent, monetary amounts, but in turn those monetary amounts represent real things in the economy: people earn incomes by educating children in school buildings. Time consistency matters too – more investment in schools this year would not only imply less of something else now, but

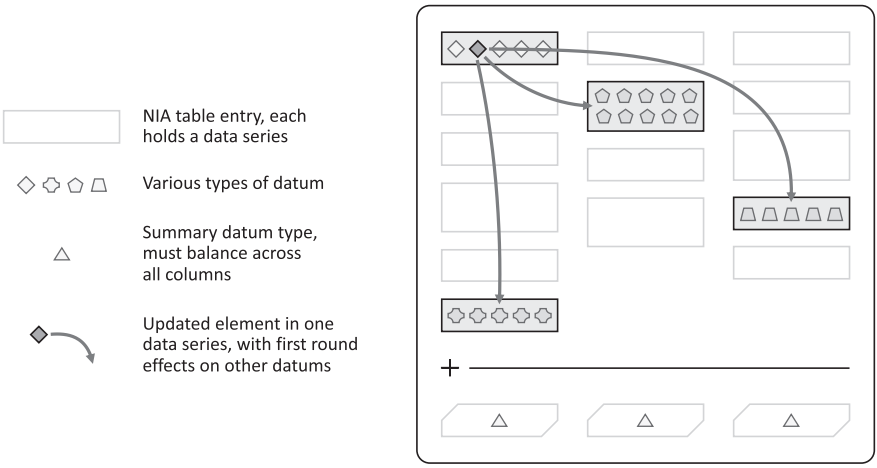


Fig. 1 Accounting kind: NIA whole data set

might also produce more returns in the future from an educated workforce, so there was also a process of making the present and future numbers consistent. As Deane remarked, the economic policy maker

> *wants to be able to see each of the constituent items in the network of national economic activity not only as a separate feature of the national accounts, but also as a factor influencing and influenced by other activities…* (Deane 1953, p 3)

Even without going into more details, it is possible to see that, in reasoning about one datum point (the numbers for investment in new schools) – it is not possible to pull the accounting principles and economic reasoning away from each other. But it is equally easy (I hope) to see that the accounting principles operate not only as a reasoning space, but also as a constraint on that reasoning about the future of that society, a specific society in time and place taking into account all the other datum points that involves. An individual datum can be transported in or out, and be mutable within the planning system, and create mutability in the represented economic system – provided only that all the consistencies hold. In other words, there are strong requirements in the accounting principles that constrain the numbers and determine the reasoning with them (see Morgan 2008).

3.2 Indicating

The second kind of data base involves so-called ‘indicators’, typically made up of a set of data series, each one indicating something of relevance for understanding the many dimensions of socio-economic aspects of life. An indicator series is one that aims to track or indicate one aspect of a complex phenomena – each characteristic of that phenomenon will have a separate data series. Sometimes these can be charted

or sited in tables together, but they are not so easily combined for both technical and subject matter reasons.[8]

There are two examples which are close to the NIA in aiming to capture, in numbers, characteristics of the aggregate economy or socio-economy. The first example, business cycle indicators, were developed in the 1920s and 1930s in literature which crossed over between academic and public domains (and are still evident in the financial press nowadays). They were numbers that were held to capture or represent characteristics of the business cycle at the level of the nation state, a phenomenon that was difficult to define beyond the idea that it was cycles in the economic activity of an economy. While the causes and mechanisms were not so evident (and are still argued over), the community of economists had more agreement over the general characteristics of the phenomenon, yet also believed that these characteristics (and the timing of such cycles) were to some extent specific to a national economy. While all national economies would have some indicators in common (eg interest rates, exchange rates, bank deposits, exports and imports, etc), a highly industrialised economy might additionally be characterised by a set of industry indicators, while a more agricultural economy might be best represented by an additional set of primary sector indicators. A relatively small set of such indicators (perhaps up to a dozen) were taken to characterise economic activity as well as offering some insight into the timing of cycles evident in the time relations between each indicator series and thus in each characteristic element. Both the overall set, and the time relations between them were taken to indicate the nature and path of economic activity for the national economy. None of them could serve as 'proxy' for the whole economy, because they did not represent the whole economy directly or indirectly but only aspects of it. And there were technical difficulties in making combinations: they did not each follow the same pattern in the same time frame. More pertinantly, they could not easily be combined into one single indicator because, although they exhibited correlations, there was no principled way that they could be related as far as subject matter was concerned. Business cycles on the one hand operated as a rather vague concept, and on the other hand as a phenomenon of many characteristics which could not easily be patterned or drawn together into a causal network, nor measured in any direct way.[9] Indeed it was partly this problem that lead economists to prefer the greater insight offered by the joined-up system of national income accounts which became available in the 1940s and 1950s and so made business cycle indicators less important.

A similar kind of data structure, but with a much higher dimension of characteristic elements and with much broader reach of subject area, are the indicator set now being developed for the UN's Sustainable Development Goals. These replaced the

[8] Morgan and Bach (2018) explore why such data series cannot be easily or informatively combined, which might be considered in comparison to the data mash-ups of epidemiology and related fields, see Leonelli and Tempini (2018).

[9] See Boumans and Leonelli (this volume) who discuss the rather 'inflexible' characteristics of data clustering associated with business cycle indicators; they argue that these practices, in this context, are an interpretative move which has not encouraged the re-use, or aggregation, of these data for other purposes.

Millenium Development Goals (2000–2015), and are substantially more ambitious (see Morgan and Bach 2018). This set of 230 data series is designed to offer a numerical picture of every nation's socio-economic health, including now their environmental health. They consist of a bundle of separate data series, each one having an 'indicator' relationship to one of 169 'targets', each of which itself has an indirect relationship to the 17 'goals'. By indirect, I mean that the indicators don't offer measurements of, or for, one of the targets but only numbers related to one characteristic of each target; in most cases there are several indicators per target and several targets per goal. That is, both goals and targets are multidimensional and goals in particular are defined verbally and conceptually rather than in any measurable way (unlike the aggregate gross national income in the NIA accounts). For example, Goal 3 of the SDGs is aimed at increasing health and well being. It is accompanied by a set of targets concerned with maternal and infant healthiness, reducing preventable diseases, providing access to health care, and so forth. Some of these are easier to associate with numerical evidence than others. Each of the 9 targets for Goal 3 is accompanied by a set of indicators which can offer numerical evidence associated with the current situation of that target in different countries over time. These indicators – such as 'malaria incidence per 1000 population' or 'road traffic deaths' – *indicate*: they offer numerical information about some aspect of one target in relation to the goal, but they are far from measuring or representing the target let alone the overall goal that needs to be represented. This example is rather straightforward for there are lots of health-related data series that can be turned into numerical indicators. But suppose we take a more opaque Goal 16: 'promoting peaceful and inclusive societies' and ask for 'legal identity' as a target for inclusivity: we are immediately faced with difficulties in finding ways to indicate this concept. For example, how should one rank-order the various forms of legal identity, let alone find numbers for them? Registered birth and citizenship are relatively straightforward and likely have relatively good numbers collected by the state. But what about the host of in-between status such as 'the right to remain and work but not have your children have the right to school or health care'? Even assuming we had numbers that would fit those categories of people, we have no principled way to rank-ordering the categories, nor to value them in some commensurable way.

Because of the three-level 'goals-targets-indicators' system of the SDGs, these indicators have a double degree of detachment from their goals, and so distance in representing power, for those goals and targets to which they are attached (see Morgan and Bach 2018). The indicators are taken to represent the characteristics of the targets (in some form), and the targets are taken to represent the characteristics of the goals (in some form). This is a downside for the representing power of any data set. At the same time, the various indicator series remain largely independent of each other, having no formal or informal relations between them. They are not part of an interrelated causal account, although individual series might capture individual symptoms, causes or consequences of underdevelopment. (For example, high infant mortality is thought to be a *consequence* of low levels of development whereas low levels of education are thought to be a *cause* of low levels of development.) They cannot be aggregated according to any usable or principled rules as

works for the NIA, nor provide matter for functional or behavioural theorizing about socioeconomic development as we saw for the NIA. And unlike the 'index numbers' by which economists regularly measure multiply-component concepts (for example inflation, or industrial output), social scientists cannot easily turn these sustainable development indications into a single overall data series that would make sense according to measurement principles. Why not – because they are not measured in comparable units (eg money) nor is there any principled ways of deciding how to weight the various elements in the whole (eg is legal identity worth 10% of total sustainable development or 1%). They cannot be turned in any principled way in an aggregate measure like the national income, nor combined in a principled way consistent with ideas about development into one meta-data series for each country and so be available for international comparison.[10] While these data certainly contain *information* indicating characteristics of development, they should not be considered *measurements* of development.

As individual indicators, these data series and individual datum points can and do travel fruitfully from the statisticians to many users including into social scientists' research labs and are used for many varied topics not just those of development even though their status as measurements in relation to development theory is not generally easy to determine They also travel from UN usages to a variety of other users for any other purpose they choose for them: they are public numbers for public use and their usage depends in considerable part on their UN provenance that makes them trustworthy (Porter 1995). As a set of 230 different data series indicating levels of development for each country member of the UN, they provide a whole data set. As such, they most frequently appear for use in comparison purposes in social scientific work, and for certain action purposes at national level.[11] But they remain a set of data series, not an integrated whole measurement system, as depicted in Fig. 2. Consider the problem situation parallel to our NIA example: suppose a government wanted to use the SDG structure of goals and targets to create a more sustainable development path. They cannot be reasoned with for planning a development programme in a nation state because they have no internal socio-economic relations generated either by association with the kind of measuring instrument involved, nor by any behavioural or theoretical relations from their subject matter. But, the very fact that the indicator numbers are not held tightly together by internal relations between different indicators (as in the NIA), and that they might be indicating a cause or effect or symptom, means that individually they can be (and are

[10] Several data series might be 'mashed up' (see again Leonelli and Tempini 2018) into a single series for each country or region, but the informative quality of the resulting numbers would likely be low, and the country comparisons largely meaningless, for the grain of analysis is not nearly fine enough across the geographical space to be helpful. This is in contrast to the Multidimensional Poverty Index which was carefully designed to be a combined number that was informative at a finer grain than previous poverty indices (see Bach and Morgan Forthcoming).

[11] It would be a false separation to think that there are scientific uses and policy or practical uses for these indicator number or for the NIA: all these numbers are hard to come by; gathering them generally requires public resources; they are used by professional communities of practice in and out of academic institutions; and for a wide range of purposes.

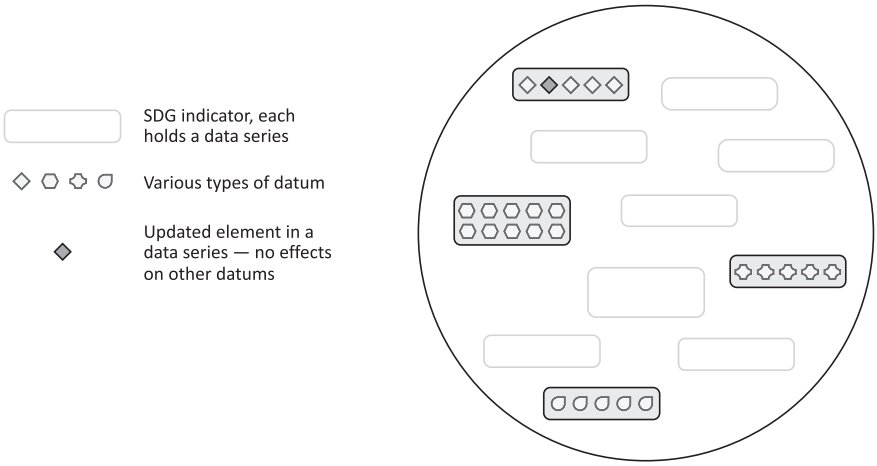


Fig. 2 Indicator kind: SDGs whole data set

easily) taken out of their group set, to be used separately for prompting action in all sorts of circles: academic and scientific in the professional sphere; and in public and international circles where the indicator data can be used for lobbying, asking for development aid, held up as exemplary for encouragement, or pronounced as dire in order to shame the government concerned. The lack of internal principles to hold the individual indicator series together makes for different characteristics of usage than individual numbers in the NIA.[12] Indicators can be used with considerable freedom without worrying about the constraints of measurement principles or where they fit in the overall subject contexts, and this is most evident when they travel from domain to domain of usage. Unlike the NIA, where every datum travelling in and out has the potential to change all the other numbers (if only to correct them), travels of the SDG's indicators cause no ripples within the rest of the indicator system of data, as indicated in Fig. 2 in comparison with Fig. 1.

4 Conclusions

Economists have developed two kinds of data to capture social-economic well-being. They are based on two different frameworks of measurement The national income accounts are designed to measure the complete set of income, expenditure and products at the level of the nation. They do so by building up from the subcategories of all these three activities which are understood to be – in the bottom line –

[12] It is possible that these independent data series in the indicators could be analysed and combined with correlated analysis within the national unit, or between/across national units. The latter possibility is not dealt with in this paper (but see also FN10).

equivalent (in economic and monetary terms). In contrast, the indicator series may look just as ordered because they are arrayed in connection with bigger targets, but they are in fact held together by no such constraints.

From these different measuring frameworks, come differences in usage. For the bundle of indicators, each of which can be used for action but not reasoning – any travelling datum has no effect on the whole. In contrast, the other kind is the highly constrained NIA which can be used for measuring the current health of the economy, and for reasoning and action in that realm, but in which any travelling datum can upset – and then must reset – the whole system. Perhaps counter-intuitively, datums from both travel easily and fruitfully into new contexts.

Not all indicator systems have this degree of bit-whole freedom. Datums from the business cycle indicators for example, tend to travel together because they indicate time-related characteristics of the same phenomena. Each datum and indicator can be taken out separately, but they gain from travelling together in a pattern, perhaps like a murmuring of starlings. In contrast, the indicators of the SDGs are more like a swarm of midges, with no recognisable pattern and no obvious relationships between the bits. Both of these indicator sets are very different in their relations to each other and to the whole compared to the national income accounts (NIA). Whereas both individual datums and series from the indicators have bit-whole relations, those for the NIA depend on their part-whole relations. The NIA parts might look like the ant-line, because if one element travels off the path for some reason (eg, for correction or updating), the rest have to fall in to make up the line. But they have more part-whole relations than just lining up, since they rely on multiple relations for their effectiveness in reasoning and analytical usage, and this relies on a well ordered hierarchy of rows and columns; thus the relation of parts within the whole is more like the hierarchy and co-ordination of the wolf-pack. Or perhaps – as Jim Griesemer suggested,[13] to bring the analogy into line with our socio-economic world: a bundle of indicators is like a flashmob of independent agents – taking a datum out or bringing one in does not upset the whole; in contrast, the national income accounts are tightly joined together so that taking out a datum would be equivalent of taking a section of piping out of a chemical plant: the whole process would need to be reassembled.

When we think of individual datum travels, one has to think first of the rest of the data set as their most intimate of travelling companions. Datums rarely travel on their own without their companions in the data series or set, but when they do, that set of interrelations – or indeed lack of such relations – within the whole data set is critical to their independence of travel and how they fit into their new contexts. That set of interrelations in turn depends on the measuring structures or instruments that were used to generate and organise the individual data series and individual datums within them.

[13] Thanks to James Griesemer for this incisive analogy – provided at the Exeter meeting in 2017 that spawned this volume.

Acknowledgments I thank Sabina Leonelli, Niccolò Tempini, participants at the Exeter conference and especially Jim Griesemer for an interesting collaborative writing experience which prompted this paper. I thank Marcel Boumans for many illuminating discussions about measurements in science and economics which have influenced the content of this paper; Michel Durinx for help with the illustrations; and Maria Bach for allowing me to draw on our two recent papers (Morgan and Bach 2018, and Bach and Morgan Forthcoming).

References

Alpers, Svetlana. 1984. *The Art of Describing*. Chicago: University of Chicago Press.

Bach, Maria, and Mary S. Morgan (Forthcoming). Measuring Difference? The United Nations' Shift from Progress to Poverty. *History of Political Economy.*

Boumans, Marcel. 2001. Fisher's Instrumental Approach to Index Numbers. *History of Political Economy* 33 (supplement): 313–344.

———. 2005a. *How Economists Model the World to Numbers*. London: Routledge.

———. 2005b. Measurement Outside the Laboratory. *Philosophy of Science* 72 (5): 850–863.

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.

Deane, Phyllis. 1948. *The Measurement of Colonial National Incomes: An Experiment.* National Institute of Economic and Social Research, Occasional Papers XII. Cambridge: Cambridge University Press.

———. 1953. *Colonial Social Accounting*. Cambridge: Cambridge University Press.

Finkelstein, L. 1982. Theory and Philosophy of Measurement. In *Handbook of Measurement Science, Vol 1: Theoretical Fundamentals*, ed. P.H. Sydenham. New York: Wiley.

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.

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

———. 1990. *The Taming of Chance*. Cambridge: Cambridge University Press.

Howlett, W.P., and Mary S. Morgan, eds. 2011. *How Well Do Facts Travel?* Cambridge: Cambridge University Press.

Leonelli, Sabina. 2011. Packaging Small Facts for Reuse: Databases in Model Organism Biology. In *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*, ed. P. Howlett and M. Morgan, 325–348. Cambridge: Cambridge University Press.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: Chicago University Press.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Leonelli, Sabina, and Niccolò Tempini. 2018. Where Health and Environment Meet: The Use of Invariant Parameters in Big Data Analysis. *Synthese* (online, June 8, 2018), https://doi.org/10.1007/s11229-018-1844-2.

Morgan, Mary S. 2001. Making Measuring Instruments. In *The Age of Economic Measurement* (edited with Judy Klein) *History of Political Economy*, Annual Supplement to Volume 33, 235–251. Duke University Press.

———. 2007. An Analytical History of Measuring Practices: The Case of Velocities of Money. In *Measurement in Economics: A Handbook*, ed. M. Boumans, 105–132. Amsterdam: Academic Press.

———. 2008. *'On a Mission' with Mutable Mobiles*. Working Paper 34, The Nature of Evidence: How Well Do 'Facts' Travel? project, Department of Economic History, LSE.

———. 2011a. Travelling Facts. In *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*, ed. Peter Howlett and Mary S. Morgan, 3–39. Cambridge: Cambridge University Press.
———. 2011b. Seeking Parts, Looking for Wholes. In *Histories of Scientific Observation*, ed. L.J. Daston and E. Lunbeck, 303–325. Chicago: University of Chicago Press.
Morgan, Mary S., and Maria Bach. 2018. Measuring Development – from the UN's Perspective. In *The Political Economy of Development Economics: A History Perspective*, ed. Michele Alacevich and Mauro Boianovsky. *History of Political Economy* 50, (supplement): 193-210.
Porter, Theodore. 1986. *The Rise of Statistical Thinking*. Princeton: Princeton University Press.
———. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.
Suppes. P. 1998 [2000]. Measurement, theory of. In *Routledge Encylopedia of Philosophy*, ed. E. Craig ed. London: Routledge. https://doi.org/10.4324/9780415249126.

Mary S. Morgan is the Albert O. Hirschman Professor of History and Philosophy of Economics at the London School of Economics; she is an elected Fellow of the British Academy and an Overseas Fellow of the Royal Dutch Academy of Arts and Sciences. Her past research has focussed on practical aspects of economic science (questions about models, measurements, observation, experiments) and on broader topics in history and philosophy of the sciences (such as the role of case studies and use of factual knowledge), with most recent books: *The World in the Model: How Economists Work and Think* (CUP 2012) and *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge* (CUP 2011). She is currently leading a project on *narrative science* (funded by the European Research Council), investigating the many functions that narratives play within the natural, human and social sciences: https://www.narrative-science.org/. She has a long-standing interest in the ways that economic and social science numbers both open windows onto our world and at the same time are used to effect change in the world.

Data Journeys Beyond Databases in Systems Biology: Cytoscape and NDEx

William Bechtel

Abstract This chapter investigates how data travels beyond databases in cell biology by focusing on Cytoscape, a platform that has been developed to represent networks, and NDEx, a database that allows for the reuse of network representations. I begin with a brief review of the databases that have been developed for data involving, for example, protein-protein interactions, that are relational and hence productively represented in networks. Given the amount of data stored in modern databases, raw network representations are typically hairballs that provides researchers little useful information other than that lots of things interact. Cytoscape was created by systems biologists to facilitate moving beyond hairballs to informative representations. It provides tools for clustering nodes and annotating them according to what is known about the objects represented. I provide examples of how Cytoscape has been deployed to develop new knowledge about biological mechanisms. Cytoscape has been made freely available, and I describe how a large interational community of researchers has created Apps that enable researchers to make a number of more specialized inferences. NDEx, created by members of the same research lab, serves as an Expo for networks—researchers can share networks they have developed and other researchers can search for networks and made them the basis for further incorporation of data or analyses.

1 Introduction

As in many fields, contemporary biologists generate vast amounts of data. Increasingly, this data is stored in large, on-line databases that procure data from curation of published literature and from high-throughput experiments. There it is accessed by researchers distinct from those who produced the data. Leonelli (2016; this volume) has developed the useful metaphors of *data travel* and *data journeys* to characterize this process of data movement. Much of the work on data journeys to

W. Bechtel (✉)
Department of Philosophy, University of California, San Diego, CA, USA
e-mail: bill@mechanism.ucsd.edu

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_7

date has focused on the preparation and travel of the data themselves, with less attention paid to the resources that are employed to analyze the data after they travel.[1] When the data specifies relations (causal, co-occurrence, etc.) between entities, this analysis often involves the construction of network diagrams in which entities are represented as nodes and relations between them as edges.[2] In the course of research, network diagrams are subject to various manipulations designed to reveal additional patterns in the data. Beyond their use in individual research projects, these networks themselves travel, providing the foundation for yet other research projects in which they are subject to further manipulation. Network diagrams are one format in which data are physically instantiated and subject to mutation as they are incorporated into network diagrams and passed on to other researchers (see Leonelli, this volume, for discussion of how data are mutated in the course of data journeys).

My focus will be on the tools that systems biologists have created to construct and operate on network diagrams and to enable networks themselves to travel. Anyone could construct a network diagram by hand from a body of data using a standard graphics package. However, such a process is laborious and the product is frozen—the researchers cannot then integrate data from additional sources or transform the diagram to reveal new patterns. Accordingly, researchers have developed software tools for creating, analyzing, and disseminating network diagrams. In Sect. 4 I will discuss Cytoscape, the most widely used platform for constructing network diagrams in systems biology. While developed in a systems biology framework, Cytoscape has itself traveled to and is actively used in numerous other scientific fields. Cytoscape provides a platform on which researchers with specific analytic needs can develop their own add-ons, referred to as apps. In Sect. 5 I will describe several apps and, using them, illustrate some of the analysis strategies employed in systems biology. In Sect. 6 I will describe the recent development of NDEx, which serves as an online exposition (expo) to which networks themselves can travel so as to be viewed by others and selectively taken up for additional journeys. As a background for focusing on network diagrams, I begin in Sect. 2 by introducing the types of data used to construct network diagrams in systems biology and in Sect. 3 describe the public databases and ontologies from which researchers extract data to create and analyze networks.

[1] Leonelli (2016, chapter 6) provides a pioneering examination of reuse. See chapters by Tempini, Chap. 13, Morgan, Chap. 6, and Griesemer, Chap. 8 in this volume, for other aspects of reuse. Tempini addresses the reuse of data for different objectives than that for which it was collected, and in particular focuses on how this often involves linkage of data from different sources such as between weather, environment, and health data. As he demonstrates, this requires manipulations that attenuate the differences due to where the data originated.

[2] Networks are just one mode of downstream analysis of data. See Cambrosio et al., this volume, for an account of knowledgebases that tailor large datasets for particular clinical applications.

2 Data Production: From Individual Experiments to High-Throughput Experiments

Through most of the twentieth century, experiments in fields like cell and molecular biology were conducted one at a time. But many of the procedures used in these experiments lent themselves to automation so that multiple variants on an experiment could be conducted in parallel. For example, when Sanger first developed techniques for sequencing amino acids in the 1950s or nucleic acids in the 1970s, he applied them to one protein or gene at a time. By the late 1980s these techniques were automated and by the 1990s automation made possible the sequencing of whole genomes of numerous species. Sequencing data identifies proteins and genes, but not what they do. Automated procedures enabled procuring other types of data related to function such as techniques that reveal whether proteins form complexes either with other proteins or with DNA or whether genetic mutations interact epistatically. I discuss only techniques detecting whether proteins can form complexes.

Much of the early twentieth century research focused on the reactions individual proteins catalyze, but in the second half of the twentieth century it became increasingly evident that proteins form complexes with each other and these are important to their catalytic function. Two techniques have proven especially useful in enabling high-throughput studies of protein-protein interactions (PPIs). The first, the yeast two-hybrid technique introduced by Fields and Song (1989), begins by transfecting yeast cells with two plasmids, each attaching to a different protein. One serves as the bait and the other as the prey and when the proteins to which they are attached interact with each other, the two domains are united and form a functional transcription factor that initiates transcription of a reporter gene. This technique identifies pairs of proteins that *can* interact, but many pairs do not do so in a given cell type. An alternative technique, affinity purification followed by mass spectrometry, starts with proteins that are actually bound into a complex in a cell and uses mass spectrometry to determine their identity (Rigaut et al. 1999). This approach identifies stable multi-protein interactions that actually occur in the cell. On the other hand, it misses more transient interactions that form and dissolve as cells carry out activities. As a result, both approaches to obtaining PPI data are actively employed.

High-throughput techniques for performing PPI studies were created shortly after automated gene sequencing was introduced and provided a means to study many of the novel genes they revealed. In the first high-throughput attempt to identify PPIs in yeast, Uetz et al. (2000) chose 192 proteins to use as baits and mated them with 6000 prey proteins. They identified 957 interactions between 1004 proteins. The following year Ito et al. (2001) performed an even larger-scale study, identifying 4549 interactions between 3278 proteins. Surprisingly, there was little overlap with the interactions identified in these two studies. I return to the Uetz et al. and Ito et al. studies to show how they were used in a pioneering network study in the next section.

3 Data Travels in Systems Biology: Databases and Ontologies

As biologists generated increasing volumes of data, they established publicly accessible databases to make this data accessible. The first databases were created for protein and gene sequence data. Dayhoff created the *Atlas of Protein Sequence and Structure* (Dayhoff and Eck 1965-1972) which she published in book form. Shortly after her death in 1984 it was made available electronically as the Protein Information [originally Interaction] Resource's Protein Sequence Database. It eventually merged into UniProt, which continues as a major source of information about proteins (The UniProt Consortium 2017). GenBank was developed in the same period for gene sequence data. Many additional databases for different types of biological data soon appeared—in 1989 the Listing of Molecular Biological Databases identified 50 databases (Lawton et al. 1989) and the number has continued to grow ever since. Annually, the first issue of *Nucleic Acids Research* reviews new and updated databases. On its website it provides a compilation of current databases, totaling 1613 in 2019. As Leonelli (this volume) notes, this process is both uncontrolled and unsustainable. In fact, each year the *Nucleic Acids Research* compilation annually eliminates discontinued URLs, including 147 in 2019.

Two of the early databases to include PPI data were the Yeast Proteome Database (YPD) and the Martinsried Institute for Protein Sequences (MIPS) database of protein interactions. A study by Schwikowski et al. (2000) illustrates how these databases were employed to construct a network from which new knowledge about yeast was extracted. They combined data from YPD and MIPS with data from the two high throughput studies noted at the end of the last section, yielding information on 2709 interactions involving 2039 proteins. Employing hierarchical clustering based on functional assignments found in the YPD and a layout procedure that located similarly connected nodes near each other, Schwikowski et al. identified one large connected network, shown in Fig. 1, plus 203 much smaller networks. In cases in which YPD contained information about a protein's cellular role, the researchers encoded it using the color of nodes: blue for membrane fusion, grey for chromatin structure, green for cell structure, yellow for lipid metabolism, and red for cytokinesis. By zooming in on parts of the network, as in panel B, they could focus on interactions between proteins that performed similar cellular roles, in this case membrane fusion, lipid metabolism, and cell structure.

An important question about any network diagram is whether the patterns it reveals are informative or simply an artifact of the representation strategy the researchers employed. To investigate this, Schwikowski et al. started with a given node to which a cellular role was assigned and asked how often one of the nodes with which it was connected in the network was assigned the same cellular role. This happened 72% of the time, (compared with, on average, 12% for scrambled networks). The authors present this as vindicating the network—had they not known the cellular role of the initial protein, they could have predicted it correctly 72% of the time based on the roles of its neighbors.

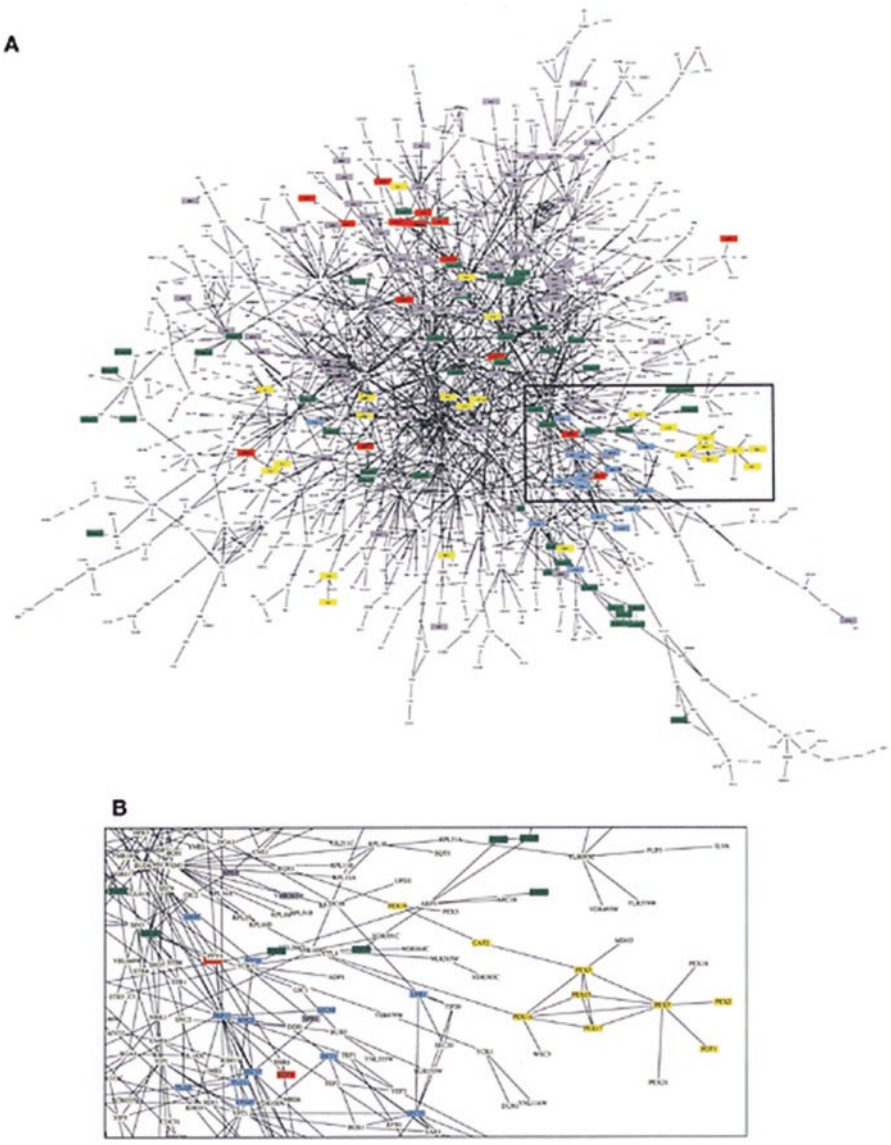


Fig. 1 Network diagram of protein interactions in yeast constructed by Schwikowski et al. 2000 drawing both upon results of high-throughput yeast two-hybrid studies and data from low-throughput studies collected in the MIPS and YPD databases. Reprinted by permission from Springer Nature: *Nature Biotechnology*, A network of protein-protein interactions in yeast, Schwikowski et al. 2000

As researchers recognized the usefulness of drawing upon large datasets in their research, many researchers created their own databases, tailored to their interests, and made them publicly available. These included the Database of Interacting Proteins (DIP) (Xenarios et al. 2000), MINT (Zanzoni et al. 2002), BIND (Alfarano et al. 2005), HPRD (Peri et al. 2003), BioGRID (Breitkreutz et al. 2003a), and IntAct (Hermjakob et al. 2004b). The infrastructure for each was relatively small—on average, they employed two full-time curators who read published papers and entered the data. In addition to primarily serving the interest of a particular laboratory, each database developed its own data structures and procedures for downloading and curating data. No single database could keep up with the rapid appearance of new datasets. As a result, researchers who wanted to use PPI data often combined data from multiple databases, developing their own tools (parsers, etc.) to do so. Recognizing the problem users faced, the curators of several databases collaborated to develop a standardized format (Hermjakob et al. 2004a). A standard format, however, made another problem even more salient. In reporting data, journal articles often failed to supply sufficient information about the entities studied or the experimental procedure used. This information is crucial for others to use and interpret the data (see Leonelli 2016, chapter 4; Rogers and Cambrosio 2007; and Boumans and Leonelli, this volume). Accordingly, the consortium generated guidelines as to the minimal information required in reporting a PPI experiment (Orchard et al. 2007). Several of the databases also began to work directly with journals so that data in new publications could be directly added to the databases. These efforts ultimately led to the development of the International Molecular Exchange (IMEx) Consortium, which among other initiatives introduced a deep curation standard aiming "to capture the full experimental detail provided in the interaction report, as this is often essential to assess interaction context and confidence" (Orchard 2012, p. 347). The initiative also sought to address another problem, that of maintaining funding for the various databases. The IMEx consortium also provided that if a member can no longer curate its databases, its records would be turned over to another member. Accordingly, when MPIDP ceased its curation efforts in 2012, it turned its records over to IntAct, which has subsequently maintained and updated them.

PPI databases have provided the data for constructing networks, but another database created during the same period, Gene Ontology (GO), has played a crucial role in allowing biologists to interpret networks. The motivation for developing GO was to develop "a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products" (Ashburner et al. 2000, p. 26) represented in the databases that had been developed for different model organisms (initially yeast, fruit fly, and mouse). GO comprises three ontologies, one for biological processes, another for molecular functions, and a third for cellular components, each providing general terms, organized hierarchically, that can be used to annotate individual genes. These ontologies are themselves undergoing continual revision and development (Leonelli 2010, 2016).

By 2000 systems biologists had a rich set of databases on which they could draw. Some, such as GenBase and UniProt, emphasized structural knowledge, but many focused on relational information, including PPI data. GO provided a common lan-

guage for annotating the entries in the different databases. These are the raw materials from which systems biologists constructed network diagrams with the goal of developing new biological knowledge.

4 Cytoscape: A Platform for Generating and Analyzing Network Diagrams

Tables in databases are great for storing and organizing data, but it is often difficult for humans to examine data tables directly and draw biologically meaningful inferences or even figure out what algorithms they might employ to generate inferences.[3] For this reason, most of the databases include a self-developed program to display the results of searches as network diagrams. These, however, typically employ a fixed format designed by the curators of the database.[4] Individual network formats support some inferences but not others. In order for users to leverage the vast amount of data contained in these databases, they need to generate network representations appropriate for their needs (see Leonelli, this volume, for a discussion of the relational nature of data).

Although several programs for creating network diagrams, including Osprey, VisANT, Gephi, and GraphViz, were developed in the first decade of the twenty-first century, Cytoscape (Shannon et al. 2003) has emerged as the most widely used. Ideker and his collaborators at the Institute for Systems Biology began developing Cytoscape in late 2001 for their own research and publicly released Cytoscape 0.8 as an opensource platform in June 2002. When Ideker moved to the University of California, San Diego, it became the center for Cytoscape development. The local team of 3–5 developers collaborates with numerous other developers at other institutions (currently including the Academic Medical Center in Amsterdam, the Institute for Systems Biology, the Institute Pasteur, the Gladstone Institute, the University of California, San Francisco, and the University of Toronto).

Although it is hard to measure actual use, in 2018 Cytoscape was downloaded on average 17,600 times per month and started on users' computers about 5000 times each day. According to Google Scholar, the standard reference used to acknowledge Cytoscape, Shannon et al. (2003), has been cited more than 14,750 times as of September 2019, most often by papers that include a network diagram generated with Cytoscape. These numbers likely significantly underestimate how frequently Cytoscape is used since many users do not explicitly acknowledge it (just as most people do not acknowledge Microsoft Excel or Adobe Illustrator even if they made extensive use of these in their research).

[3] Tables, though, sometimes enable viewers to visualize data. See Müller-Wille and Porter (this volume) for examples.

[4] The exception is BioGRID, whose developers also created Osprey, a network visualization program (Breitkreutz et al. 2003b). However, development of Osprey has ended and its webpage suggests researchers use Cytoscape.

Cytoscape, now in version 3.7.1, is an open-source, freely available java-based software package that runs on individual computers. It is a key platform of the National Resource for Network Biology and its development team continues to add new features to facilitate investigations directed at a range of topics such as representing networks at multiple scales and representing dynamic changes in cellular network organization in disease. An even larger community of computationally oriented biologists from around the world generates apps (initially referred to as plug-ins) that extend Cytoscape's capacities for analyzing networks. These are made available through the Cytoscape App Store, hosted on the Cytoscape website (http://cytoscape.org). In this section I will describe how Cytoscape is used to construct and modify network diagrams. In the subsequent section I will discuss apps and how they support analyses of networks.

Figure 2 provides a schematic overview of the Cytoscape architecture. The Cytoscape Window contains both the tables of node and edge attributes, from which Cytoscape constructs the network diagram, and the network diagram itself. Other components operate on the tables and graphs. I will not elaborate on the Graph Editing and Selection component. It performs functions much like those contained in the File and Edit components of word processing programs: opening stored networks or creating new ones, selecting, deleting or hiding, or copying nodes or edges, etc.

Visual Mapper (later termed VizMapper and in Cytoscape 3.5 renamed Style) and the Layout Engines take their input from the Node and Edge Attribute Tables. An Edge Attribute Table is shown in the screenshot in Fig. 3; a similar table defines

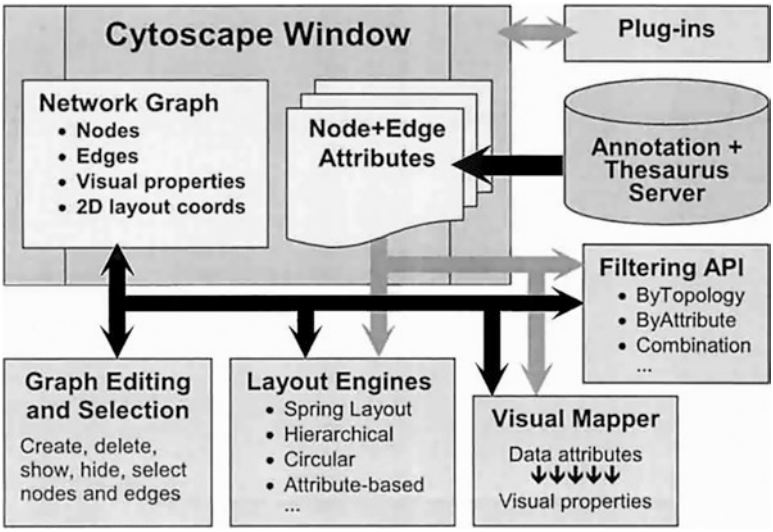


Fig. 2 Schematic overview of the Cytoscape architecture reprinted from Shannon et al. 2003. Although the labels for some of the components have changed, the overall architecture has not. Reprinted with permission of Trey Ideker

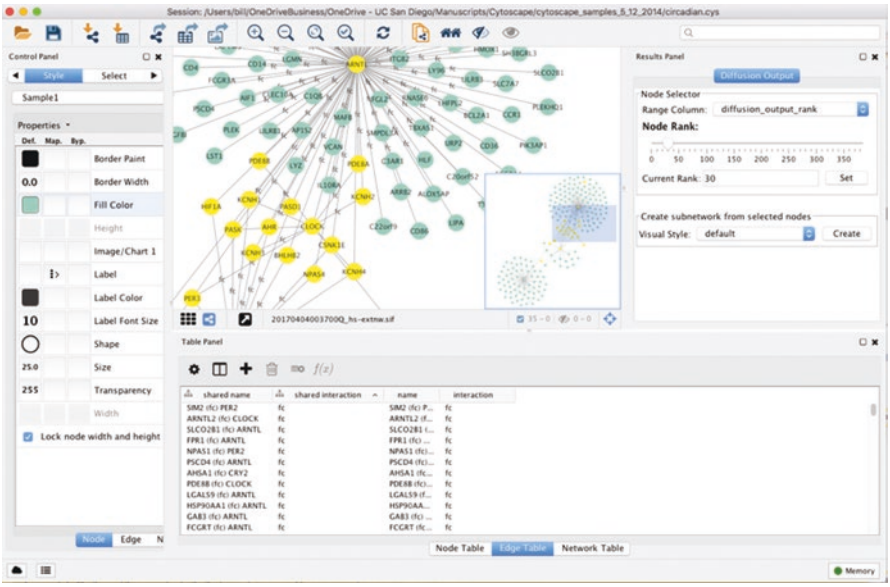


Fig. 3 Screenshot of Cytoscape 3.5. The window at the bottom shows the Edge Table from which the diagram in the upper window is generated. The window on the left shows the assignments of visual properties to nodes in Style. Screenshot used with permission of Trey Ideker

the nodes. A researcher can generate these tables based on data he or she has collected or from data downloaded from one or more of the databases discussed in the previous section. At a minimum, these tables must identify the entities to be represented by the nodes and the relations to be represented by edges, but they may also identify a variety of attributes of the entity (e.g. its concentration) or relation. The tables can also include annotations (e.g., cell location or cell function) procured from sources such as GO.

Style, shown on the left in Fig. 3, maps features specified in the table unto visual properties of nodes and of edges. Thus, an investigator can map attributes or annotations specified in the node and edge tables to labels or to visible features such as shape, size, and color. If color, for example, is used to indicate biological processes as specified in GO and size is used to represent the level of expression of a gene, the viewer can quickly see patterns in how these attributes and annotations vary.

There are many ways to lay out nodes in a 2-dimensional representation—nodes can be positioned randomly, around a circle, in a grid, or in a hierarchical arrangement. It is often useful to group nodes by their values on a particular annotation such as biological process or cellular component. When used with a circular layout, this results in nodes that share an attribute being located close together around the circle. There is great flexibility in how nodes are laid out and the choice affects what patterns the researcher can identify. For example, it is easier to see that several nodes are highly interconnected or are all connected to another set of nodes when they are positioned near each other. Spring-embedded layouts do this by treating edges like

springs (Eades 1984): connected nodes that are far apart are drawn together, but if they get too close, they are repelled a bit. For each of these strategies for laying out nodes there are a variety of algorithms, each of which generates a somewhat different result. After an algorithm is applied, the user can also manually move one or a selected group of nodes. Researchers find it useful to try out different layout strategies to find one that generates interpretable patterns.

Since the goal of network analysis is to generate biologically interpretable results, researchers derogatorily refer to networks such as shown in Fig. 4a as *hairballs*. Although the data is represented, it is not presented in a manner that can be interpreted biologically. Merico et al. (2009) illustrate how, by altering visual features and layout in Cytoscape, to transform this hairball into an informative network diagram revealing components of mechanisms involved in chromosome maintenance and duplication in yeast (Fig. 4b). Figure 4a was generated from curated data of PPIs (represented as edges) from both low- and high-throughput experimental studies retrieved from BioGRID. The nodes represent proteins and their colors indicate their location in the chromosome: red, replication fork; green, nucleosome; blue, kinetochore; yellow, other chromosome components. The use of color in Fig. 4a is already a step away from a pure hairball, but the network diagram offers no mechanistic insight. By applying a spring-embedded layout in which edges are assigned forces so as to draw highly connected nodes closer together and yet keep them from getting too close, the authors transformed Fig. 4a into 4b. Being highly connected, the nodes for proteins in the kinetochore, nucleosome, and replication fork are now situated adjacent to each other. VizMapper (Style) used data about how much gene expression changes over the cell cycle to determine node size. In addition, the width of the edges is determined by the Pearson correlation between transcript profiles. Looking at the network diagram one can readily see that many green

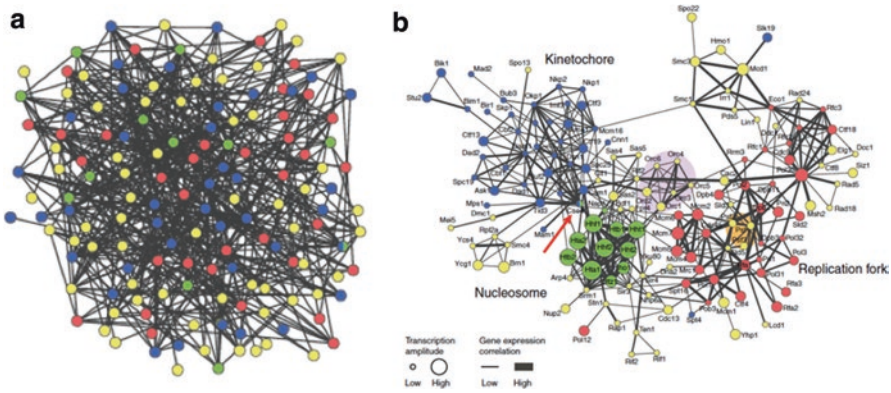


Fig. 4 (**a**) A hairball network diagram based on PPIs among proteins involved in chromosome maintenance and duplication in *Saccharomyces cerevisiae*. (**b**) The network has been transformed into an informative network diagram. Reprinted by permission from Springer Nature: *Nature Biotechnology*, How to visually interpret biological data using networks, Merico et al. 2009

nodes are large and connected with numerous thick edges, indicating that the expression of proteins in the nucleosome is changing together during the cell cycle.

Now that the nodes are laid out in an informative manner, a researcher can zoom in to local regions and make his or her own inferences about parts and operations. A commonly used inference strategy is guilt-by-association—if neighbors of a node without an annotation share a common annotation (in this case, for a cellular component), the researchers infer that the unannotated node should receive the same annotation. The three proteins shown in the region shaded in orange in Fig. 4b, Psf1, Psf2 and Psf3, are colored yellow since GO did not assign them a cellular component annotation below the level of chromosome. The layout procedure, however, situated them among the red nodes that have the replication fork annotation. Employing guilt-by-association, the researchers inferred these proteins should be assigned that annotation as well. Merico et al. report that although these proteins are not so annotated in GO, research already published showed that they belonged to the GINS complex in the nucleosome that is responsible for assembling the DNA replication machinery. Guilt-by-association led the network researchers to make a correct assignment.

The layout algorithm also enables the identification of new mechanisms. The nodes labeled Orc1, Orc2, Orc3, Orc4, Orc5 and Orc6 are located together (in a region shaded in violet) apart from the three regions of nodes annotated to cellular components. The authors infer that they form a distinct mechanism and report that although these nodes lacked specific annotations in GO, "they are known members of the yeast origin recognition complex (ORC), responsible for the loading of the replication machinery onto DNA" (p. 922). In this case again the inference is supported.

Cytoscape thus provides researchers the ability to transform tables into network diagrams, assign visible features to attributes and annotations of entities and their relations, and determine how the nodes and edges will be laid out. Exploration with different approaches (e.g., changing whether an attribute is represented by the shape or color of nodes) is often important to finding informative patterns. This would be very cumbersome if researchers had to construct each network diagram by hand but relatively easy with Cytoscape.

5 Further Analyzing Networks: Cytoscape's App Store

As I have noted, Cytoscape provides a platform for other researchers to construct apps to perform specific analyses for their own purposes but also make the resulting apps available to others. In this way Cytoscape serves multiple groups of users who have different research agendas and require different tools for their execution. Many of the apps are the focus of journal publications that describe the procedures employed in the app and one or more examples of its use (I have identified such publications for several of the apps discussed below). In Spring 2017 there were

more than 180 apps in the Cytoscape App Store that work with Cytoscape 3.X.[5] Some apps support the import and integration of data from specific databases that researchers might wish to represent in networks. For example, KEGGScape, GeneMania, ReactomeFIViz, and STRING, draw results from these different databases into Cytoscape. Bisogenet integrates and imports data from multiple databases such as DIP, BIOGRID, BIND, MINT, and IntAct. AgilentLiteratureSearch allows users to directly query published literature for PPIs and incorporate the results into a Cytoscape network. Apps such as BiNGO and ClueGO facilitate annotation of nodes and edges using Gene Ontology.

Yet other apps provide layout and visualization algorithms that extend beyond what is offered in the core. For example, Cy3D generates three-dimensional views of networks while CyAnimator supports the construction of animations. With respect to layout, GOlorize enables the use of GO annotations to direct the layout of nodes so that the network is interpretable in terms of biological functions while DeDaL facilitates using principal components analysis in developing layouts, aligning one network with another, and morphing between selected layouts so as to find ones that are biologically interpretable.

Yet other apps support particular analyses of networks useful for specific lines of research. I will first discuss two classes of analysis apps: those used to compute a variety of standard network measures and those designed to identify clusters or modules in a given network. I will then offer two illustrations of how particular apps contribute to a better understanding of biological processes.

Apps for Computing Network Measures Graph theorists have developed an extensive set of measures to characterize networks. For purposes of this exposition, I will focus only on networks with undirected edges. Some of the most common measures are *mean shortest path length*, the *clustering coefficient*, and *node degree distribution*. The length of a path between two nodes is the number of edges that are traversed in going from one to the other; the mean shortest path length is the mean for all pairs of nodes of the shortest (or characteristic) path lengths between them. It provides a measure of how quickly effects can travel through the network. The nodes to which any given node is connected are its neighbors and the clustering coefficient characterizes the degree to which the neighbors of a node are connected to one another. Finally, node degree refers to the number of connections a given node has to other nodes. Of particular interest are networks in which node degree is not distributed normally but according to a power law. In such a case, some nodes are highly connected to other nodes, and serve as hubs, whereas most nodes have few connections. NetworkAnalyzer (Assenov et al. 2008) computes these and many other statistics that are used to characterize networks, displaying the results in histograms or scatterplots. Apps such as CytoHubba identify hubs.

[5] Another 132 Apps were written for Cytoscape 2.X but have not been recoded to work with Cytoscape 3.X. This was a serious cost of completely revising the Cytoscape's program interface in 2013, which was done in part to improve the architecture through which apps interact with the core program.

Apps for Identifying Clusters For many research objectives it is valuable to identify nodes that are especially highly interconnected. These clusters, sometimes referred to as *modules*, often reflect groups of components that perform a common activity—that is, work as a mechanism. The apps Molecular Complex Detection (MCODE) (Bader and Hogue 2003) and ClusterMaker2 (Morris et al. 2011) identify clusters. Modules may be organized hierarchically, sometimes with different types of connections at different levels. When Bandyopadhyay et al. (2008) developed a network based on both PPI and genetic interactions they found that PPIs tended to link nodes in modules while genetic interactions generated higher level clusters. Srivas et al. (2011) implemented the procedure Bandyopadhyay et al. employed in the app PanGIA.

5.1 Applying an App for Identifying Active Modules

Most clustering algorithms view networks as static structures, but Ideker et al. (2002) sought to identify nodes that organize into clusters or mechanisms only in specific circumstances such as when particular genes are mutated or yeast are grown on specific media. In an earlier paper, Ideker et al. (2001) has investigated the galactose (GAL) utilization mechanism in yeast. They started with PPI and protein-DNA interaction data to construct a network of 348 genes with 362 interactions. They grew colonies of wild-type and nine mutant strains, each lacking one known GAL gene, on media containing or lacking 2% galactose, measured global mRNA changes and protein concentration changes across the conditions, and plotted these on the network. As Cytoscape had not yet been developed, they used the LEDA toolbox developed at the Max-Planck-Institut für Informatik (Mehlhorn and Näher 1999) to construct the network shown in Fig. 5a. Arrows represent protein-DNA interactions and straight edges PPIs. The nodes are shown in clusters corresponding to genes that exhibited similar changes in expression over all perturbations and the clusters are labeled by their biological functions. Darker shading of nodes indicates increased and lighter shading decreased expression. The size of the nodes reflects the magnitude of change in the case in which gal4 (the node colored in red) is knocked out in the presence of galactose. The network diagram reveals that the expression changes resulting from the perturbation is more correlated in connected proteins than among randomly selected proteins, a result Ideker et al. further confirmed with statistical analysis.

In the 2002 study, Ideker et al. sought to identify modules in which expression changed the most in specific conditions. Having developed Cytoscape, they represented the network in it and developed an analysis strategy that became one of the first Cytoscape apps, jActiveModules. The analysis first computes a z-score for the degree of change in expression of each gene in a particular condition, indicated by the shading of the nodes in Fig. 5b. It then identifies subnetworks of genes under or over expressed and rank-orders them in terms of activity. The top five subnetworks are indicated in Fig. 5b by common coloring of the node border and the attached

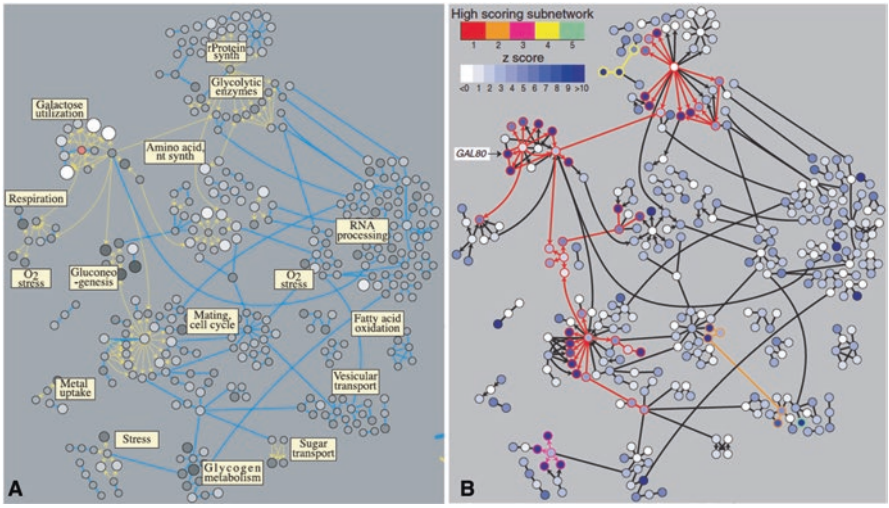


Fig. 5 Comparative network diagrams: (**a**) from Ideker et al. 2001 and (**b**) from Ideker et al. 2002. Both show the same 362 associations between genes whose expression was increased or decreased when grown with or without 2% galactose. In the diagram on the left, darker nodes indicate increased expression when gal4 (shown in red) is knocked out. The edges shown in color other than black in the diagram on the right indicate the subnetworks that were most altered when gal80 was knocked out. A. From Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science, 292*, 929–934. Reprinted with permission from AAAS. B. reprinted from Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F., Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics,* 2002, Volume 18 Suppl 1, S233–240, by permission of Oxford University Press

edges. Ideker et al. interpret the subnetworks active in a particular condition as mechanisms involved in transmitting signals and performing regulatory functions. In the example shown, GAL80 (the only labeled node) is deleted. The adjacent node, GAL4, is a hub with protein-DNA connections to seven other genes. This suggests the hypothesis that GAL80 influences these genes through its effect on GAL4, a hypothesis for which there was already empirical support (Lohr et al. 1995).

5.2 Applying an App for Modeling Diffusion

Whereas jActiveModules was one of the first apps developed for Cytoscape, Diffusion (Carlin et al. 2017) is one of the most recent. Diffusion implements a distinctive strategy for discovering underlying clusters that correspond to mechanisms that has proven effective in fields such as cancer research in which researchers confront extremely heterogeneous data. For example, in The Cancer Genome Atlas study of 500 tumors of various types, individual tumors exhibited from 20 to 300 somatic mutations, with the genes mutated varying substantially across samples of

the same type of tumor. This made it difficult to determine which mutations might play a causal role. To address this problem, Vandin et al. (2011) developed a strategy of mapping mutated genes onto a PPI network and treating them as hot spots from which simulated heat could diffuse. In many cases, heat diffusing from different nodes would converge on the same cluster of nodes. These nodes were hypothesized to represent a mechanism or pathway that, when disrupted through any of the mutations, leads to cancer. The approach was further developed by Hofree et al. (2013), who used propagation in networks to stratify cancer populations in ways that corresponded to patient survival. Heat diffusion algorithms are computationally extremely demanding. Thus, the designers of Diffusion linked the app locally installed on an individual researcher's computer to an internet service that performs the computation. Using Diffusion within Cytoscape, the user can visually select nodes as heat sources, invoke the service, and then visualize the diffusion results.

Carlin et al. employed Diffusion to better understand why one melanoma cell line responds to the drug Vemurafenib (LOX-IMVI) while another is resistant. They use a network generated from the NCI Pathway Interaction Database (an amalgamation of expert-curated cancer pathways) and initiated diffusion from six genes with known relations to the drug: *BRAF*, *PDGFRB*, *NRAS*, *HGF*, *MAP 2 K1*, and *MAPK1*. Diffusion identified a subnetwork of 53 nodes and 448 edges. Cytoscape was then used to filter the top 10% of nodes activated after diffusion. Based on combining the results of multiple queries followed by filtering, Carlin et al. determined that *TSC2* and *BLNK* are mutated in the resistant but not the sensitive cell lines and proposed that this might explain the difference.

6 Network Expo: NDEx

In the previous two sections I have characterized how tools like Cytoscape allow for data that has traveled to databases to travel one step further and be used in network analyses. But is that the end of the line? In this section I show how network diagrams themselves can also travel. Traditionally, network diagrams have been distributed as static visual representations and those who wanted to analyze them further had to recreate them for themselves. But networks generated with Cytoscape and similar programs can be stored in structured data formats in which they can then be distributed to other users, who may then incorporate additional data into the network or perform a different type of analysis (e.g., a different clustering procedure) to the existing network. While such sharing can be carried out informally by authors,[6] the Network Data Exchange (NDEx) is providing a platform for doing this on a large scale.

[6] A collaboration between Elsevier and Cytoscape created the Interactive Network Viewer which allowed authors to make networks available in online publications in a viewer with some capacities for readers to further explore the network or download it to Cytoscape. This project is no longer active.

NDEx was introduced in 2014 as "an online commons" (Pillich et al. 2017) or expo that functions much like World Expos. In this case, the exhibits are the networks that provide original interpretations of data. By uploading their networks, researchers can showcase them and others can download them for use in their own work. The developers further characterize NDEx as "a step toward an ecosystem in which networks bearing data, hypotheses, and findings flow easily between scientists" (Pratt et al. 2015). The project employs its own group of developers in Ideker's lab at UC San Diego and is supported by the National Cancer Institute, the National Resource for Network Biology, the California Stem Cell Agency, Pfizer, Janssen, and Roche.[7]

At its core, NDEx functions much like Google Docs or Dropbox. Networks are added to NDEx either from other online sources such as Pathway Commons, which draws data from a wide range of databases including BIND, DIP, and BioGRID that were discussed in Sect. 2, or by individual users via either direct file import or from Cytoscape. Individual users store their own networks and have control over who can access them—they can keep them private, share them with designated others, or make them public. Sharing with a group of researchers allows a group to collaborate in further developing a network. If made public, other users might use the network as the basis for their own work and upload new versions for others to access. Each network that is added to NDEx is assigned a Universally Unique Identifier (UUID) so that it can be easily referenced. If someone modifies a public network and saves it, it is assigned a new UUID. NDEx is distinct from other online network repositories such as KEGG and Pathway Commons in that users manage their own networks rather than the networks being managed by the organization that maintains the resource. To facilitate visualizing and indexing networks as well as interactions with Cytoscape, NDEx employs the Cytoscape Cyberinfrastructure network exchange format, CX, to store information. CX, however, maintains the semantics of the format employed by the creator of the network.[8]

For networks to be useful to others, it is important that depositors provide sufficient information about how they were created and the data that was used (databases are updated regularly and attempts to reconstruct networks will not necessarily yield the same results unless the same iteration of the database is used). Accordingly, NDEx maintains a provenance history that contains this information. The history also includes information about other networks that were used in constructing a particular network.

For NDEx to provide a useful expo, other users must be able to find networks that are relevant to them. Thus, when networks are uploaded, NDEx indexes text strings for network descriptions, the user and group that manages the network, the

[7] Legally, the Cytoscape Consortium, a 5.0.1cs corporation, owns Cytoscape and NDEx, along with NeXO and Cytoscape.js. It contracts with the various pharmaceutical companies and subcontracts with UC San Diego.

[8] WikiPathways provides a useful comparison case with NDEx. WikiPathways is based on the Wiki model in which everyone collaborates on a common public document. It is also limited to small networks and allows for content that is not represented in a network.

genes or proteins represented by the nodes, the relations represented by the edges, and references cited. Users can initiate searches from NDEx homepage by entering names of cell processes or names of genes or proteins. This will bring up a table listing a number of networks. Figure 6 shows the results of a search for three circadian genes, *per2*, cry2, and *bmal1*. This returned 165 networks in which at least one of these genes is included. The table shows the name of the network, the number of nodes and edges, whether the network is public or private, the owner, and the date it was last modified. When one hovers a mouse over the name of a network, a popup window appears with a description of the network if one has been provided. If there is an icon in the Ref. column, it links to a publication in which the network appeared. One can proceed to download the network by selecting the icon with a white downward arrow.

Clinking on a network name brings it up in a window (if there are too many edges, a sample of 500 edges will be displayed). Users can choose instead to see a listing of the edges in a table view. The screen also shows either network info (e.g., when it was created, its UUID address) or the provenance history. A search box enables users to query particular nodes and select a number of edges out from those nodes. The network selected in Fig. 6 has 195 nodes and 4534 edges. Entering CRY2 and distance 1 returns the more restricted network shown in Fig. 7. Selecting the nodes PER2, CRY2, and the two edges connecting them, brings up information about the nodes, including links to UniProt, GenBank, and publications providing evidence for the edges.

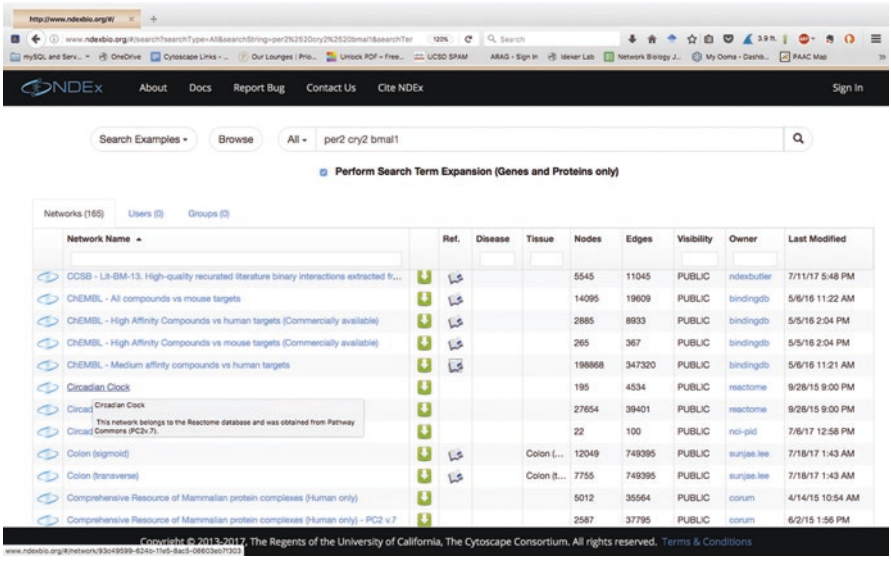


Fig. 6 Screen shot of NDEx after search for networks that include *per2*, *cry2*, or *bmal1*, three prominent mammalian circadian genes

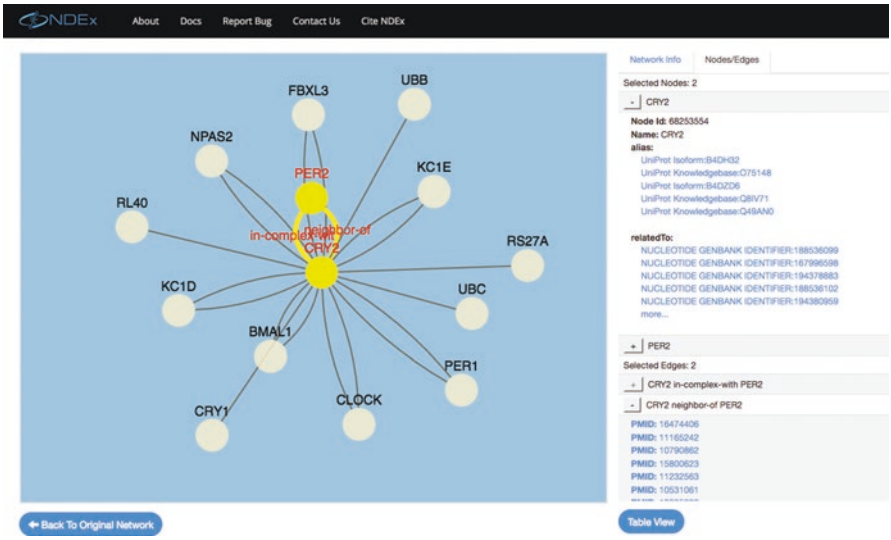


Fig. 7 Screen shot of the network selected in Fig. 6 after a query requesting nodes directly connected to CRY2

NDEx has been designed to integrate smoothly with Cytoscape. From within Cytoscape, one can use the app CyNDEx to query networks in NDEx and import selected ones. CyNDEx also allows users to export networks developed or modified in Cytoscape to NDEx. Once a network has been imported from NDEx to Cytoscape, a researcher can use it to continue the inquiry for which it was originally designed by carrying out additional analyses or accept the analysis offered and incorporate further data into the network.

The developers of NDEx have advanced a bold vision of how NDEx can provide "new models of scientific publication." It provides an expo "in which live data structures replace static diagrams and supplemental files." Drawing upon these live data structures, other biologists can create new networks that serve their own ends and create new expositions in NDEx. For NDEx to realize these goals, network biologists must be willing to share their networks. There is evidence that they will as use of NDEx is showing steady growth. From July 2015 until March 2016 the number of unique visitors per month increased from 151 to over 1200. As of July 2017 there were 3190 public networks, 810 registered users and 37 groups, although not all of these have uploaded networks to NDEx. The developers are pursuing a number of strategies to encourage greater use such as making NDEx a platform on which authors may make networks in their papers available to reviewers. To the extent that NDEx is successful as an expo of networks, network diagrams will be both products of inquiry and inputs for future inquiries.

7 Conclusions

In systems biology and many other fields, relational data travel from individual researchers to publicly accessible databases, from which they are accessed and employed by subsequent researchers. I have focused on the resources that systems biologists have created to enable further data journeys. These resources are allowing researchers both to represent and extract interpretations from the data and to share the products of their research so that other researchers can build upon them. These tools enable data and the analyses constructed from them to continue to travel far beyond the initial database to which they were uploaded.

My focus has been on the increasingly popular use of network representations of relational data. Networks are not just an attractive format in which to represent data. As I have developed in earlier publications, they are employed in novel ways to make discoveries about biological mechanisms. In recent decades, philosophers of biology have characterized the research strategies by which biologists in a variety of fields search for mechanisms to explain phenomena of interest (Bechtel and Richardson 1993/2010; Craver and Darden 2013). Most of these strategies start with hypothesized mechanisms and decompose them to find their constituents. Network biology pursues a different strategy, starting with data about how biological entities are related to each other (e.g., which proteins interact), identifying mechanisms as local clusters within the network and appealing to them to explain biological phenomena (Bechtel 2017, 2019).

Key to network biology is the construction of network representations and the application of tools to analyze these representations. Since its introduction in 2002, Cytoscape has emerged as a freely available and widely used platform for creating and analyzing network representations. The core of Cytoscape allows researchers to import databases of relational data and generate network representations employing a variety of different layouts that enable specific inferences from the data and different ways to annotate the representation to incorporate yet additional information. A user can, for example, quickly switch between different layouts until he or she finds one that provides insight into the data. Of central importance are algorithms used to find clusters of nodes that are then interpreted as potential mechanisms.

The construction of a revealing network representation is often just the starting point for further analysis. The core of Cytoscape provides a range of tools intended for use on a wide variety of network studies (extending, for example, to the social sciences). But Cytoscape also provides a platform for other researchers, often with interests limited to specific domains, to develop their own analytic tools in the form of apps. By providing an App store, the developers of Cytoscape have encouraged researchers to make these available to yet other researchers.

Cytoscape and its apps are powerful tools for researchers to reuse data that has been deposited into the growing number of databases developed by biologists. A particularly valuable feature is allowing researchers to readily integrate data from a

variety of different databases into a single network that can then be analyzed in different ways. Until recently, however, these network representations and the analyses performed on them represented the end of data journeys—they might be published, but anyone who wanted to carry on the inquiry would have to procure the network in a useable format from the researchers or reconstruct it for themselves. By providing an easily searchable expo of networks that other users can access, add data to, and further analyze (using Cytoscape or another platform), NDEx enables data to travel yet further. Since users can both download networks and upload their revised network, data can be recirculated potentially indefinitely.

Resources such as databases, Cytoscape and its apps, and NDEx, constitute important infrastructures that are increasingly relied upon by contemporary biologists. These tools supplement traditional experimental tools, allowing results to travel widely and to be analyzed by multiple researchers using different techniques for network analysis. They thereby contribute in novel ways to the development of scientific knowledge.

Acknowledgements I thank the editors, Sabina Leonelli and Niccolò Tempini for helpful comments on this manuscript. Further, I thank Benjamin Sheredos, Rebecca Hardesty, Jason Winning, and other members of the Philosophy of Science in Practice Study Group at UC San Diego as well as participants in the workshop on Varieties of Data Journeys at Exeter University in November 2017 for their valuable comments and suggestions on earlier drafts of this paper. In addition, I thank Barry Demchak, former Project Manager for Cytoscape, and Dexter Pratt, Director of Software Development in the Ideker Lab, for their comments and suggestions and Trey Ideker for inviting me to sit in on his lab meetings.

References

Alfarano, C., C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, et al. 2005. The Biomolecular Interaction Network Database and Related Tools 2005 Update. *Nucleic Acids Research* 33: D418–D424.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, et al. 2000. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics* 25: 25–29.

Assenov, Y., F. Ramirez, S.E. Schelhorn, T. Lengauer, and M. Albrecht. 2008. Computing Topological Parameters of Biological Networks. *Bioinformatics* 24: 282–284.

Bader, G.D., and C.W. Hogue. 2003. An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinformatics* 4: 2.

Bandyopadhyay, S., R.M. Kelley, N.J. Krogan, and T. Ideker. 2008. Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data. *PLoS Computational Biology* 4: e1000065.

Bechtel, W. 2017. Using the Hierarchy of Biological Ontologies to Identify Mechanisms in Flat Networks. *Biology and Philosophy* 32: 627–649.

———. 2019. Analyzing Network Models to Make Discoveries About Biological Mechanisms. *British Journal for the Philosophy of Science* 70: 459–484.

Bechtel, W., and R.C. Richardson. 1993/2010. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

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.

Breitkreutz, B.J., C. Stark, and M. Tyers. 2003a. The GRID: The General Repository for Interaction Datasets. *Genome Biology* 4: R23.

———. 2003b. Osprey: A Network Visualization System. *Genome Biology* 4: R22.

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.

Carlin, D.E., B. Demchak, D. Pratt, E. Sage, and T. Ideker. 2017. Network Propagation in the Cytoscape Cyberinfrastructure. *PLoS Computational Biology* 13: e1005598.

Craver, C.F., and L. Darden. 2013. *In Search of Mechanisms: Discoveries Across the Life Sciences*. Chicago: University of Chicago Press.

Dayhoff, M.O., and R.V. Eck. 1965-1972. *Atlas of Protein Sequence and Structure*. Silver Spring: National Biomedical Research Foundation.

Eades, P. 1984. A heuristic for graph drawing. *Congressus Numerantium* 42: 149–160.

Fields, S., and O. Song. 1989. A Novel Genetic System to Detect Protein-Protein Interactions. *Nature* 340: 245–246.

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.

Hermjakob, H., L. Montecchi-Palazzi, G.D. Bader, J. Wojcik, L. Salwinski, et al. 2004a. The HUPO PSI's Molecular Interaction Format--A Community Standard for the Representation of Protein Interaction Data. *Nature Biotechnology* 22: 177–183.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, et al. 2004b. IntAct: An Open Source Molecular Interaction Database. *Nucleic Acids Research* 32: D452–D455.

Hofree, M., J.P. Shen, H. Carter, A. Gross, and T. Ideker. 2013. Network-Based Stratification of Tumor Mutations. *Nature Methods* 10: 1108–1115.

Ideker, T., V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, et al. 2001. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science* 292: 929–934.

Ideker, T., O. Ozier, B. Schwikowski, and A.F. Siegel. 2002. Discovering Regulatory and Signalling Circuits in Molecular Interaction Networks. *Bioinformatics* 18 (Suppl 1): S233–S240.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, et al. 2001. A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98: 4569–4574.

Lawton, J.R., F.A. Martinez, and C. Burks. 1989. Overview of the LiMB Database. *Nucleic Acids Research* 17: 5885–5899.

Leonelli, S. 2010. Documenting the Emergence of Bio-Ontologies: Or, Why Researching Bioinformatics Requires HPSSB. *History and Philosophy of the Life Sciences* 32: 105–125.

———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Lohr, D., P. Venkov, and J. Zlatanova. 1995. Transcriptional Regulation in the Yeast GAL Gene Family: A Complex Genetic Network. *The FASEB Journal* 9: 777–787.

Mehlhorn, K., and S. Näher. 1999. *Leda: A Platform for Combinatorial and Geometric Computing*. New York: Cambridge University Press.

Merico, D., D. Gfeller, and G.D. Bader. 2009. How to Visually Interpret Biological Data Using Networks. *Nature Biotechnology* 27: 921–924.

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.

Morris, J.H., L. Apeltsin, A.M. Newman, J. Baumbach, T. Wittkop, et al. 2011. clusterMaker: A Multi-Algorithm Clustering Plugin for Cytoscape. *BMC Bioinformatics* 12: 436.

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.

Orchard, S. 2012. Protein Interaction Data Curation: The International Molecular Exchange (IMEx) Consortium. *Nature Methods* 9: 345–350.

Orchard, S., L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, et al. 2007. The Minimum Information Required for Reporting a Molecular Interaction Experiment (MIMIx). *Nature Biotechnology* 25: 894–898.

Peri, S., J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, et al. 2003. Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research* 13: 2363–2371.

Pillich, R.T., J. Chen, V. Rynkov, D. Welker, and D. Pratt. 2017. NDEx: A Community Resource for Sharing and Publishing of Biological Networks. *Methods in Molecular Biology* 1558: 271–301.

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.

Pratt, D., J. Chen, D. Welker, R. Rivas, R. Pillich, et al. 2015. NDEx, the Network Data Exchange. *Cell Systems* 1: 302–305.

Rigaut, G., A. Shevchenko, B. Rutz, M. Wilm, M. Mann, et al. 1999. A generic Protein Purification Method for Protein Complex Characterization and Proteome Exploration. *Nature Biotechnology* 17: 1030–1032.

Rogers, S., and A. Cambrosio. 2007. Making a New Technology Work: The Standardization and Regulation of Microarrays. *The Yale Journal of Biology and Medicine* 80: 165–178.

Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of Protein-Protein Interactions in Yeast. *Nature Biotechnology* 18: 1257–1261.

Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, et al. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* 13: 2498–2504.

Srivas, R., G. Hannum, J. Ruscheinski, K. Ono, P.L. Wang, et al. 2011. Assembling Global Maps of Cellular Function Through Integrative Analysis of Physical and Genetic Networks. *Nature Protocols* 6: 1308–1323.

Tempini, Niccolò. this volume. 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.

The UniProt Consortium. 2017. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Research* 45: D158–D169.

Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, et al. 2000. A Comprehensive Analysis of Protein-Protein Interactions in Saccharomyces cerevisiae. *Nature* 403: 623–627.

Vandin, F., E. Upfal, and B.J. Raphael. 2011. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology* 18: 507–522.

Xenarios, I., D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, et al. 2000. DIP: The Database of Interacting Proteins. *Nucleic Acids Research* 28: 289–291.

Zanzoni, A., L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, et al. 2002. MINT: A Molecular Interaction Database. *FEBS Letters* 513: 135–140.

William Bechtel is Distinguished Professor of Philosophy and a Member of the Center for Circadian Biology and the Interdisciplinary Program in Cognitive Science at the University of California, San Diego. His research focuses on philosophical issues in cell and molecular biology, systems biology and circadian biology. In his book *Discovering Complexity* (1993/2010, with Robert Richardson), he argued that explanations in many fields of biology take the form of identifying a mechanism responsible for a selected phenomenon. He developed this perspective in detail for cell biology in *Discovering Cell Mechanisms* (2006) and for cognitive science and neurobiology in *Mental Mechanisms* (2008). In subsequent work with Adele Abrahamsen, he has explored the use of computational modelling to understand the functioning of interactive biological mechanisms such as the circadian clock that exhibit complex dynamic behaviour. He has also investigated strategies for network representation in systems biology and their use in discovering mechanisms within larger interactive systems. Most recently, he has expanded his focus on how productive biological mechanisms (e.g. muscles and metabolic pathways) are controlled by heterarchically organized control mechanisms within cells and multicellular organisms and how such control enables production mechanisms to support the autonomy of these organisms.

A Data Journey Through Dataset-Centric Population Genomics

James Griesemer

Abstract I describe a data journey drawn from a case study of research in human population genomics. The case is framed in dialogue with a project on what has been called the "re-situation" of scientific knowledge (Morgan 2014). The kind of journey described elicits a missing concept—"data*set*-centric" biology—in the conversation around the emergence of "big data" and data-centric biology (Leonelli 2016) and its contrast, "traditional" or "small data" biology. I distinguish data*point*-centric from data*set*-centric practices. The case study is about the development, use, and amendment of data sets in one lab's pursuit of human genome diversity studies. I offer a model of data journeys to interpret the case. The model is comprised of three kinds of components: scientific data structures, data representations, and data journey narratives. The case study illustrates two visualizations that frame the dataset journey.

1 Traveling Findings and Data Journeys in Human Population Genomics

In this chapter, I make a case for a "middle ground" landscape of data *set*-centric biology as an important setting for data journeys in twenty-first century science, adding "middle sized" facts to the big and the small (Howlett and Morgan 2011, Leonelli 2016). Communities of specialists in fields practicing dataset-centric biology are organized around data *sets* rather than dissociable, individually retrievable data *points*, even though the dissociability of the latter is key to the data journeys of dataset-centric biology. For dataset-centric biology, if datapoints are disaggregated from their context in a dataset, datapoints may lose value or meaning as datasets add value and change meaning. Scientific focus on datasets prods dataset-centric sciences down toward a "craft" scale of operation rather than up to an "industrial" scale: in dataset-centric biology, datapoints are not interchangeable parts, nor independently valuable "widgets" in a datapoint-as-product economy of science. At

J. Griesemer (✉)
Department of Philosophy, University of California, Davis, Davis, CA, USA
e-mail: jrgriesemer@ucdavis.edu

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_8

craft scale, datapoints are more like individualized parts of whole dataset products and less like anonymized members of possibly arbitrary or merely conventional sets.

In a broad sense, the data journeys in human population genomics of interest in this chapter begin with tissue specimen collection, proceed to extraction of DNA from specimens, and eventually result in sequencing, production of digital sequence records, and archiving of the records. My focus here, however, is on the journey *after* digital data is produced: how these records are collected into datasets that can travel (or not), just as Leonelli (2016) has documented how genomics datapoints can travel. These journeys must of course be planned, including developing protocols for subject sampling and specimen collection, but here I focus on journeys of datapoints and datasets derived from DNA already extracted and archived. After tissue collection and curation, extracted DNA specimens are allowed to circulate in a limited fashion to qualified research labs. The labs then conduct or arrange sequencing so as to use the digital data in a range of biomedical and ancestry studies. Once the data gets into digital form, the datasets can have a life of their own. This "workflow" can be summarized by distinguishing: (1) a "field" setting in which a study design is put into action to produce "data," (2) a lab setting in which specimens or data are put in motion to produce findings and reports, and (3) a community setting in which findings are put into circulation in various social worlds that become evaluated as "facts" or sent back into scientific workflows to be reworked, reinterpreted, reevaluated (Fig. 1). My case study focuses on the latter: the use of genomic DNA data to infer ancestry relations among human populations.

The case is part of a project on what has been called the "re-situation" of scientific knowledge (Morgan 2014). The kind of journey described elicits a missing concept—"data*set*-centric" biology—in the conversation around the emergence of "big data" and data-centric biology (Leonelli 2016) and its contrast, "traditional" or "small data" biology. I distinguish data*point*-centric from data*set*-centric practices. The case study is about the development, use, and amendment of datasets in one lab's pursuit of human genome diversity studies.

The data journey I re-trace here begins with sequence data analyzed in a paper by Noah Rosenberg et al. (2002) in *Science* magazine: "Genetic Structure of Human Populations." This paper reports "big findings," that is, findings about worldwide ancestry relationships derived from analysis of a substantial collection of datapoints in a dataset using advanced analytical methods and theoretical models. The paper also reports (or refers to) "small findings," e.g. findings of particular sequences detected in particular DNA samples. Some of the small findings are presented simply by citation of the datasets used in the analysis leading to the big findings, based on sequencing cell line panel DNA collected for the Human Genome Diversity Project (HGDP).

Data for the HGDP that supplied the Rosenberg lab came from 1064 lymphoblastoid cell lines (LCLs) cultured from blood samples collected from people of different localities or regions around the world by a variety of laboratories interested in participating in the shared effort (Cann et al. 2002). These collection efforts were heterogeneous. Specimens were eventually deposited and archived at the Center for the Study of Human Polymorphism (CEPH), in Paris, which provides samples of

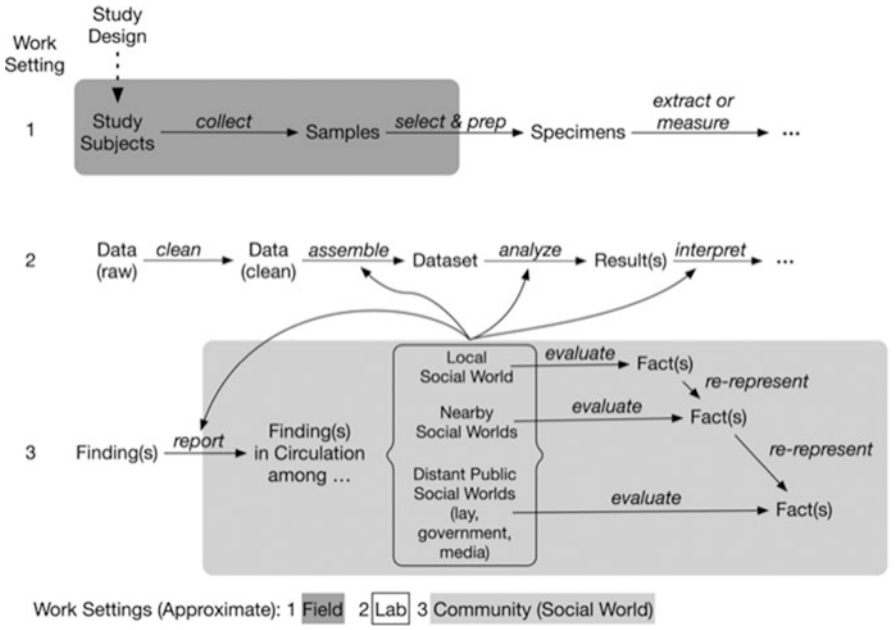


Fig. 1 Diagram illustrating the kind of work flow from a study design, to field work (*stage 1*) producing specimens or raw data, which is then assembled into datasets, analyzed and interpreted as yielding findings in the lab (stage 2), that are then circulated via talks, publications and online media in various social worlds (stage 3) that evaluate findings, elevating some of them to the status of facts and returning others for reconsideration, reinterpretation, and reevaluation. Many points in such processes feed into future study designs or the modifications of ongoing studies

extracted DNA to qualified researchers. These users of HGDP-CEPH specimens then generated data by sequencing the DNA (or in some cases RNA) or by arranging for third parties to do the sequencing.[1] Attention to data in the HGDP, like data in the Human Genome Project (HGP) more broadly, reflects emerging sensibilities of data-centric biology. DNA sequences—"digital" data derived from DNA samples—are the main form of data used to reconstruct ancestry in population genomics. Over the course of the 1990s and 2000s, this data—the data *points*—became increasingly archived in online databases of the kinds Leonelli (2016) describes.

That said, the kind of data *journey* of the sequence data in the HGDP data *sets*, is quite different, in mode of travel, in the organization and standardization of data practices, and in the institutionalization of the data packaging practices that govern the work. It is a data *set* journey—of datapoints between datasets and datasets within and among projects—as much or more than a journey of datapoints into and out of a centralized database.

[1] E.g. the Mammalian Genotyping Service of the Marshfield Clinic Research Institute (Marshfield Clinic Research Institute 2014).

One of the big findings reported concerns relationships between clusters of similar genetic sequence markers and continent-scale geographic distribution of humans. The finding is big enough to be reported in the abstract of the paper (Rosenberg et al. 2002, 2381):

> … without using prior information about the origins of individuals, we identified six main genetic clusters, five of which correspond to major geographic regions, and subclusters that often correspond to individual populations.

The 2002 publication was a landmark and its findings, methods, and conceptual presuppositions widely debated (Horton 2003). The model-based clustering algorithm implemented in analytical software the authors and their collaborators built, program STRUCTURE (Pritchard et al. 2000), assumes a pre-defined number of clusters and then allocates datapoints to clusters based on patterns of genetic similarity. The methodology is to allocate sample individuals to clusters by similarity across a collection of loci—sequences that are either shared or not shared between individual samples. The particular clusters to which individuals are assigned emerge in the clustering process and can then be compared to the "pre-defined" population labels from which the samples came.[2] The "big fact" of continental geographic patterns of human ancestral groups in circulation since the eighteenth century was affirmed by Rosenberg et al. (2002) in a novel way: based on genotype sequence distributions without reference to the pre-defined population labels. The paper is easy to read, contra the authors' intentions, as endorsing a presupposed biological concept of race by conflating a geographic interpretation of genetic classification with race on the grounds that the pre-defined populations (either from the sampling design or in the analysis) somehow biased the results.[3] The analysis is subtle and interpretation tricky.

The analysis leading to the big finding was also contested for its theoretical presuppositions (e.g. by Serre and Pääbo 2004) and defended (e.g. Rosenberg et al. 2005). Some challenges to the results questioned the sampling methodology that produced the HGDP-CEPH cell lines. Others challenged that the analysis was flawed mainly due to theoretical presuppositions regarding whether human genetic variation can be assumed to be organized in more or less discrete "clusters," perhaps with some admixture, or rather in more or less continuous "clines," perhaps with some clumping and isolation. There has been discussion of the analytical methodology as well, including examination of the models and algorithms used by STRUCTURE, alternative cluster algorithms, and alternative multivariate statistical approaches (see Sect. 4 below).

[2] Part of the methodological controversy about this research concerns the sampling methods used to collect samples in the first place and part with whether and how DNA donors "self-identify" with population labels assigned as "meta-data" to the DNA sequence data. Our larger project will address the latter topic in detail (Griesemer and Barragán 2019).

[3] See Wills 2017 for an analysis of "rhetorical appropriations" of the article; see Wade 2014 for a journalist's reading of the paper as supporting a concept of race as "clusters of variation."

It is not my purpose to characterize how well this big "fact" of continental differences (variously as a story of race, ethnicity, or genetic variation) has traveled through the centuries or spread among disciplines or societies, nor to assess the critical charges by post-colonialist thinkers, even while I fully agree that issues of race and ethnicity are far more important in the grand schemes of human cultures and societies than is reconstruction of the data journeys of the datapoints, their uptake in datasets, or interpretations of narratives of facts related to the journeys of the constructed datasets. Nevertheless, my interest *here* is to understand scientific practices involved in using the kinds of data that fuel the work of producing big findings, rather than the findings themselves.

2 Scientific Data Structures

In contrast to the big findings—the stuff of "results" and "discussion" sections of published scientific papers—key small findings mentioned or referred to in Rosenberg et al. (2002) concern the genotypes at the particular loci of the particular sample subjects used to assemble the genome diversity dataset for the analysis. These small findings are, in effect, "asserted" by reference, via the computer files in which the data are represented and recorded, to "scientific data structures." These data structures are displayed in the files and described in "materials and methods" sections, figures, tables, information supplementary to main publications, and software manuals. The data structures and files link sample subject identifiers to sequence data, e.g. diversitydata.stru, which is described in another file, diversityreadme.txt.[4]

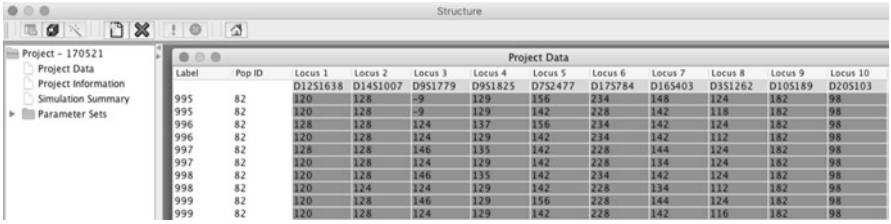
The representation of genotypes in the diversitydata.stru file is clear but indirect, involving pointers (labels) to sequence data records stored in centralized databases such as GenBank. GenBank labels for DNA sequences appear as names of loci in the data file.[5]

Genotypes for each sample individual are coded in labels for the two alleles at each locus represented in the file: 377 loci in this dataset × 2 alleles for each diploid sample individual, with two rows in the data table for each sample subject, one row for each of the paired chromosomes. The allele at the first sequenced locus for sample individual 995, for example, from "Karitiana Brazil AMERICA," (Pop ID 82) is an allele coded as "120" (Fig. 2).

Allele encodings report "genotypes (measured in base pairs)" (Rosenberg et al. 2002), that is, by integer labels: "Each allele at a given locus should be coded by a unique integer" (Pritchard et al. 2010, p. 6). "120" encodes a unique allele at locus

[4] Rosenberg maintains downloadable copies of the exact data used in the original paper at the Rosenberg Lab website (Rosenberg Lab 2018).

[5] Another downloadable file, diversityreadme.txt, contains "meta-data" information about how diversitydata.stru is organized. The reference to "the structure program" is to the software, called "STRUCTURE," authored by some of the authors of Rosenberg et al. (2002).



Label	Pop ID	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7	Locus 8	Locus 9	Locus 10
		D12S1638	D14S1007	D9S1779	D9S1825	D7S2477	D17S784	D16S403	D3S1262	D10S189	D20S103
995	82	120	128	-9	129	156	234	148	124	182	98
995	82	120	128	-9	129	142	228	142	118	182	98
996	82	128	128	124	137	156	234	142	124	182	98
996	82	120	128	124	129	142	234	142	112	182	98
997	82	128	128	146	135	142	228	144	124	182	98
997	82	120	128	124	129	142	228	134	124	182	98
998	82	120	128	146	135	142	234	142	124	182	98
998	82	120	124	124	129	142	228	134	112	182	98
999	82	120	128	146	129	156	228	144	124	182	98
999	82	120	128	124	129	142	228	142	116	182	98

Fig. 2 Screen shot of records in a dataset visualization in program STRUCTURE, after I cleaned (pruned) out meta-data from the file downloaded from the Rosenberg Lab's dataset web page, so the software could read the data file. STRUCTURE is a free software package described by Prichard et al. (2000) and downloadable at http://web.stanford.edu/group/pritchardlab/structure.html. The dataset used by Rosenberg et al. (2002) is downloadable from the Rosenberg Lab "Data sets" webpage: https://rosenberglab.stanford.edu/datasets.html

D12S1638. Sample subject 995 happens to have the same allele, "120," on both chromosomes and is thus homozygous for that locus.

A different data file, diversityloci.txt, associates GenBank sequence identifiers such as D12S1638 with Marshfield Screening Set labels (AFMB002VD5) linking the sequence to the tissue sample from which it was sequenced. This link represents and visualizes an early part of an "omics"-like datapoint journey from samples to sequences in the workflow of population genomicists. In turn, the GenBank identifier points to a record in NCBI's Nucleotide Database, "a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB," (NCBI 2019). reflecting a datapoint journey from a HGP reference sequence contributed to this centralized, online-accessible database. The GenBank sequence label D12S1638 is itself a reference to an actual sequence of 233 nucleotides, reported at the NCBI web site.[6]

These several files, maintained at the Rosenberg lab website as "the data" (and meta-data), correspond to a simple relational data structure that points in one direction to the tissue sample sources in the Marshfield Screening set of CEPH-curated cell lines and points in the other direction to the DNA sequences generated from those cell lines that are eventually encoded in datasets in the Rosenberg lab (and potentially uploadable to GenBank's NCBI Nucleotide Sequence Database).

For the text-based cluster analysis methods implemented in program STRUCTURE, which are used to analyze the dataset in Rosenberg et al. (2002), and for the project of studying allele polymorphisms in these sequences, all that matters is that the text used to label the sequences, e.g. "120," be unique.[7] Whether the

[6] The complete reference sequence for locus D12S1638 can be retrieved from a NCBI Nucleotide Sequence Database Fasta search report https://www.ncbi.nlm.nih.gov/nuccore/Z53031.1?report=fasta. Accessed 5 June, 2018.

[7] The mathematical method at the heart of the software's algorithm, latent Dirichlet allocation, is also used for topic modeling in digital humanities (see Blei and Lafferty 2009). There are journeys of models and software within and among fields to be tracked alongside the data journeys described here.

software actually compares sequence "data" or rather encodings of genotypic differences in text labels for these similarities or differences is irrelevant to the form of analysis and findings presented in the publication, although quite relevant to how we might interpret their datapoint journeys and what other uses or "re-situations" might be made of the datapoints and datasets.

The software, program STRUCTURE, is also downloadable from the laboratory of Jonathan Pritchard, one of its authors, now at Stanford University (Pritchard Lab 2019). The downloadability of the data *set*, which visualizes a scientific data structure, and analytical software from a *local* but accessible website, i.e. a lab web site rather than a community- or government-maintained online database, is a feature of the kind of dataset-centric practice I suggest is now widespread in contemporary biology. This dataset archiving practice occupies a middle ground between the non- or poorly-circulating datasets of hypothesis-centric traditional practices and the highly accessible datapoints archived in centralized databases of the datapoint-centric sciences. It is notable that while web links for this kind of local hosting of datasets and software tend to break as researchers move from one research organization (typically, a university) to another, links to the datasets, software, and references do mostly get reestablished and are relatively speaking "findable" (by internet search) if not by archiving in stable, centrally located internet resources of a federal government (e.g., NCBI, CEPH) or major NGO (e.g., Coriell, Marshfield, Simons).

3 Dataset Journey Representations: Two Visualizations

Datapoint and dataset structure representations for the Rosenberg et al. (2002) paper were already introduced in Fig. 2. What I am *not* talking about is the widely noted and discussed figures in Rosenberg et al. (2002, Figures 1 and 2) and other publications using program STRUCTURE (and in its early versions, the separate visualization software, DISTRUCT). These are visualizations of the *output* of the dataset analysis which are interpreted to produce "big findings."

The description of this dataset in the supplemental information to the paper already narrates a dataset journey by relating the dataset constructed and analyzed for the publication from its source material in DNA extracted from one of the Marshfield screening sets of tissue samples used as sources of DNA. I describe that narrative in the next section. Here, I describe two data visualizations that are central to dataset journey narratives.

Figure 2 displayed a fragment of the Rosenberg et al. (2002) dataset in the form it takes when the dataset file is opened with the Apple MacOS graphical interface implementation of program STRUCTURE, version 2.3.4, after I did some "cleaning" or "pruning" of the "raw" data file. There was a data journey even from the "raw-raw" data—that is, the downloadable data file as archived on the Rosenberg

lab's dataset web page.[8] The "raw-raw" data file contains redundant "meta-data," i.e. data that is not used by program STRUCTURE for data analysis, but which makes the data file more human-readable without following cross-references to other data files, as described above. This meta-data about "pre-defined" populations embedded in the dataset is also used to interpret what genotype similarity clusters *mean* so as to formulate big findings.

Indeed, this meta-data added to the data file is redundant because it is also linked by a data field in each data record to the "population code," e.g. "82" standing for "Karitiana Brazil AMERICA," which also appears in a separate "meta-data" file called diversitycodes.txt.[9] This meta-data must be removed from the data file in order for STRUCTURE to read it.

So far, I have considered datapoint and dataset representations in data tables (stored in computer data files). I turn now to visualized representations of datapoint and dataset *journeys*. These journey visualizations are not narratives themselves, i.e. stories of the travels of points and sets through and to various sets, publications and research projects. Rather, visualizations of scientific data structure representations can *facilitate* data journeys as "chronicles" promoting certain sorts of dataset "travel narratives" in a research community. These visualizations "chart the territory" or "map the waters" in which dataset "ships" can travel from research project to research project.

Thus far, I have mentioned the journeys of samples to specimens to datapoints in dataset assembly, visualized by the kinds of data files discussed above. Next, I describe two kinds of visualizations of data *set* journeys linking different datasets into sequences or chronologies.

3.1 Example: Lab Web Page Dataset Journey Visualization

Rosenberg's lab "diversity" web page links to a "Data sets" web page with a link titled: "HGDP-CEPH human genome diversity cell line panel" (Rosenberg Lab 2018). The main "Data sets" page shows that the Rosenberg lab maintains data sets mostly on humans, but includes non-humans (chickens) and also links to datasets "hosted by collaborating labs."[10]

This diversity web page provides links to many of the maintained datasets for human data. It *also* visualizes a kind of data journey itself. The web page does this as a structured framework of boxes/panels—a vertical, textual "triptych"—in the

[8] I discovered the raw data file was not in a format program STRUCTURE could process directly by trial and error, as have many other naïve users. For evidence, see the Google Groups FAQ: https://groups.google.com/forum/#!forum/structure-software. Accessed 13 August, 2019.

[9] Additional figures can be viewed in an expanded version of this chapter at: http://philsci-archive.pitt.edu. For diversitycodes.txt see Rosenberg Lab (2018).

[10] Chicken breeds with known population structure are used to test "the utility of genetic cluster analysis in ascertaining population structure," see Rosenberg et al. 2001.

web page. Each panel includes a descriptive title, summary dataset description, references to sources, and links to downloadable dataset files. The panels start with the HGDP 2002 dataset from Rosenberg et al. (2002) at the bottom of the page (reading up to the top of the page to follow the journey chronologically) or start with the most recently archived dataset of exome data from 2013 (reading down the page from top to bottom to retrace the lineage of current work back to source datasets). The triptych is headed (at the top) by a summary of the "lineage" of datasets from 2013 back to 2002: "[2013] [2011] [2009] [2008] [2006] [2005] [2002]."

Each panel title indicates the character of the dataset as a modification from HGDP 2002, e.g. "HGDP+other 2013 microsatellites", indicating that 645 autosomal microsatellite loci were added to the original 377 of the HGDP 2002 study in the study published by Pemberton et al. (2013). The web page overall visualizes the journey of the HGDP 2002 datapoints in the 2002 dataset in summary form as each new dataset (or version) is assembled from previous ones, sometimes noting variation from other, related or similar datasets referenced in the literature.[11]

3.2 Example: Excel Spreadsheet Dataset Journey Visualization

In 2006, Rosenberg published a paper attempting to frame the story of a dataset journey in terms of a different kind of visualization than the vertical triptych in his Lab's "Data sets" webpage. Interestingly, because this was also a project concerning the HGDP 2002 dataset, the 2006 project also appears as a place in the dataset journey in that triptych visualization, titled "HGDP 2006 relatives" (Rosenberg Lab 2018).

Rosenberg (2006) seeks to put some order into the proliferation of datasets serving human population genomics ancestry reconstructions by offering a naming convention for datasets and an assessment of which of the datasets that his lab assembled are appropriate for what kinds of work, based on their characteristics *as* datasets.

Rosenberg's dataset visualization is in the form of an Excel Spreadsheet (Fig. 3) that offers a different kind of triptych than the one previously discussed.

The spreadsheet lists individual HGDP sample donors by sample number (e.g., sample donor 995 discussed above). The population codes and "meta-data" of population names, sample locations (usually nation-states) and large scale regions follow. Meta-data information on the sex of the donor is also included. Then, a series of columns are used to indicate whether each donor's sample (in the form of DNA sequence datapoints) is included in datasets that figured in the research projects marked by publications cited in the column headings.

Wherever a "1" appears in the rows of these columns, the individual's DNA sequence data is included among the records of the dataset used in that column's publication. By scanning across the columns from left to right, one can see when a

[11] See additional figures in the expanded version of this chapter at: http://philsci-archive.pitt.edu

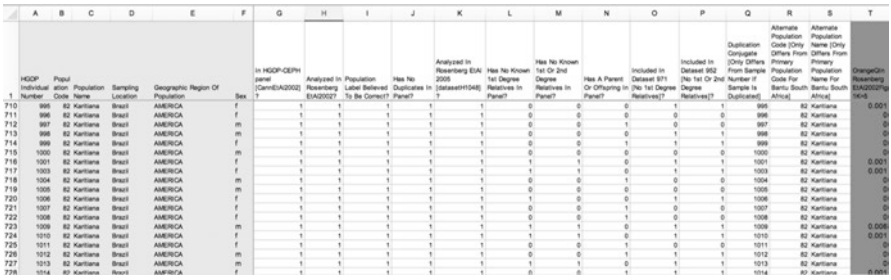


Fig. 3 Screen shot of a fragment of the Rosenberg (2006) spreadsheet "SampleInformation.xls". The figure displays a "triptych" or rather 10-ptych (columns G-P) of points of embarkment/disembarkment of datapoints originating in the HGDP-CEPH LCL cell line panel and ending in dataset H952, which has dropped all data (and records) that include close (1st or 2nd degree) relatives. The spreadsheet is downloadable from Rosenberg Lab (2018). It is not included as supplemental information to the published paper. https://rosenberglab.stanford.edu/data/rosenberg2006ahg/SampleInformation.xls. Accessed 26 August 2019

particular datapoint embarked or disembarked the research program (sequence of research projects) in the Rosenberg Lab. The stops along the journey are from the HGDP-CEPH sample set, to the dataset analyzed in Rosenberg et al. (2002) to the dataset analyzed in Rosenberg et al. (2005), to the dataset called H971 to the dataset called H952.

4 Data Journey Narratives: Datapoints and Datasets

A data journey narrative appears in a particular research publication to tell the story of the dataset that arrived at the research project reported in the publication and is analyzed *there*. Such narratives have the form of stories about "how the dataset got to its destination," after a perhaps circuitous route through other research projects, labs, programs, or specialties.

Dataset journey narratives support a form of narrative explanation (Currie 2018). However, because they are narratives of *dataset* journeys, the target of explanation is not some phenomenon in nature, but rather an explanation of the use of a particular dataset in a particular research project.

The aim is to explain how and why a particular dataset "arrived" at this particular destination, given a particular research project. Dataset journey narratives are needed to persuade an audience to accept the dataset as appropriate for data analysis and thus to accept the results as findings worthy of circulation.

4.1 Dataset Assembly Narrative

Rosenberg et al. (2002) describe a dataset derived from 1056 individuals from 52 "pre-defined" populations, sequenced at 377 autosomal microsatellite loci. The 1056 individual DNA samples are a different set than the samples delivered to the lab from CEPH because not all of those samples could be used for Rosenberg et al.'s purposes. As they write (Rosenberg et al. 2002 supplemental, 1):

> The data set that we analyzed differs from the HGDP-CEPH Human Genome Diversity Cell Line Panel of 1064 individuals in its inclusion of Japanese individual #1026, whose cell line could not be produced owing to technical problems, and its exclusions of She #1331, who was not genotyped, and 8 individuals whose populations had samples of size 1 or 2 (#993, #994, #1028, #1030, #1031, #1033, #1034, #1035). Individual #1410, who is not included in the Cell Line Panel, was genotyped, but as the only representative of his population, was not analyzed. The loci studied, from Marshfield Screening Set #10 (http://research.marshfieldclinic.org/ genetics/sets/combo.html), include a mixture of 377 polymorphic di-, tri-, and tetra-nucleotide repeat loci spread across all 22 autosomes (2, 19), with 3.8% missing data. Genotyping was performed by the Mammalian Genotyping Service (19).

This kind of attention to precisely what dataset is being assembled for a particular investigation is central to the kind of data journey of interest here. Consideration is given to why individual datapoints may or may not embark on the journey. The goal is to use as much of the HGDP-CEPH world-wide sample tissue collection as possible to reflect as much of the world-wide genetic diversity sampled and to provide the most robust inferences of ancestry relations possible, given the available data and background knowledge at the time.

Datasets assembled for specific projects seek to answer questions or test hypotheses. In the case of Rosenberg et al. (2002), the question is whether STRUCTURE can reveal population diversity through study of genetic diversity data without appeal to "self-identified" population membership of sample donors. The datapoints and dataset are described, their provenance and relations to previously assembled datasets are also described, and the reasoning behind the beginnings and endings of journeys *of particular datapoints* (or *specimens*, in the early stages of these data journeys) is given.

The reasons the data journey takes particular twists and turns are a mix of kinds, starting from the usual kinds of "cleaning" of "raw" data familiar from other contexts and discussed above. "Japanese individual #1026" was included in the Rosenberg study even though the extracted DNA was not derived from the CEPH cell line diversity panel due to technical problems with the CEPH cell line. Other tissue samples were not sequenced and hence could not supply data. Samples that were included in the Rosenberg study collectively have 3.8% missing data, i.e. sequences missing for particular loci within the 377 loci sequenced for each individual. Missing data reduces the resolution and precision of the analysis, but not so much that the whole data record for those individuals must be excluded from the analysis. Some data, in other words, fails to be generated from specimens while other data is dropped when the records in which they are coded are eliminated from consideration for various reasons. These are typical kinds of "missing data."

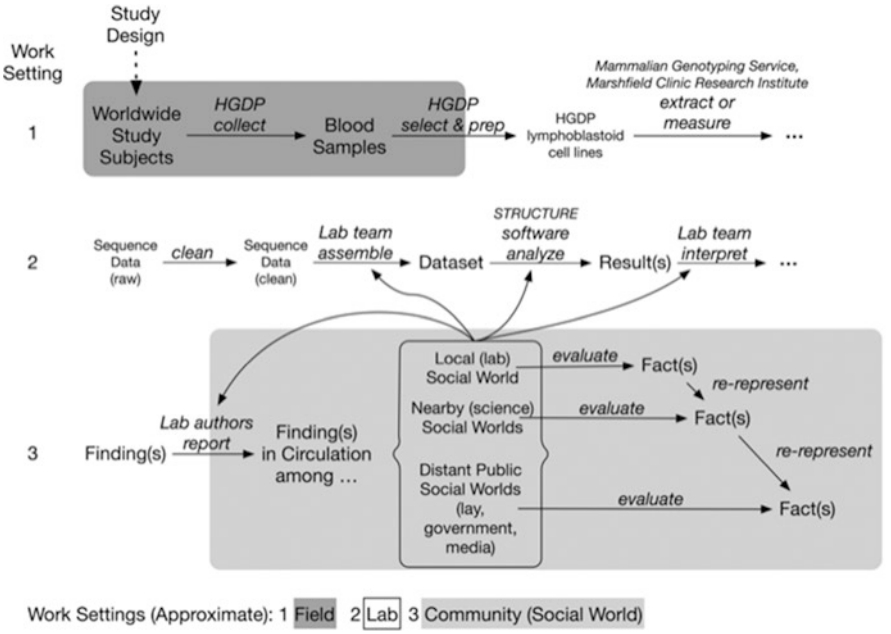


Fig. 4 Workflow diagram following the format of Fig. 1, illustrating specific elements of the dataset assembly and use of data in the Rosenberg et al. 2002 study

Of more interest is when researchers drop DNA sequences in the transition from specimens to data because samples don't meet *theoretical* requirements of their "model-driven" analysis tools. Population genetics theory (and statistical sampling theory) says inferences will be poor for populations represented by only one or two specimens (i.e. sample size n = 1 or 2), so they are not included in the dataset, although they are included in the HGDP-CEPH donor blood tissue specimens, lymphoblastoid cell lines, and DNA sample "screening sets."[12] This kind of hiatus or end to a datapoint and sub-dataset journey is the tip of an iceberg of ways in which data may be "cleaned" or "pruned" in the processing steps leading from material samples to "raw raw" data to "raw" data to "cooked" or processed data.[13] Figure 4 illustrates a workflow for dataset assembly in the work of the Rosenberg Lab following the outline of Fig. 1.

[12] Different investigators and labs set different local sample size thresholds based on varying theoretical requirements for their specific research purposes, so whether a given datapoint can continue on a dataset journey depends on the lab and the project.

[13] The cleaning metaphor supports a useful contrast between "raw" and "cooked" data, even if Bowker (2005, p. 184) is right that "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."

4.2 Dataset Journey Narrative

Of still more interest are beginnings and endings of the journeys of data *points* that result from further analyses inspired by working with the data *set*. These further practices support stories of data journeys of datapoints from dataset to dataset and journeys of datasets from research project to research project. They are *dataset* journeys: a voyage of the *Beagle* rather than Darwin's voyage or FitzRoy's voyage; voyages of the starship *Enterprise* rather than Kirk's voyage or Spock's voyage.

Samples are gathered together; information is collected from samples and assembled into a dataset; the data journey begins with a scientific study of the dataset. Small and big findings arise and emerge from this traditional kind of scientific work. In addition, medium-sized facts arise about the dataset itself, where a medium-sized fact is a relational fact over the group of datapoints, or a fact derived from the set, but not extending or applying beyond the sample specimens that led to the group of datapoints. Medium-sized facts contrast with Leonelli's (2016) small facts or findings corresponding to individual datapoints and with big facts or findings derived from the analysis of the whole dataset in the light of a theory, question or hypothesis.

Because of the technical nature of the work of comparing genetic sequences, results of model-driven analysis in hypothesis-centric research often reveal salient features *of the dataset*, e.g. features that identify particular datapoints or small groups of datapoints as exceptional.[14] These are "medium-sized" facts or findings about the dataset itself, and thus about the sample set or sample sub-sets. These medium-sized facts can drive dataset journeys less visible than the big fact journeys in which scientists *use* data and whose reports grab the headlines when the science is perceived to have important scientific implications, societal impact or is otherwise controversial.

One of these less visible data journeys concerns individual 995 from the Karitiana in Brazil. The challenge in her journey was due to her traveling companion, individual 996. Individual/datapoint 995 from the Karitiana remained on the dataset journey from 2002 to 2006 at least, but when it was inferred that individual 996 was probably 995's sister (due to the level of genetic similarity), one of them had to get off the ship (dataset). Rosenberg (2006) introduced the convention to drop whichever among pairs of such datapoints had arbitrarily been given the higher-numbered label, so Ms. 996's journey ended while Ms. 995's continued. In other cases, whole families had to exit the journey for analogous reasons. This is not how the data journeys would go if socio-cultural anthropologists rather than geneticists were arranging the journeys, given the fundamentally different orientation of the two disciplines to family-level data. For anthropologists, families represent important units in the organization of cultures, but in the context of population-level genomics, they

[14] Compare Tempini, this volume a, b, on analogous discoveries of middle-sized facts about environmental public health datasets, and Hoeppe, this volume, on discovery of "artifacts" in radio telescope datasets.

represent complications to sampling assumptions needed to apply theory to data and thus are to be avoided.

The character of the journeys of the datapoints in the Rosenberg et al. (2002) dataset does not become apparent until one looks at some of the destinations to which the *dataset* traveled. Here, I focus more on dataset journeys *within* the practices of the Rosenberg Lab and its collaborations and less on data journeys out into the wider specialty and beyond where others can download Rosenberg et al.'s data and software and try to repeat the analysis reported in the publication or construct new datasets from old. My goal in this chapter is modest: to formulate the idea of dataset-centric biology, display some of its narrative forms and visualizations, and underscore its potential value for understanding the organization of contemporary sciences, using an illustrative case, not to establish its generality or reach.

In 2005, Rosenberg et al. (2005) published a defense of their methods and findings in the 2002 paper. They "expanded their earlier dataset" from "377 to 993 markers" so they could evaluate critical responses (e.g. Serre and Pääbo 2004) that human populations are ordered in clines, not clusters. Since this was mostly an expansion, with new datapoints joining the journey, including datapoints of kinds other than microsatellite data, I will not further discuss this paper. I note simply that in 2005 a bunch of new travelers joined on, so we can think of datasets as both structures serving as vehicles for the travel of datapoints and as destinations: datapoints travel from dataset to dataset, getting on or getting off different ships at various "stops."

A different paper, by Ramachandran et al. in 2005, is more interesting for present purposes. Certain features of some of the datapoints in the 2002 study were noted, causing some of them to be dropped and others to be added for this study. The account of the dataset structure in the "Materials and Methods" section (p. 15942) is instructive. In this quotation, note that reference (11) is to Rosenberg et al. (2002).

> **Data.** The data set that we analyzed consists of 1,027 individuals from the HGDP-CEPH Human Genome Diversity Cell Line Panel (10). Several individuals from the collection of 1,056 individuals studied by Rosenberg et al. (11) were excluded from the present analysis. These included the following: (i) no. 1026, who was studied by Rosenberg et al. (11) but who was not in the HGDP-CEPH panel; (ii) nos. 770 and 980, who were identified by Rosenberg et al. (11) as likely labeling errors; (iii) nos. 589, 652, 659, 826, 979, 981, 1022, 1025, 1087, 1092, 1154, and 1235, each of whom was identified by Mountain and Ramakrishnan (12) as a duplicate sample of another individual included in the panel; (iv) nos. 111 and 220, who were identified by Mountain and Ramakrishnan (12) as duplicates of each other but whose population labels differed; and (v) 21 individuals from the Surui population, an extreme outlier in a variety of previous analyses (11, 13, 14). Individuals not studied by Rosenberg et al. (11) but analyzed here included the following: (i) no. 1331, whose genotypes had been unavailable at the time of the Rosenberg et al. (11) study; (ii) nos. 993, 994, 1028, 1030, 1031, 1033, 1034, and 1035, who were previously excluded as members of populations with small sample sizes but who were grouped for the present analysis into Southwestern Bantu (individuals no. 1028, 1031, and 1035) and Southeastern Bantu (individuals no. 993, 994, 1030, 1033, and 1034) populations. Thus, the present data set includes two additional populations along with all populations studied by Rosenberg et al. (11) except Surui for a total of 53 populations.

In addition to the kinds of data "cleaning" mentioned previously, this paper dropped a whole population, the 21 individuals sampled from the Surui in Brazil, who live near the Karitiana by the way, as an "extreme outlier." 21 individual data points were dropped from the journey because of a characteristic of that population as a whole—bad traveling companions one might say. This points to the dataset as itself a "fact" or finding produced by the analyses cited. I describe such facts as "medium" sized because they form the basis for the analyses leading to big facts, but are facts about the datasets themselves, analogous to the way the small facts of interest here are facts about individual sample subjects.

Equally interesting is the continuation on the dataset journey of datapoints 1028, 1031, 1034 and 993, 994, 1030, 1033, and 1034 who didn't make the earlier segment of the journey from HGDP-CEPH sample set to the dataset of Rosenberg et al. (2002), but who were allowed to get back into the research program and the overall dataset journey at a different research project and publication "stop" due to the small sample size threshold set by Rosenberg's project. Ramachandran et al. regrouped them into Southwestern and Southeastern Bantu, in effect defining new populations by means of a statistical procedure and adding population labels ("metadata") in the lab rather than as a result of "self-reporting" or "data collection" in the field. In effect, they were interpreted as coming from different places than their original "relevance labels" (place of origin) designated, so they in effect, got new "visas" to travel by Ramachandran et al. (see Leonelli 2011 and 2016 on relevance and reliability labels).[15]

These and other papers appearing between 2002 and 2005 prompted Rosenberg to publish the 2006 paper described above (Sect. 4.2). It visualizes datapoint journeys to and among datasets in a spreadsheet format. Although this paper can be read as part of the other visualization of dataset journeys in the Rosenberg lab (on the datasets web page), this paper can alternatively be read as a new kind of publication in this specialty: a data "curator" paper, signaling a kind of work analogous to that of the specialized data curators in the bio-ontology projects Leonelli (2016) discusses. Instead of tracking changes to datasets within the "materials & methods" or "supplementary" sections of publications of a research project, Rosenberg (2006) is a publication aimed at tracking datasets and, more importantly, proposing standards for naming and using these datasets. This implies a new level of attention to the

[15] M'charek 2005 writes about the "passports" DNA samples needed to pass from one part of the forensics lab she studied to another. I use the related metaphor of "visa." The difference of metaphors is that the passport is a license to travel. The visa is a license to travel in a specific place for a specific period of time. To continue the metaphor, DNA sequences or their tissue samples get "passports" when they are enrolled as samples in the CEPH bio-repository. To get a visa to be included in a particular dataset, the "receiving" country—research group in this case—has to approve. Approval can turn on questions of "desirability" (un-sequence-able tissue samples are undesirable; duplicates are undesirable) or for "theoretical" reasons (sample size too small). Barragán, on the other hand, writes about dataset curating practices in terms of data noise and data silencing as life scientists confront genomic datasets with archaeological, ethnographic, ethnohistorical and linguistic datasets about pre-Columbian and contemporary indigenous groups in northern South America (Barragán 2016, 2017).

ways in which data visualizations (and narratives) set data in motion and contribute to data travel among research projects.

The curation of HGDP-derived datasets in Rosenberg (2006) is not for the sake of online database management and curation of sequence *datapoints*, accessible in the way the "omics" databases are. Rather, it attempts to curate, by documenting in a publication, both the dataset that was initially assembled for the 2002 study *and the journeys* of the datapoints among datasets as a widening circle of researchers used and tinkered with the 2002 dataset to produce new datasets. Differently put, researchers such as Rosenberg (and perhaps those involved in the HGDP more broadly) seem to be taking a new and active interest in conceptualizing and representing the "middle-ground" dataset landscape in which many of their data-centric practices are enacted.

5 A Model of Dataset Journeys and Conclusions

I don't pretend to have done more than scratch the surface of a case study of dataset-centric human population genomics. What I hope to have illustrated is that there is a "middle ground" data landscape between the traditional hypothesis-driven use of data as familiarly described by philosophies of "scientific method" and the new ground of data-centric science described so well by Leonelli. I have gestured at ways in which individual datapoints in datasets, at least in human population genomic diversity studies, make data journeys that are of neither of Leonelli's two kinds, but which resemble them in some respects and to some degree and differ in other respects. Perhaps other question-driven scientific specialties are also influenced by what is newly afforded in the rapidly changing landscape of computational and online digital methods, so there may be many forms of dataset-centric scientific practices waiting to be described. Morgan's study (this volume) of two kinds of data journeys in economics regarding national income accounts and indicator series also concern humans and population data, though with a very different subject matter and principles for dataset formation and use than the biological genomics studies considered here.

In this chapter, I have characterized data journeys in terms of a model comprised of three kinds of components: data structures, data visualizations and data journey narratives. The details of specific scientific practices involved in producing and using these components do matter, if we are to understand these data journeys in middle-ground landscapes of datasets and how they might inform big findings and facts. This is particularly true of genomic ancestry projects like HGDP and biomedical projects like personalized genomic medicine. A further result of this case study is important for present purposes to signal a connection of dataset-centric biology to characteristic features of emerging data-centric "omics" research practices: the emergence of a "bioinformatics" practice alongside the basic, craft research process of asking and answering questions, posing and testing hypotheses.

A distinct and notable line of investigation emerged in population genomics in roughly the time frame 2002–2006 around detection of close relationships among individuals with sequence data in genetic datasets of this kind, both for ancestry and biomedical studies (e.g. Boehnke and Cox 1997; Epstein et al. 2000). This literature, reviewing both datasets and software and modeling approaches, flourished to the point that there are now review articles "benchmarking" different relatedness inference methods (e.g. Porras-Hurtado et al. 2013; Ramstetter et al. 2017). This is evidence of a "standards" specialization emerging within dataset-centric population genomics analogous to the kind of "infrastructure" supporting a bioinformatics specialization that Leonelli (2011, 2016) discusses for data-centric "omics" biology (see also Tempini 2017, this volume a, b).

Moreover, Rosenberg's efforts in (2006) are, I suggest, aimed at supporting a narrative that *steers* the dataset *journeys* of particular datapoints. This is not quite like the curation that goes on in the world of "omics," because the target is *datasets* that are purpose-built and question-driven. The corresponding findings reported in this emerging dataset curation literature are medium-sized, regarding these datasets themselves. The normative directions derive from the standards concerning what sorts of findings or "big" facts can or should be derived from datasets of particular kinds or with particular characteristics.[16]

The data journey discussed here is not quite like the ones Leonelli describes, nor like many of those detailed in Howlett and Morgan (2011) on traveling facts. The journey of the *dataset* is driven in part by the conventional publication system in which peer-reviewed publications of findings using these datasets (together with ancillary visualizations in web pages, spreadsheets and supplementary material) draw attention to the datasets themselves and provoke scrutiny of the datapoints. This scrutiny may extend, moreover, to science studies analysts tracing dataset and datapoint journeys in terms of the components of a model in which data structures, data visualizations and data journey narratives mobilize datapoints in dataset journeys. These journeys may encourage re-use of the dataset or construction of related or alternative datasets, adding and dropping datapoints, thus driving the data journey(s) forward. A different story will be needed for the drivers of "sample sets" such as blood donor samples, cell lines, and extracted DNA sample sets because the differences in materiality matter. The contingency of such sample sets being *available* to feed the production of datasets is critical to dataset journeys.[17]

Dataset journeys, classification schemes and data visualizations designed to maintain and manage them in contemporary biology are driven by a hybrid system of formal, institutionalized, community-sanctioned publishing and quasi-"samizdat" or "self-publishing" systems of personal, individual, laboratory, and university-sponsored websites for distributing datasets and software as well as publications. Unsurprisingly, there is also an emerging effort to institutionalize these kinds of

[16] On the links between data, classification systems and standards, see Bowker and Star (1999).

[17] It remains to be seen whether the model described here applies to sample journeys as well as to data journeys. Thanks to Carlos Andrés Barragán for emphasizing this point.

publication as well, in data journals and dataset archiving services. There is, nevertheless, less standardization of data formats in *dataset* curation and publication as displayed in this case study, even if there is substantial standardization of some of the data content of datapoints due to the rise of data-centric biology and centralized, shared databases for datapoints.[18]

The lower degree of standardization is no doubt partly due to the fact that nearly every population geneticist running a lab today is (or is becoming) a coder who writes their own software in their own way, typically built to read and analyze data formatted anachronistically for their own lab's purposes. It is a relatively manageable problem for others to gain access to such data and tools: if the software and the dataset can be downloaded and the provenance and versioning meta-data for the software is curated along with the dataset, one can (with effort) get the original software to analyze the original dataset. Nevertheless, it *is* a problem. And it entails different kinds of practices and workflows than biological research had required before the data and software coding revolutions of the last few decades.[19]

It means that data journeys may require *software journeys:* particular software versions (and perhaps operating systems or whole virtual machine execution environments) may have to chaperone datasets in order for scientific analyses to be repeated and re-evaluated. Indeed, software versioning is a form of software journey in this middle-ground landscape between the small landscapes of datapoints and small facts on the one hand, and the big landscapes of research findings and big facts on the other.[20]

One more comparison of *dataset*-centric biology with the bioinformatics dimensions of *datapoint*-centric biology will display some similarities and highlight differences. Rosenberg also engages in dataset packaging practices which parallel Leonelli's (2011, 2016) labeling story. Relevance labels, which signal the value of datapoints for particular kinds of journeys and analyses, are included in the dataset (or linked to it) by coding what are called "pre-defined" populations as part of the data records. These are names like *Karitiana*, for the name of the people/place of a certain culturally specific, geographically localized group of people; like *Brazil*, for the name of the nation-state in which the Karitiana are (largely) thought to reside at present; and like *AMERICA*, for the name of the "region" or "continent" of which the relevant nation-state is considered part (see Barragán 2016). As we saw, these "pre-defined" populations played no role in the cluster based *inference* of ancestry

[18] See Tempini, this volume a, b, for a case where infrastructures are built to systematize, institutionalize and standardize the sourcing, hosting, manipulation and generation of datasets. See also Tempini (2017). Morgan's two cases (this volume)—national income accounts and UN indicators of national "health"—also suggest different subject matters and principles may require or lead to different respects and degrees of both standards and infrastructure.

[19] A recent trend in bioinformatics is to solve this problem by making the entire "execution environment" of a whole computational "scientific workflow" the basic unit to be prepared for data journeys. Rather than just data, or software or both, this workflow-centric biology involves creating whole execution environments of data, software and computer operating system as the "basic units" (Meng and Thain 2017).

[20] Thanks to Jason Oakes for pressing this point.

relations in Rosenberg et al. (2002) directly, though they surely did play a role in attracting the attention of those who conducted the initial *sampling* effort because the collectors were interested in sampling human genetic diversity, especially among small groups that might soon disappear. It is no accident that the HGDP-CEPH samples are not (all) drawn from nation-state capital cities, for example, nor from a conventional grid of equally spaced sample locations defined by the geometry of the Earth (constrained by availability of time, money, skill, and interest of collectors in sampling at a particular geographic "scale"). The HGDP-CEPH sample panel was made after several years of inconclusive internal battle over what would be an appropriate sampling protocol for the HGDP (see NAS 1997, for example), but it is not the focus of interest and concern here.

Leonelli's "reliability" labeling practices are also included in Rosenberg's dataset curation practices, though the latter do not appear in "evidence codes" stored in an online accessible "bio-ontology" or "database." Rather, they appear in the "Materials and Methods" sections of "ordinary" scientific papers or coded in archived, downloadable "data" (i.e. meta-data) files devoted to answering a research question or testing a model-driven hypothesis. Cross-referencing a DNA sequence dataset via joining ID field, "Pop ID," is perhaps assurance of both reliability and readability of the data file.

It is common to describe the sources and methods used to generate a dataset in any scientific paper worthy of the name. In the case of human population genomics diversity papers, this extends to discussion of individual datapoints and, increasingly, to a methods literature of papers like Rosenberg (2006) devoted to curation of datasets apart from the research papers devoted to reporting the "big"-fact findings of question-driven research projects. Interestingly, unlike the methods sections of ordinary "omics" papers from molecular biology labs, precious little, if any, space in the Materials and Methods sections is devoted to reporting on the protocols and technologies used to actually generate the sequence data. This may seem surprising, but the data curation tasks for these dataset-centric research programs are less concerned with reporting on *sequence* data reliability than on sequence *dataset* reliability for the question at hand.[21]

In the illustrative case of dataset-centric research discussed here, there are two aspects of the case that may require recalibrating the concept for use beyond my case study of a human population genomics data journey. First, the research is in the population sciences. Population sciences by their nature deal with collections of "individuals" (members of populations). There is a sense of compositionality of the relevant data that is integral to this kind of research. The very idea of a population is that it be composed of members (or parts, depending on one's metaphysics). Surely attention in such contexts is focused on datasets since *collections* of datapoints tend to be used to represent data about populations, e.g. through statistical reasoning that treats the collected data as a sample from a population whose

[21] Studies of ancient human DNA are something of an exception, since the quality of sequence data deriving from ancient, even fossil, specimens is a special problem. See e.g. Veeramah and Hammer (2014) for a relatively recent overview of whole genome sequence data.

unknown properties are subjects of theoretical inquiry, or through some other mode of aggregation, extrapolation or inference from information about members to a set or population. Inquiry may even focus on properties of individuals *qua* members of a population, in a form of research known in some fields as "downward causation," whereby properties of the group cause (or determine) properties of the members. So perhaps the notion that the case discussed here illustrates dataset-centric biology may not generalize beyond population sciences.

A second kind of particularity of the case study is the way it focuses on humans. Data in human biology can be difficult to collect for familiar reasons of ethical or legal restraint or constraint, difficulty of access, expense, entanglement with political, social or cultural differences between researchers, sponsors and potential "subjects," and for many other reasons (Barragán 2012). The constraints may be quite different than for social science data collection about humans (e.g. Morgan, this volume). Biological datasets collected from human subjects thus tend to be more "precious" to researchers than data collected from non-humans (though not always of course—natural history is often pursued in out of the way places that can be hard, expensive, or unpleasant to get to and work in). Human genome diversity data on members of the Karitiana in South America, for example, are critical for the story of human diversity in ways that make these people much more than mere "sample subjects" (see Barragán 2016).

The virtues of "model organisms" include features that tend to make data collection easy, cheap, and fast, and the data, in consequence, relatively disposable. As the unit cost of DNA sequencing falls with advances in technology, on top of scaling and standardizing effects of commercialization, researchers may find it easier to collect new fruit fly specimens, extract new DNA samples, and generate new collections of sequence data for their project-specific uses, than to rely on data already generated by other labs (that may have used doubtful or out-of-date methods, or with questionable expertise, or based on samples less specifically suited to a different project's questions and purposes).

I conclude by noting that the case study analysis and model of datapoint and dataset journeys sketched here indicates not only that new modes of data-centric science are emerging, but that old ones are transforming—particularly around the packaging, vehicles, conveyances, and infrastructure that gets organized or reorganized to put research subjects, specimen samples, extracted materials, and data *points* and *sets* into motion on new kinds of journeys to new kinds of destinations.

Acknowledgments I thank Sabina Leonelli and Niccolò Tempini for inviting me to contribute to their project on "Varieties of Data Journeys" and to Mary Morgan in particular for an interesting collaborative writing experience which helped shape the way I think about data points and data sets. I thank Sabina and Mary for encouraging my larger collaborative project on the re-situation of scientific knowledge with Elihu Gerson, Jason Oakes, Carlos Andrés Barragán, and Alok Shrivastava. I thank Rasmus Winther for prompting my initial interest in this topic. I thank Mary

Morgan for conversations that led to the paper and the conference audience in Exeter for comments. Carlos Andrés Barragán and Sabina Leonelli made many very helpful comments, clarifications and corrections that greatly improved the manuscript. Members of Bill Bechtel's lab group and Bob Westman at UC San Diego provided helpful feedback at a critical stage. Members of my joint lab group with Roberta Millstein, especially Roberta, Michael Hunter, and Tiernan Armstrong-Ingram, provided important input and discussion. I thank Elihu Gerson for feedback, encouragement and general support. I thank the people of California and NSF Grant SES 1849307 for financial support.

References

Barragán, C.A. 2012. Molecular Vignettes of the Colombian Nation: The Place(s) of Race and Ethnicity in Networks of Biocapital. In *Racial Identities, Genetic Ancestry and Health in South America*, ed. Sahra Gibbon, Ricardo Ventura Santos, and Mónica Sans, 41–68. New York: Palgrave Macmillan.

———. 2016. *Lineages Within Genomes: Situating Human Genetics Research and Contentious Bio-Identities In Northern South America*. PhD Dissertation, Department of Anthropology/ Science & Technology Studies Program (STS), University of California, Davis.

———. 2017. *Substantiating Genetic and Cultural Continuity: Partial Connections Between Genomic, Archaeological and Linguistic Datasets*. Paper presented at the Annual Meeting of the International Society for the History, Philosophy and Social Studies of Biology (ISHPSSB) and the Associação Brasileira de Filosofia e História da Biologia (ABFHIB). Panel: The Re-situation of Scientific Knowledge, Organized by J. R. Griesemer. July 17, 2018. São Paulo: ISHPSSB/ABFHIB.

Blei, D. and J. Lafferty. 2009. Topic Models.. http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf. Accessed 14 June 2018.

Boehnke, M., and N. Cox. 1997. Accurate Inference of Relationships in Sib-Pair Linkage Studies. *American Journal of Human Genetics* 61: 423–429.

Bowker, G. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, G., and S. Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Cann, H., C. de Toma, L. Cazes, M. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. Ferrara, J. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. Herrera, X. Huang, J. Kidd, K. Kidd, A. Langaney, A. Lin, S. Mehdi, P. Parham, A. Piazza, M. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. Weber, H. Greely, M. Feldman, G. Thomas, J. Dausset, and L. Cavalli-Sforza. 2002. A Human Genome Diversity Cell Line Panel. *Science* 296 (5566): 261–262.

Currie, A. 2018. *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. Cambridge, MA: MIT Press.

Epstein, M., W. Duren, and M. Boehnke. 2000. Improved Inference of Relationship for Pairs of Individuals. *American Journal of Human Genetics* 67: 1219–1231.

Griesemer, J., and C. A. Barragán. 2019. *Standard Grant: A Case Study of How Re-Situation of Scientific Knowledge from Human Population Genomics Works*. NSF grant SES-1849307, 2019–2021.

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.

Horton, R. 2003. Paper of the Year. *Lancet* 362: 2101–2103.
Howlett, P., and M.S. Morgan, eds. 2011. *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*. Cambridge: Cambridge University Press.
Leonelli, S. 2011. Packaging Small Facts for Re-Use: Databases in Model Organism Biology. In *How Well Do Facts Travel?: The Dissemination of Reliable Knowledge*, ed. Peter Howlett and Mary S. Morgan, 325–348. Cambridge: Cambridge University Press.
———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.
M'charek, Amade. 2005. *The Human Genome Diversity Project: An Ethnography of Scientific Practice*. Cambridge: Cambridge University Press.
Marshfield Clinic Research Institute. 2014. http://www.marshfieldresearch.org/about/welcome, http://www.marshfieldresearch.org/irdl/research-support. Accessed 26 Aug 2019.
Meng, H., and D. Thain. 2017. Facilitating the Reproducibility of Scientific Workflows with Execution Environment Specifications. *Procedia Computer Science* 108C: 705–714.
Morgan, M.S. 2014. Resituating Knowledge: Generic Strategies and Case Studies. *Philosophy of Science* 81 (5): 1012–1024.
———. 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.
NCBI (National Center for Biological Information). 2019. https://www.ncbi.nlm.nih.gov/nucleotide/. Accessed 26 Aug 2019.
Pemberton, T., M. DeGiorgio, and N. Rosenberg. 2013. Population Structure in a Comprehensive Data Set on Human Microsatellite Variation. *Genes, Genomes, Genetics* 3: 891–907.
Porras-Hurtado, L., Y. Ruiz, C. Santos, C. Phillips, A. Carracedo, and M. Lareu. 2013. An Overview of STRUCTURE: Applications, Parameter Settings, and Supporting Software. *Frontiers in Genetics* 4: 1–13. https://doi.org/10.3389/fgene.2013.00098.
Pritchard Lab. 2019. *Structure Software*. http://web.stanford.edu/group/pritchardlab/structure.html. Accessed 26 Aug 2019.
Pritchard, J., M. Stephens, and P. Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155: 945–959.
Pritchard, J., X. Wen, and D. Falush. 2010. Documentation for *Structure* Software: Version 2.3. http://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/structure_doc.pdf. Accessed 6 June 2018.
Ramachandran, S.O., C. Deshpande, N. Roseman, M. Feldman Rosenberg, and L. Cavalli- Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS* 102 (44): 15942–15947. www.pnas.org, 10.1073 pnas.0507611102.
Ramstetter, M., T. Dyer, D. Lehman, J. Curran, R. Duggirala, J. Blangero, J. Mezey, and A. Williams. 2017. Benchmarking Relatedness Inference Methods with Genome-wide Data from Thousands of Relatives. *Genetics* 207: 75–82. https://doi.org/10.1534/genetics.117.1122.
Rosenberg, N. 2006. Standardized Subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, Accounting for Atypical and Duplicated Samples and Pairs of Close Relatives. *Annals of Human Genetics* 70: 841–847.
Rosenberg Lab. 2018. https://rosenberglab.stanford.edu/diversity.html. Accessed 26 Aug 2019.
Rosenberg, N., T. Burke, M.W. Feldman, P. Friedlin, M.A.M. Groenen, J. Hillel, A. Mäki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical Evaluation of Genetic Clustering Methods Using Multilocus Genotypes from 20 Chicken Breeds. *Genetics* 159: 699–713.
Rosenberg, N., J. Pritchard, J. Weber, H. Cann, K. Kidd, L. Zhivotovsky, and M. Feldman. 2002. Genetic Structure of Human Populations. *Science* 298 (5602): 2381–2385.
Rosenberg, N., S. Mahajan, S. Ramachandran, C. Zhao, J. Pritchard, and M. Feldman. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genetics* 1 (6): e70. https://doi.org/10.1371/journal.pgen.0010070.

Serre, D., and S. Pääbo. 2004. Evidence for gradients of human genetic diversity within and among continents. *Genome Research* 14: 1670–1685. http://www.genome.org/cgi/doi/10.1101/gr.2529604.

Tempini, Niccolò. 2017. Till Data Do us Part: Understanding Data-Based Value Creation in Data-Intensive Infrastructures. *Information and Organization* 27: 191–2010.

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.

Veeramah, K., and M. Hammer. 2014. The Impact of Whole-Genome Sequencing on the Reconstruction of Human Population History. *Nature Reviews Genetics* 15: 149–162.

Wade, Nicholas. 2014. *A Troublesome Inheritance: Genes, Race and Human History*. New York: Penguin Books.

Wills, Melissa. 2017. Are Clusters Races? A Discussion of the Rhetorical Appropriation of Rosenberg et al.'s "Genetic Structure of Human Populations.". *Philosophy Theory, and Practice in Biology* 9 (12): 1–24.

James Griesemer is a Distinguished Professor and Chair of the Department of Philosophy at the University of California, Davis, and Member of the UC Davis Science and Technology Studies Program, the Center for Science and Innovation Studies, the Cultural Studies Graduate Group, the Population Biology Graduate Group and the Center for Population Biology. He is also Past President of the International Society for History, Philosophy and Social Studies of Biology and a Member of the KLI in Klosterneuburg, Austria. His primary interests are philosophical, historical and social understanding of the biological sciences, especially evolutionary biology, genetics, developmental biology, ecology and systematics. He has written on a wide variety of topics in history, philosophy and social studies of biology, including models and practices in museum-based natural history, laboratory-based ecology, units and levels of inheritance and selection in evolutionary biology and visual representation in embryology and genetics. He is currently writing a book, *Reproduction in the Evolutionary Process*, which develops a theory of reproduction more comprehensive than current philosophical accounts of inheritance, with applications to theoretical problems ranging from the nature and origin of living systems, evolutionary transitions, eco-evo-devo and cultural change.

Part III
Sharing: Data access, Dissemination and Quality Assessment

Sharing Data, Repairing Practices: On the Reflexivity of Astronomical Data Journeys

Götz Hoeppe

Abstract This chapter probes into how scientists' discursive interactions are oriented not only to others' arguments but also toward achieving an agreement on what data are like and how they ought to be used. It does so by attempting a reading of an episode of data re-use from recent astronomy that is mindful of researchers' interactional and discursive work. I focus on the presumed detection, in 2004, of a galaxy at record distance from Earth. The original data became public at the time of publication and were soon re-used and supplemented with new observations by other teams. Data re-using scientists sought to reconstruct the practices used in making the discovery claim, and found them at fault. This allowed them to suggest the repair of data and of data use practices, which were subsequently taken up by the scientists who had claimed the discovery. I argue that this work was enabled by astronomy's discipline-specific architecture for observation, of which objectual, technological and institutional elements provide contexts and resources for achieving the reflexive repair of data and data use practices. These astronomers experience data journeys more as reflexive loopings in screen-mediated work than as itineraries across physical sites or geographies.

1 Introduction

As Sabina Leonelli notices in her introduction to this volume, Bruno Latour's notion of immutable mobiles – 'objects which have the properties of being mobile but also immutable, presentable, readable and combinable with one another' (Latour 1986, 7) – has been a useful starting point for making sense of data journeys in the sciences. In this contribution I take Latour's notion as a point of departure for probing into how digital data become 'tools for communication' (Leonelli 2016, 69) in astronomical research, oriented not only to the production of specific results but also to the repair or correction of data analysis practices. In doing so I take note of

G. Hoeppe (✉)
Department of Anthropology, University of Waterloo, Waterloo, Canada
e-mail: ghoeppe@uwaterloo.ca

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_9

how data journeys in astronomy are shaped by its disciplinary setting in terms of researchers' shared object of interest (the sky), their use of digital infrastructures and data standards, as well as their largely shared access to telescopes and data. This has pervasive effects on the mobility and uses of data in astronomy. One of these is how it enables practices to be reflexive, that is, how earlier observations and interpretations can be witnessably revised in sequences of action.[1]

Hans-Jörg Rheinberger (1997, 106) has observed that, by making (traces of) transient events durable and available in many places and at various times, immutable mobiles are 'able to retroact on other graphematic articulations – and, what is most important, not only on those from which they have originated.' Drawing on William Ivins (1953) and Elizabeth Eisenstein (1979), Latour (1986, 19–20) can be read as illustrating this retroaction with the impact of printing technology on early modern astronomy, which made it possible for astronomers to notice differences and inconsistencies in data, allowing them to use new observations to re-assess prior ones.

The retroaction that Rheinberger describes is worth a closer look if one seeks to gain insights into contemporary data uses as social and material practices. For one thing, it brings the sequentiality and temporality of scientific work into focus. New data can lead researchers to re-consider prior records. They can spot differences where data were expected to show 'the same,' alerting data users to details of the unavoidably local and contextual production and interpretation of data. In its course, data may be used-as-is, be dismissed, or repaired.[2]

When conceived as machine-generated 'inscriptions' (Latour and Woolgar 1979), digital data may appear to be text-like, a form of writing. The transmission of writing has been commonly regarded as fundamentally distinct from dialogical exchanges in co-presence (Peters 1999). Sybille Krämer expresses this view starkly when she writes that '[t]ransmission is precisely not dialogical: the goal of technical communication is emission or dissemination, not dialogue. We can thus clearly distinguish between the personal principle of understanding and the postal principle of transmission' (Krämer 2015, 23). As conversation analysts have demonstrated, talk-in-interaction (whether in face-to-face situations or mediated through telephones or screen-based media) is shaped by the ongoing repair of utterances: fellow conversationalists routinely resolve the meaning of indexical, context-dependent utterances in the 'here and now' of their interaction, and thus maintain mutual understanding and communicative order concurrently. For example, a speaker may correct an utterance upon noticing her recipient's misunderstanding – a case of self-repair. In doing so participants maintain intersubjectivity (Schegloff 2006).

By contrast, uses of texts appear to be subjected less to the 'tyranny of accountability' (Enfield and Sidnell 2017) characteristic of social interaction in co-presence

[1] I shall elaborate on this ethnomethodological usage of reflexivity later in this text. Always understood as temporal and sequential, it is different, for example, from the postmodern concern of ethnographers about their role in doing fieldwork.

[2] The removal of an artifact and the (re-)construction of missing metadata would be two kinds of repair of scientific data.

(Deppermann 2015). Interpreting texts is less constrained than the interpretation of utterances in conversation, but also more necessary (McHoul 1982; Livingston 1995). But because of this, certain features of texts become more prominent and consequential for assuring the success of communication at a distance, including the resort to numbers (Porter 1995; Heintz 2007).

Some work on writing argues that the schism between transmission and dialogue is not as radical in practice as Krämer and others posit in principle. Thus, Dorothy Smith (2001, 175–176) suggested to conceive of the social, organizational and institutional uses of texts, especially of printed materials, as

> text-reader conversations in which, unlike real-life conversations, one side of the conversation is fixed and unresponsive to the other's responses. (…) However the reader takes it up, the text remains as a constant point of reference against which any particular interpretation can be checked. It is the constancy of the text that provides for the standardization effect. (…) Text-reader conversations are embedded in and organize local settings of work. (…) In standardizing one 'party' to every text-reader conversation, the terms of all conversations with the 'same' text are standardized. Among participants, an open-ended chain is created: text-reader-reader-reader-.

Much like Latour (1986), Smith explores the consequences of the spread of 'identical copies' to multiple sites, yet she focuses on the institutional, regulatory and always again locally situated uses of texts. If digital media technologies provide new possibilities for communication, one may wonder if, in scientists' work with digital data, the schism of transmission and dialogue is likewise challenged.

Building on studies of social interaction and Alfred Schütz's (1967) phenomenology of the social world, Charles Goodwin illustrates how social actors perform 'co-operative, accumulative action on materials provided by predecessors who are not present' (Goodwin 2018, 248). He argues that this pertains characteristically to scientific data production (Goodwin 2013, 8). Witnessing the training of an astronomy PhD student I observed that the work of combining data from different telescopes is not only sequential, temporal, and contextual, but also reflexive (Hoeppe 2014). That is, past actions and interpretations were commonly re-assessed, and repaired as this unfolding work was oriented to the (re-)construction of natural order. For example, when the output of an algorithm for parameter estimation was assessed and deemed implausible (yielding galaxies that were 'too bright for their distance'), calibration exposures were re-inspected, resulting in the identification of an artifact of straylight that was subsequently subtracted to yield better calibrated 'science images' on which the algorithm was re-run. Involving such instances of repair this work bears a resemblance with repair in talk-in-interaction and correction in instructional settings as it has been studied by ethnomethodologists and conversation analysts (Macbeth 2004; Schegloff 2006).[3] It also resonates with studies

[3] Ethnomethodology is a sociological approach to the study of human sense-making practices rooted in phenomenology. Following Garfinkel (1967), it inquires into how people achieve mutual understanding and social order through practices that are inevitably embodied, witnessable, temporal and sequential. See Lynch (1993: 15–17) for a refined account of ethnomethodological reflexivity.

that have expanded and elaborated this notion of repair to address the maintainance of infrastructures and socio-material orders (Henke 2000; Graham and Thrift 2007; Schaffer 2011; Sims and Henke 2012).

My aim in this chapter is to make the notions of repair and reflexivity fruitful for the study of data journeys in the natural sciences. I do so by attempting a reading of an episode of data re-use from recent astronomy. I focus on the presumed detection, in 2004, of a galaxy at record distance from Earth. The original data became public at the time of publication and were soon re-used and supplemented with new observations by other teams. I inquire into how data re-using scientists sought to reconstruct the practices used in making the discovery claim, and found them at fault. Doing so allowed them not only to suggest the repair of data (such as removing artifacts) but also the repair of data use practices, which were subsequently taken up by the scientists who had claimed the discovery. I shall argue that this work was enabled by astronomy's discipline-specific 'architecture for observation,' of which objectual, technological and institutional elements provide contexts and resources for achieving the reflexive repair of data and data use practices. Before describing and interpreting this episode (in Sects. 3 and 4) I sketch the architecture of astronomical observation in which it unfolded (Sect. 2).

While I draw mainly on published sources, the episode I describe happened when I worked as an editor and staff-writer of the popular astronomy magazine *Sterne und Weltraum*. I wrote two pieces about it (Hoeppe 2004, 2005). This magazine's editorial offices are located at the Max Planck Institute for Astronomy in Heidelberg (Germany), a leading research institute, where I benefitted from witnessing rumour about the claimed discovery and assessments of it. This chapter is also informed by my subsequent 18 months of ethnography on digital astronomical research practices, conducted between 2007 and 2010, followed by re-visits between 2010 and 2017, as well as by my own graduate training in astrophysics.

2 An Architecture for Observation: Enabling Reflexive Uses of Data

Seeking to gain insights into data journeys in contemporary astronomy as a social and material practice, I first identify three recurrent disciplinary aspects that come to matter therein: It is marked by astronomers' shared practices of observing and re-observing objects in the sky (a), by their data being almost exclusively digital and available in a standard format (b), and by the shared access to many observing facilities and much observational data (c). The first of these – an object or environment, of sorts – is specific to astronomy (although reference to shared environments or objects is common in other disciplines as well). The other two – a set of technologies and social institutions – are shared to a certain degree with other scientific disciplines.

Together these aspects contribute essentially to what I shall call the architecture of contemporary astronomical observation. It is a relatively stable, and partly

institutionalized, configuration that is shared by diverse users throughout various projects. Today encompassing all branches of astronomy, this architecture has been shaped by the use of satellite observatories and radio telescopes (Hoeppe, in preparation). Data need not be digital or public to be able to travel, nor does the sky have to be fixed for this to succeed, but in contemporary astronomy the three aspects – (a), (b) and (c) – are central to researchers' experience.[4] Here I prefer 'architecture' to the notion of 'knowledge infrastructure' (Edwards 2010; Borgman 2015; Hoeppe 2019a) for drawing attention to the discipline-specific, situated and material setting of observational astronomy and its pervasive effects on the mobility and uses of data.

My use of 'architecture' is informed primarily by Michael Lynch (1993) and Charles Goodwin (2010). Drawing on work by Gurwitsch, Merleau-Ponty and Foucault, Lynch (1993, 132) inquired into how acts of observation are shaped and constrained by disciplinary 'archi-textural environments' that comprise buildings, laboratory set-ups and other equipment. Goodwin (2010, 107) conceives of an 'architecture for perception' as 'a physical object that embodies a solution to a repetitive cognitive task posed in the work of the community using it.' My use of 'architecture' resonates more loosely, but still pertinently, with Knorr-Cetina's (2003) notion, informed in turn by Fligstein (2001), of the reflexive architecture of financial markets, wherein traders engage (and co-constitute) a shared object (a financial market) through mediating digital technologies.

2.1 Object: 'Astronomy is About Observing and Re-Observing Sources on the Sky'

In a blog post, New York University astronomer David W. Hogg (2008) noted in passing that '[a]ll of astronomy and astrophysics is built on the observation and reobservation of sources on the sky.' Doing so is contingent on the stability or 'immutability' of the sky that has been a commonplace for astronomers since Antiquity (Evans 1998). While some objects are known to move in respect to this apparently stable background, most celestial objects can be found again by reference to patterns of stars or celestial coordinates. These are dominant organizing principles for accessing observational data.

Whereas some ancient Greek philosophers famously imagined the astronomical sky to be a material sphere surrounding all observers on Earth (Aristotle 1939), contemporary astronomers tend to define it as 'a two-dimensional distribution of intensity of electromagnetic radiation' (Léna 1989, 245). But it only becomes a 'two-dimensional distribution' when thus represented using media like paper, photographs or digital technologies. The epistemic benefits of observing and re-observing objects in the sky are contingent on this use of media. In using diverse

[4] These three aspects do not characterize astronomical work exhaustively. Other elements of this architecture would be, for example, the implicit cosmology (Hoeppe 2014) that astronomers share, as well as widely shared tools, including the SExtractor code mentioned below.

media, astronomers' 'mundane reason' is oriented to reflexively producing consistent representations of the 'same' sky despite ever-present noise and artefacts in their data (Hoeppe 2014, 2019b; cf. Pollner 1987). In such work, cartographic reference posits the uniqueness of the world as a methodological maxim (Giere 2006) – an assumption that facilitates robustness reasoning in astronomy (cf. Wimsatt 2012 [1981]; Wylie, chapter "Radiocarbon Dating in Archaeology: Triangulation and Traceability", this volume).

2.2 Technology: Astronomical Data Are Digital, and Utilize a Standard Format

A second aspect of contemporary astronomy's architecture of observation is technological. Unlike the enormous diversity of materials that biologists, oceanographers or archaeologists can use (Leonelli 2016; Halfmann, this volume; Wylie, this volume), almost all data in contemporary astronomy are digital recordings of cosmic radiation. To unpack the specific salience of the digital for the travel of data, it is necessary to refine Latour's (1986) notion of immutable mobiles, which included, among others, hand-drawn maps, machine generated inscriptions and printed tabulations. Rheinberger (2011, 344) suggests that the traces produced in laboratory experiments become 'data proper' (and proper immutable mobiles) only when they can be easily stored and retrieved. In my reading, he appears to be close to suggesting that 'data proper' are symbols. In Peirce's (1992 [1894]) classification of the relation between signs and their objects, traces are indices and represent their object by contiguity. Photographs are indices as well as icons, signs which correlate with their objects by resemblance. Beyond this, digital photographs are also symbols, since – constituted by arrays of numbers, in binary format or otherwise – they use notational conventions. This resonates with an understanding of the digital as the 'encoding' of 'information' that permits its subsequent retrieval without loss (e.g. Dourish 2017, Chapter 1).

Invented in 1969, Charged-Coupled Devices (CCDs) are found in most digital cameras and at all observatories today (Smith and Tatarewicz 1985; McCray 2014). These detectors use the photoelectric effect to produce grid-shaped pixel images which can be read out and then stored, retrieved or transmitted as digital files. Not only are they very sensitive, and – once cooled with liquid nitrogen to reduce quantum noise in the detector – can be exposed for several hours. CCDs also are very linear, recording incoming light in direct proportion to the exposure time. This implies that their outputs are directly amenable to arithmetic calculations, including the pixel-by-pixel addition, subtraction and division of images, with generative uses for epistemic work (Hoeppe 2019b). The linearity of CCDs also allows astronomers to calculate the exposure time necessary for reaching a specific sensitivity. This encourages conceiving of data in terms of the 'abstract time' of exposures and facilitates scheduling observing time – a requirement for the institutionalization of

service mode observing, in which observatory staff members produce data for absent data users (Hoeppe 2018).

In 1979, astronomers defined FITS (Flexible Image Transport System), a shared data format to 'transfer regularly gridded astronomical image data between different locations' (Grosbøl et al. 1988, 359; cf. also McCray 2014). It was quickly adopted and endorsed by all major observatories and space agencies. FITS files are calculable objects which link metadata to images and tables; they have been the dominant data format in astronomy for more than 30 years. The FITS format has shaped astronomers' understanding of what their data are like.[5] Its dominance contrast with the diversity of data formats in disciplines like biology (Leonelli 2016), the Earth sciences (Halfmann, this volume) and economics (Morgan, this volume).

2.3 Social Institutions: Sharing Instruments and Data

The third aspect of astronomy's architecture for observation is institutional.[6] Since the 1960s, a dominant fraction of astronomical data has been produced by public observatories built and operated using tax money. In their process of allocating observing time, peer-review committees at major observatories and space agencies consider proposals from a diverse, international community of academic users. Current practices of observation and data management are deeply informed by how satellite telescopes and radio observatories have been operated since the late the 1970s. These data have been digital throughout. Produced mostly at public institutions, they were made exclusively available to applicant users only for a period of proprietary use (typically 6 or 12 months), after which they became public. The commitment to do so instigated the formation of public data archives. Another defining element of the operation of satellite and radio observatories was the introduction of service-mode observing (Hoeppe 2018). Authors of observing projects can use data earlier, but they do not have preferential access to the local context of data production, including the 'tacit knowledge' of observatory staff members.

3 Re-Using Data to Assess an Astronomical Discovery Claim

Given this background I now consider a discovery claim and its subsequent evaluation, in which the original data, available publicly at the end of a period of proprietary use, were re-used and re-assessed in the light of additional observations.

[5] The dominant status of FITS as astronomy's unique data format has been challenged recently.

[6] Here I adopt Hart's (2001, 136) convenient definition that an 'institution is an established practice in the life of a community or it is the organization that carries it out.'

3.1 Record Distance: "A Lensed Galaxy at z = 10.0"

In 2004, a group of five astronomers led by Roser Pelló of the Observatoire Midi-Pyrénées in Toulouse (France) announced the discovery of a galaxy at record distance from Earth (Pelló et al. 2004a). These researchers had used detectors at three large telescopes to observe clusters of galaxies, which, because of their considerable mass, are thought to act as gravitational lenses which focus the light emitted from faint distant background sources. By utilizing this 'gravitational telescope,' they hoped to exceed the sensitivity of previous searches for the most distant galaxies. What astronomers call redshift (abbreviated as z) is a measure of how much the wavelengths of the light emitted by cosmic objects are stretched due to cosmic expansion, shifting specific spectral features to longer wavelengths. Adopting a specific cosmological model allows computing both the distance and the look-back time, that is, how long this light has traveled to reach observers on Earth. Pelló et al. claimed to have discovered a galaxy at redshift 10.0 behind the galaxy cluster Abell 1835, corresponding to a look-back time of more than 13 billion years. This was a momentous claim, given that spectroscopically confirmed, and thus presumably reliable, record-redshifts had increased more or less steadily from $z = 5.7$ in 1993 to 'only' $z = 6.5$ in 2004, with a few redshift 7 candidates awaiting spectroscopic confirmation (Hu and Cowie 2006).

The Toulouse team relied on two lines of evidence. The first was a series of digital pixel images taken through a series of broad-band filters (each transmitting light of a specific wavelength range) in visible and near-infrared light using the Wide-Field/Planetary Camera (WFPC2) of Hubble Space Telescope (HST), the 3.6-meter Canada-France Hawaii Telescope (CFHT) on Mauna Kea (Hawaii) and, with the Infrared Spectrometer And Array Camera (ISAAC) at one of the European Southern Observatory's (ESO) four 8-meter Very Large Telescopes (VLT) on Paranal (Chile). These data, throughout in FITS format, were obtained in service mode. Pelló et al. first reduced the digital images of Abell 1835, detected objects using SourceExtractor (Bertin and Arnouts 1996), a code widely used in the community, and assembled a catalogue of photometric measurements of the detected sources in the exposures of all the filters used.

As in other attempts to find distant, young galaxies, Pelló et al. then searched for a discontinuity in the observed spectral energy distributions. To qualify as candidate high-redshift galaxies, objects had to be detected at longer (near-infrared) wavebands only, but not at shorter (visible) ones. The 'break' in-between, ascribed to the observed wavelength of the redshifted Lyman α spectral emission line of hydrogen, was expected from previous observations of distant galaxies and simulated model spectra.

Object #1916 in Pelló et al.'s catalogue was the most promising candidate. It was not detected in visible light, but in three near-infrared wavebands, with an apparent 'jump' between the so-called J-band (around 1.26 μm) and the H-band (around 1.65 μm; Fig. 1). This suggested a redshift around 10 to Pelló et al., even though detections in each single detection were only marginally statistically significant.

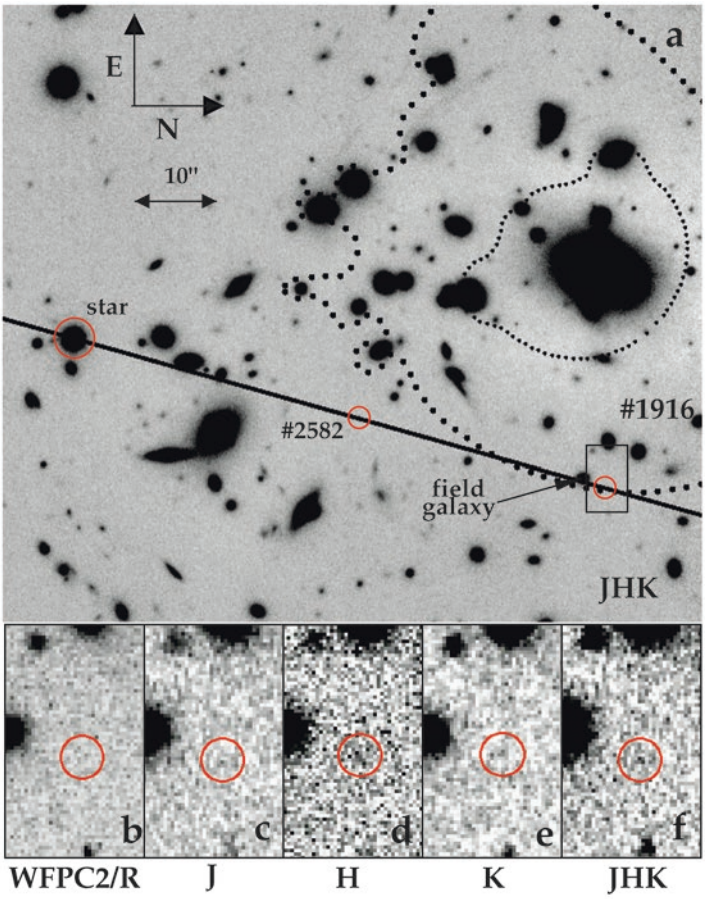


Fig. 1 Figure 1 of Pelló et al. (2004a), showing digital photographic negatives of exposures of the Abell 1835 galaxy cluster using the Infrared Spectrometer And Array Camera (ISAAC) at the Very Large Telescope (VLT, Chile; above) with exposures of the field around the candidate high-redshift galaxy #1916, as taken with the WFPC2 camera on board the Hubble Space Telescope in the visual R band (bottom left) and the near-infrared J-, H-, and K-bands using ISAAC (bottom right). Pelló et al. claim the detection of #1916 in the J-, H- and K-bands. (Reproduced with permission © ESO)

The Toulouse team's second line of evidence was a spectroscopic analysis. They recorded spectra of #1916 in the J-band, also with the ISAAC instrument at the VLT. Long exposures taken in two different observational set-ups suggested to them a statistically significant signal of a spectral line at a wavelength of 1.337 μm. Interpreting it as the redshifted Lyman α emission, they inferred a redshift of 10.0 for #1916. Pelló et al. argued that finding a galaxy at such a high redshift, whose light was emitted only 460 million years after the big bang, was in accordance with theoretical models of galaxy formation and cosmology. On March 1, 2004, ESO

published a press release entitled ‘VLT smashes the record of the farthest galaxy known.’[7] It was widely taken up by popular news media.

3.2 Three Hot Pixels

Pelló et al.’s ISAAC/VLT observations became public through ESO’s data archive website on March 3, 2004, one year after the observations were recorded, and 2 days after publication of the press release. Several scientists retrieved the data for scrutinizing the analysis and for re-assessing the data in light of additional observations. Soon thereafter the Toulouse team’s second line of evidence was challenged. Stephen Weatherley from Imperial College London and colleagues processed the spectroscopic data with an independent approach (Weatherley et al. 2004). After failing to confirm the spectral line, they tried to identify the discrepancy with the analysis of the Toulouse team of Pelló et al. (2004a), which they refer to as P04, by replicating their procedure:

> To find the cause of the discrepancy between our results for the Lyα line and those reported by P04, we re-reduced the data following the principles of P04, i.e. subtracting frames in pairs, then wavelength calibrating the frames, rebinning onto a linear wavelength scale. In this process we made a careful check for bad data. We identified three variable hot pixels[3] [pixels which did not record incoming light linearly and have to be excluded from the analysis] which result in spurious positive flux in four of the sky-subtracted frames in the region of the emission line. We confirmed that these are very easily identified when the frames are registered to the nearest pixel, but are harder to spot when the data are rebinned in the wavelength calibration step. The summed spurious positive flux, when averaged into the entire data set, corresponds approximately to the flux measured by P04; therefore these variable hot pixels plausibly account for the difference between our results and those of P04.
>
> [3] These have coordinates (28, 761), (28, 836), (919, 790) in the raw frames. (Weatherley et al. 2004, L32)

Weatherley et al. recognized that one step in the reduction procedure adopted by Pelló et al. (2004a) – ‘rebinning the data onto a linear wavelength scale’ – had caused them to fail to identify the three hot pixels as artifacts that, in a proper analysis, had to be removed from the data. In other words, Weatherley et al. could replicate the signal reported by Pelló et al. only if making what they thought was a mistaken use of the data. By listing the positions of the hot pixels in the raw frames in a footnote, Weatherley et al. made the Toulouse team accountable in detail to their treatment of the raw data.

[7] http://eso.org/public/news/eso0405/ (accessed 20 April 2018).

3.3 A Transient Source?

It did not take long until the Toulouse team's first line of evidence (an object detected with the photometric properties of a high-redshift galaxy) was challenged as well. Only in combination with the photometry measured through broad-band filters was the high-redshift interpretation of the spectral line plausible. A single spectral line itself would not have provided substantial evidence for any galaxy's redshift, since the spectra of young, intensively star-forming galaxies exhibit several perspicuous spectral lines at widely different wavelengths. Their individual detection would point to different, and generally smaller, redshifts. Pelló et al.'s claim that the spectral line detected of #1916 was the redshifted Lyman α emission of a galaxy critically depended on the detected discontinuity in emission between the near-infrared J and H wavebands.

However, as pointed out by a team led by Malcolm Bremer of the University of Bristol (UK), both of these detections were 'not highly significant' (Bremer et al. 2004, L1). Shortly after the publication of Pelló et al.'s paper, Bremer and his colleagues were granted two blocks of Director's Discretionary Time[8] for using the NIRI (Near Infra-Red Imager) camera at the 8-meter Gemini North telescope on Mauna Kea (Hawaii) to obtain a deeper exposure of #1916 in the H-band. In their resulting publication, Bremer et al. (2004) state that they aim to 'better constrain the H-band photometry (…) and to investigate the morphology of the source under the excellent seeing conditions that are often attainable at Gemini-North' (Bremer et al. 2004, L2). Thus, they write that they are not merely out to replicate Pelló et al.'s claim but seek to refine their interpretation.

Even though Bremer et al.'s (2004) Gemini NIRI observations had been taken under excellent conditions and being significantly deeper, i.e. more sensitive, than the ones taken for Pelló et al. at the VLT, they failed to detect #1916 in the H-band. Their paper is a comprehensive exercise in making sense of this non-confirmation. They did so by first re-reducing Pelló et al.'s H-band data, which they showed side-by-side along with their deeper H-band image (Fig. 2), confirming that their photometric calibration agrees well with that of Pelló et al. Next, Bremer et al. set out to probe whether, with their method and new data, they could have accidentally failed to detect #1916. For doing so they placed artificial objects into their digital exposures and demonstrated that, using their source detection and photometry algorithms, they could retrieve the properties of these objects, illustrating the soundness of their measurements. As such, they called the discontinuity between the J- and H-band fluxes into question, and with it a critical piece of evidence for the redshift of 10.0. Maintaining a cautious and considerate tone throughout, Bremer et al. dis-

[8] Demonstrating that one aims to conduct observations on a 'hot and highly competitive topic' is one legitimate rationale for submitting a proposal for Director's Discretionary Time at the European Southern Observatory. See: https://www.eso.org/sci/observing/policies/ddt_policy.html (accessed 14 September 2017).

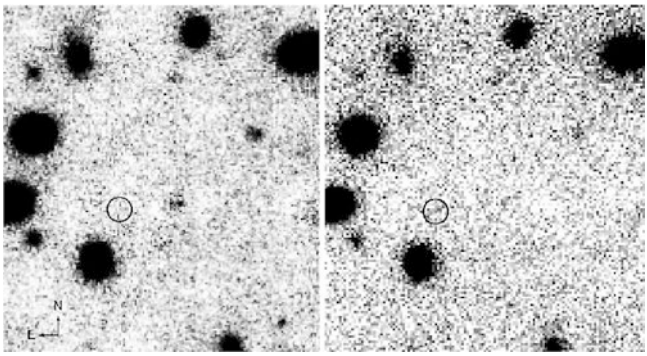


Fig. 2 Figure 1 of Bremer et al. (2004), showing their re-reduction of Pelló et al.'s (2004a, b) H-band image taken with ISAAC at the VLT (right) along with new H-band observations made with the NIRI camera at the Gemini North telescope at Mauna Kea (Hawaii). Bremer et al. emphasize that they have used the same display parameters as Pelló et al. Note that these images are rotated relative to those shown in Fig. 1. (© AAS. Reproduced with permission)

cuss that #1916 may not existent or be intrinsically variable, considering if a transient object in the outer solar system could have been spotted in some exposures. They conclude that 'the reality of any source at this position [of #1916] has to be strongly questioned' (Bremer et al. 2004, L4).

The lack of a detection at visible wavelengths was another piece of Pelló et al.'s evidence for the high redshift of #1916, an argument informed by model spectra energy distributions of young star-forming galaxies. To probe this further, members of Bremer's team, now under the lead of Matt Lehnert of the Max Planck Institute for extraterrestrial Physics in Garching (Germany), succeeded to obtain Director's Discretionary Time at the VLT to obtain additional deep imaging in the (visible) V-band. They wrote: 'A V-band detection would be decisive: it would demonstrate beyond any doubt that the source is *not* at $z = 10$' (Lehnert et al. 2005, 81, emphasis in original). Other than in their previous paper, their objective now appears to challenge Pelló et al.'s discovery claim. Despite going deeper than Pelló et al.'s previous V-band images, which had been taken with the Hubble Space Telescope, and with assessing their detection limit by again placing faint artificial objects into their digital exposure and retrieving them using algorithms, Lehnert et al. fail to detect #1916 in the V-band. They note that, '[f]ormally, a nondetection is consistent with the candidate having a redshift of 10' (Lehnert et al. 2005, 82), and then embarked on a long critical discussion of how a transient source, such as a supernova explosion or an object moving in the outer solar system, could have conspired to produce the signal that Pelló et al. claimed, finding none of these scenarios compelling.

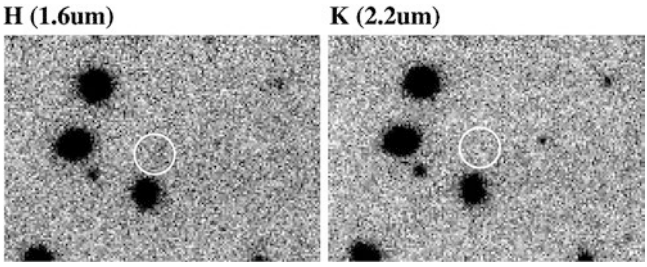


Fig. 3 Figure 1 of Smith et al. (2006), showing re-reductions of Pelló et al.'s (2004a, b) H and K near-infrared images of the field around the position of the high-redshift galaxy candidate #1916. Note that these images are rotated relative to those shown in Fig. 1. Using the same data, Smith et al. fail to replicate Pelló et al.'s H- and K-band detection. (© AAS. Reproduced with permission)

3.4 Lost in the Noise

Yet another group of astronomers combined new observations of #1916 with a re-analysis of Pelló et al.'s ISAAC/VLT data. For an independent study of Abell 1835, Graham P. Smith of the California Institute for Technology and colleagues at the University of Arizona (USA) had been granted spectroscopic observations using LRIS, the Low-Resolution Imaging Spectrograph at the 10-meter Keck telescope on Mauna Kea (Hawaii), and infrared images taken with the Spitzer Space Telescope, a satellite observatory. These researchers were able to modify their observing run with LRIS so as to include the position of #1916, and to search for it in the Spitzer images which had been scheduled prior to Pelló et al.'s discovery announcement. Neither of these observations yielded a detection at the position of #1916. It is noteworthy that Smith had been the principal investigator of the Hubble Space Telescope WFPC2 observations of A1835 that Pelló et al. (2004b) (re-)used.

Smith et al. (2006) then went on to re-analyze Pelló et al.'s H- and K-band data (see Fig. 3). After not detecting #1916 with what they regarded as a proper analysis set-up, they experimented with alternative algorithm settings (smoothing the images, varying the size of the detection area etc.) to find out under which conditions Pelló et al.'s near-infrared images would yield the detection they claimed. Doing so was similar to Weaverley et al.'s (2004) re-analysis of Pelló et al.'s ISAAC/VLT spectra. Smith et al. (2006) wondered how the apparently elongated shape of #1916 (as seen in Fig. 1, center of the bottom panel, and Fig. 2, right image) could be reproduced. They found that only, and inappropriately, searching for objects at an angular scale smaller than the resolution of the exposures would yield the stated detection at the position of #1916. Doing so would make it one of 500 comparably large statistical fluctuations across the field, each of which could have been mistakenly held for a detection. They conclude that 'there is no statistically sound evidence for the existence of #1916' (Smith et al. 2006, 580).

3.5 The Toulouse Team Responds to Its Critics

Progressively faced with these accounts, the Toulouse team first endorsed Bremer et al.'s speculation that #1916 might be variable and announced a more detailed investigation (Péllo et al. 2004b).[9] Two years later they presented a comprehensive analysis of their search for distant galaxies in the fields of the galaxy clusters A1835 and AC114 (Richard et al. 2006). It includes improved photometry of #1916, which they rename as A1835#8, and, in separate online material, a newly estimated redshift: $z = 7.38$, which is much lower than the claim of a record redshift (Richard et al. 2006, *Online Material*, p. 4). Citing Lehnert et al. (2005) and Smith et al. (2006), Richard et al. acknowledge in their main paper that 'the photometric properties of this source are still a matter of debate' and notice that 'its nature (and hence also its redshift) presents a puzzle' (Richard et al. 2006, 873). They drop it from their list of high redshift galaxy candidates without addressing the alternative analyses of Bremer et al. and Smith et al., whose data had meanwhile become public.[10]

All critics of the Toulouse team acknowledged communications with Roser Pelló in their publications (Weatherley et al. 2004, L29, L30; Bremer et al. 2004, L4; Lehnert et al. 2005, 84; Smith et al. 2006, 581). In their 2006 paper, the Toulouse team in turn acknowledges its critics' 'useful comments and discussions', including Graham Smith and his co-author Egichi Egami (Richard et al. 2006, 879). A closer reading of their paper suggests that the Toulouse team's refined data analysis is informed by their critics. The *Online Materials* to their paper are particularly interesting. There they describe improvements in the data reduction and attend carefully to the assessment 'false-positive detections.' Not only did they now probe their completeness statistics with inserting (and algorithmically retreiving) artifical stars into their digital images (Richard et al. 2006, 867), as Bremer et al. (2004) had done (see above). They also argue for a careful analysis of the noise properties of near-infrared images that echoes the comments and recommendations of Smith et al. (2006). These *Online Materials* thus communicate the Toulouse team's adoption of specific sequential operations of work with near-infrared exposures first adopted by secondary data users. As such, members of the Toulouse team repaired (or corrected) its data analysis practices.

On September 27, 2010, the European Southern Observatory added a note to the 2004 press release on its website, stating that the 'identification of this object with

[9] Since the field observed by Pelló et al. (2004a) is located in a position on the sky where models of gravitational lensing in the gravitational field of A1835 predict large magnifications of sources at a wide range of cosmic distances the probability of detecting variable sources is increased.

[10] It is only in a non-peer reviewed venue, ESO's quarterly magazine *The Messenger*, that members of the Toulouse team defended their analysis against the criticism of Bremer et al. (2004), Lehnert et al. (2005) and Smith et al. (2006). Notably, this paper (Schaerer et al. 2006) is co-authored by Egichi Egami, a co-author of Smith et al. (2006). It did not elicit a response in a peer-reviewed publication.

a galaxy at very high redshift is no longer considered to be valid by most astronomers.'[11]

4 Discussion and Conclusions

This discovery claim and its subsequent dismissal is an episode of astronomical data journeys that involved 18 astronomers and data from seven different detectors attached to four large ground-based telescopes (one in Chile, three in Hawaii) and two satellites. These diverse data 'met' in 'cartographic' digital images, as well as in discipline-specific representational spaces: in tables listing measured radiation fluxes as a function of wavelengths, and in their graphical representation as spectral energy distributions (SEDs), typically with model SED shapes overlaid (as in Fig. 1 of Pelló et al. 2004a). Once its proprietary period had ended, Pelló et al.'s (2004a) ISAAC/VLT data were being successively re-analysed in the light of additional observations, and the question turned to what Pelló et al. had done with the data to see what they saw. Given that their observations were done in service mode, Pelló et al. did not have preferential access to the local context of data production at the observatory.

To see (or not to see) #1916 in the reduced images was distinctly shaped by specific equipments and work practices (cf. Lynch 2013). Bremer et al. and Smith et al. present images of their re-reductions of Pello et al.'s VLT/ISAAC data used for the discovery claim alongside reductions of their supplementary data. The critics insist that one has to make specific identifiable and describable mistakes to make #1916 visible as a high-redshift galaxy. Weatherley et al. (2004) claim that the presumed spectral line becomes visible only when three hot pixels are not properly deleted from the data set, and Smith et al. (2006) found that only when parameters are set to values they consider inappropriate did the search algorithm identify #1916 as a proper source. All participants agreed that at least two lines of evidence were necessary to claim the discovery of a high-redshift galaxy, a shared demand for the robustness of evidence (see the chapters by Halfmann, Parker and Wylie).

Pelló et al.'s (2004a) discovery announcement elicited the critical responses and was as such generative of a sequence of actions. The unfolding 'text-reader conversation' (Smith 2001) was marked by a series of comparisons involving re-analyses of Pelló et al.'s VLT/ISAAC 'raw' data (as available on the observatory website) and re-assessments of the initial detection. The results of these re-analyses were made witnessably visible (see Figs. 2 and 3). This conversation was not entirely virtual, with scientists reading each other's papers and working with the original data set in

[11] http://eso.org/public/news/eso0405/ (accessed 20 April 2018).

different ways. As mentioned above, all critics acknowledge communications with Roser Pelló, the lead author of the Toulouse team.[12]

While any description of action is unavoidably incomplete at some level of detail (Livingston 2008, 161), the critics of the Toulouse team point to omissions of descriptive detail in the Pelló et al. (2004a) article that could challenge their replicability. Thus, Weatherley et al. (2004, L31) miss a proper description of Pelló et al.'s bad pixel rejection methods, Bremer et al. (2004, L3) bemoan the unspecified observing time of the H-band exposure, and Smith et al. (2006, 576) note that Pelló et al. 'neither explain how they reduced the [Hubble Space Telescope WFPC2] data nor how the detection limit was calculated.' However, these critics claim to have been able to re-construct what Pelló et al. had done nevertheless (perhaps thanks to Roser Pelló's clarifications; see above) – at least to their own satisfaction and expectation of what they themselves could be held accountable to. In this sense, the open access to data made analysis practices available for inspection by other researchers. This opens the way to a deeper mutual understanding, and possibly agreement, of what proper procedures for using these data are.

As such, this episode can be read as an instance of the repair of data use practices. Members of the Toulouse team ended up learning from secondary users of 'their' data, making their revised understanding witnessable in the *Online Materials* of their Richard et al. (2006) article. It seems, then, that it was through the (separate) circulation of a discovery claim and the 'raw' data on which it was based that practices could travel from data re-users 'back' to those for whom the data were originally recorded. The 'raw' data themselves were not repaired, but remained fixed as the first element of a 'text-reader conversation' (Smith 2001). The work described was reflexive, inasmuch as past actions were re-interpreted in the light of new data and analyses, and made witnessable as such. In terms of its mediated character and its episodic temporality that extended over 2 years, the repair of practices in this episode was markedly different from conversational repair or instructional correction (Macbeth 2004; Schegloff 2006). However, as argued previously for cases of maintaining, or re-establishing the functioning of, motor boats (Sohn-Rethel 1990), buildings (Henke 2000), scientific instruments (Schaffer 2011), infrastructures (Graham and Thrift 2007) and credibility (Sims and Henke 2012), the notion of repair is illuminating in its orientation to social, material and natural orders.

The architecture for observation that I described in Sect. 2 provided resources for the assessment and repair of data and data use practices. First, there are its objectual features. The 'immutability of the heavens' has been instrumental already for assembling the data set that the Toulouse team gathered over a period of 2 years (Pelló et al. 2004a). The use of celestial coordinates for achieving reference was not described as being problematic in this episode. Only in respect to the possibility that Pelló et al. may have detected a transient source were time-variable phenomena,

[12] I restrict my discussion to articles that appeared in peer-reviewed journals. The chronology of events is unavoidably affected different periods of review, re-submission and publication.

such as small objects moving in the outer solar system or supernovae, invoked (by Lehnert et al. and Smith et al.) as interpretive resources.

Secondly, there are its technological and medial features. The importance of the digitality of data is illustrated not only by its apparent mobility (through information infrastructures), which – like the FITS data format – is presumed throughout and not mentioned in the publications cited, but also by the possibilities of analysis afforded by this medium, including Smith et al.'s experimenting with inserting artificial objects into their images and their detailed assessment of the statistical properties of noise in their infrared images. The Toulouse team later adopted these techniques.

Thirdly, this episode was shaped institutionally not only by the open access to Pelló et al.'s VLT/ISAAC data after the proprietary period, which made it possible for others to reconstruct and criticize their actions. With the exception of having access to the data earlier, Pelló et al. used ESO's data archive just like those who later scrutinized, and contested, their discovery claim.

The possibility of re-using data for making sense of what the Toulouse team had done to see what they saw arguably contributed to avoiding a discourse in which a discovery claim was directly confronted with counter-evidence, resulting in its dismissal. As interest turned from the presumed discovery of a specific galaxy at record distance to the viability of the method of using galaxy clusters as 'gravitational telescopes' for such work, the reputation of the Toulouse team was not damaged beyond repair. Indeed, its members have continued to do much respected research in the field.[13] Since their data had been taken by observatory staff in service mode, Pelló et al. could not be blamed for lacking technical skill or manipulative intentions in producing their data. Although Pelló et al. were informally blamed for having issued an overly bold and ultimately mistaken claim, nobody accused them of fraud. Galison (2003) and Leahey (2016) have pointed out that scandals of fraud are rare or even absent in contemporary astronomy, ascribing this mostly to the dearth of commercial interest and the large team sizes in the discipline. Going beyond this claim it seems that if there is a particular ethos of sharing in astronomy, it may well be constituted by the 'tyranny of accountability' (Enfield and Sidnell 2017) of this work with open access data in astronomy's architecture for observation.

Acknowledgements I am very grateful to the participants of the Data Journeys workshop in Exeter for their insightful comments. Sabina Leonelli, Niccoló Tempini, Hans-Jörg Rheinberger, Catelijne Coopmans, David Teira, Roser Pelló and two anonymous reviewers provided very helpful comments on earlier versions which much improved this chapter. I thank Roser Pelló, Graham Smith and Malcolm Bremer for the permission to reproduce their figures. Financial support was provided by the Social Sciences and Humanities Research Council of Canada (Insight Grant #435-2018-1397).

[13] I infer this from the number of citations to the papers of its group members, the allocation of observing time and the funding received as indicated on the website https://obswww.unige.ch/Research/starbursts/ (accessed 13 September 2017). The work of Richard et al. (2006) led to the team being granted additional observing time with the Hubble Space Telescope.

References

Aristotle. 1939. *On the Heavens*. Trans. W.K.C. Guthrie. Loeb Classical Library, vol. 338. Cambridge, MA: Harvard University Press.

Bertin, E., and S. Arnouts. 1996. SExtractor: Software for Source Extraction. *Astronomy and Astrophysics Supplement Series* 117: 393–404.

Borgman, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Bremer, Malcolm, Joseph B. Jensen, M.D. Lehnert, N.M. Förster Schreiber, and Laura Douglas. 2004. Gemini H-Band Imaging of the Field of a z = 10 Candidate. *Astrophysical Journal* 615: L1–L4.

Deppermann, Arnulf. 2015. Retrospection and Understanding in Interaction. In *Temporality in Interaction*, ed. A. Deppermann and S. Günthner, 57–94. Amsterdam: Benjamins.

Dourish, Paul. 2017. *The Stuff of Bits*. Cambridge, MA: MIT Press.

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

Eisenstein, Elizabeth. 1979. *The Printing Press as an Agent of Change*. Cambridge: Cambridge University Press.

Enfield, N.J., and Jack Sidnell. 2017. *The Concept of Action*. Cambridge: Cambridge University Press.

Evans, James. 1998. *The History and Practice of Ancient Astronomy*. Oxford: Oxford University Press.

Fligstein, Neil. 2001. *The Architecture of Markets*. Princeton: Princeton University Press.

Galison, Peter. 2003. The Collective Author. In *Scientific Authorship*, ed. M. Biagioli and P. Galison, 325–355. New York: Routledge.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs: Prentice-Hall.

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

Goodwin, Charles. 2010. Things and their Embodied Environments. In *The Cognitive Life of Things*, ed. Lambros Malafouris and Colin Renfrew, 103–120. Cambridge: MacDonald Institute for Archaeological Research.

———. 2013. The Co-operative, Transformative Organization of Human Action and Knowledge. *Journal of Pragmatics* 46 (1): 8–23.

———. 2018. *Co-Operative Action*. Cambridge: Cambridge University Press.

Graham, Stephen, and Nigel Thrift. 2007. Out of Order: Understanding Repair and Maintenance. *Theory, Culture & Society* 24 (3): 1–25.

Grosbøl, P., R.H. Harten, E.W. Greisen, and D.C. Wells. 1988. Generalized Extensions and Blocking Factors for FITS. *Astronomy and Astrophysics Supplement Series* 73: 359–364.

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.

Hart, Keith. 2001. *Money in an Unequal World*. New York: Texere.

Heintz, Bettina. 2007. Zahlen, Wissen, Objektivität: Wissenschaftssoziologische Perspektiven. In *Zahlenwerk: Kalkulation, Organisation und Gesellschaft*, ed. Andrea Mennicken and Hendrik Vollmer, 65–85. Wiesbaden: Verlag für Sozialwissenschaften.

Henke, Christopher R. 2000. The Mechanics of Workplace Order: Toward a Sociology of Repair. *Berkeley Journal of Sociology* 44 (1): 55–81.

Hoeppe, Götz. 2004. Licht vom Ende der Dunkelzeit. *Sterne und Weltraum, May* 2004: 16–17.

———. 2005. 'Entfernteste' Galaxie verschwunden. *Sterne und Weltraum, March* 2005: 13.

———. 2014. Working Data Together: The Accountability and Reflexivity of Digital Astronomical Practice. *Social Studies of Science* 44 (2): 243–270.

———. 2018. Tensions of Accountability: Scientists, Technicians and the Ethical Life of Data Production in Astronomy. *Science as Culture* 27 (4): 488–512.

———. 2019a. Mediating Environments and Objects as Knowledge Infrastructure. *Computer Supported Cooperative Work* 28 (1–2): 25–59.

———. 2019b. Medium, Calculation, Play: On Digital Images in Scientific Practice. *Social Studies of Science* 49 (5): 758–784.

Hogg, David W. 2008. http://hoggresearch.blogspot.ca/2008/03/budavari-and-szalay.html. Accessed 4 Sept 2017.
Hu, Esther M., and Lennox Cowie. 2006. High-Redshift Galaxy Populations. *Nature* 440: 1145–1150.
Ivins, William M. 1953. *Prints and Visual Communication*. Cambridge, MA: Harvard University Press.
Knorr-Cetina, Karin. 2003. From Pipes to Scopes: The Flow Architecture of Financial Markets. *Distinktion* 7: 7–23.
Krämer, Sybille. 2015. *Medium, Messenger, Transmission*. Trans. Anthony Enns. Amsterdam: Amsterdam University Press.
Latour, Bruno. 1986. Visualization and Cognition: Thinking with Eyes and Hands. *Knowledge and Society* 6 (1): 1–40.
Latour, Bruno and Steve Woolgar. 1979. *Laboratory Life*. Rev. Edn. Princeton: Princeton University Press.
Leahey, Erin. 2016. From Sole Investigator to Team Scientist: Trends in the Practice and Study of Research Collaboration. *Annual Review of Sociology* 42: 81–100.
Lehnert, M.D., N.M. Förster Schreiber, and M.N. Bremer. 2005. Deep Very Large Telescope V-Band Imaging of the Field of a z = 10 Candidate Galaxy: Below the Lyman Limit? *Astrophysical Journal* 624: 80–84.
Léna, Pierre. 1989. Images in Astronomy: An Overview. In *Evolution of Galaxies. Astronomical Observations*, Lecture Notes in Physics, ed. I. Appenzeller, H.J. Habing, and P. Lena, 243–282. Berlin: Springer.
Leonelli, Sabina. 2016. *Data-Centric Biology*. Chicago: University of Chicago Press.
Leonelli, Sabina. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
Livingston, Eric. 1995. *An Anthropology of Reading*. Bloomington: Indiana University Press.
———. 2008. *Ethnographies of Reason*. Aldershot: Ashgate.
Lynch, Michael. 1993. *Scientific Practice and Ordinary Action*. Cambridge: Cambridge University Press.
———. 2013. Seeing Fish. In *Ethnomethodology at Play*, ed. Peter Tolmie and Mark Rouncefield, 89–104. Aldershot: Ashgate.
Macbeth, Douglas. 2004. The Relevance of Repair for Classroom Correction. *Language in Society* 33: 703–706.
McCray, W.Patrick. 2014. How Astronomers Digitized the Sky. *Technology and Culture* 55 (4): 908–944.
McHoul, Alex. 1982. *Telling How Texts Talk: Studies in Reading and Ethnomethodology*. London: Routledge.
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.
Peirce, Charles Sanders. 1992 [1894]. What is a Sign? In *The Essential Peirce: Selected Philosophical Writings. Volume II (1894–1913)*, ed. Peirce Edition Project, 4–10. Bloomington: Indiana University Press.
Pelló, Roser, D. Schaerer, J. Richard, J.-F. LeBorgne, and J.-P. Kneib. 2004a. ISAAC/VLT Observations of a Lensed Galaxy at z = 10.0. *Astronomy and Astrophysics* 416: L35–L40.
———. 2004b. Very-High Redshift Lensed Galaxies. In *Impact of Gravitational Lensing on Cosmology*, IAU Symposium 225, ed. Y. Mellier and G. Meylan, 373–386. Cambridge: Cambridge University Press.
Peters, John Durham. 1999. *Speaking into the Air*. Chicago: University of Chicago Press.
Pollner, Melvin. 1987. *Mundane Reason*. Cambridge: Cambridge University Press.
Porter, Theodore M. 1995. *Trust in Numbers*. Princeton: Princeton University Press.
Rheinberger, Hans-Jörg. 1997. *Toward a History of Epistemic Things*. Stanford: Stanford University Press.
———. 2011. Infra-Experimentality: From Traces to Data, from Data to Facts. *History of Science* 49: 337–348.

Richard, J., R. Pelló, D. Schaerer, J.-F. Le Borgne, and J.-P. Kneib. 2006. Constraining the Population of 6 < z < 10 Star-forming Galaxies with Deep Near-IR Images of Lensing Clusters. *Astronomy and Astrophysics* 456: 861–880.

Schaerer, Daniel, Roser Pelló, Johan Richard, Eiichi Egami, Angela Hempel, Jean Francois Le Borgne, Jean-Paul Kneib, Michael Wise, Frédéric Boone, and Francoise Combes. 2006. Searching for the First Galaxies through Gravitational Lensing. *The Messenger* 125: 20–23.

Schaffer, Simon. 2011. Easily Cracked: Scientific Instruments in States of Disrepair. *Isis* 102 (4): 706–717.

Schegloff, Emanuel A. 2006. Interaction: The Infrastructure for Social Institutions, the Natural Ecological Niche for Language, and the Arena in which Culture is Enacted. In *Roots of Human Sociality: Culture, Cognition and Interaction*, ed. N.J. Enfield and S. Levinson, 70–96. Oxford: Berg.

Schütz, Alfred. 1967. *The Phenomenology of the Social World*. Trans. George Walsh. Evanston: Northwestern University Press.

Sims, Benjamin, and Christopher R. Henke. 2012. Repairing Credibility: Repositioning Nuclear Weapons Knowledge After the Cold War. *Social Studies of Science* 42 (3): 324–347.

Smith, Dorothy E. 2001. Texts and the Ontology of Organizations and Institutions. *Studies in Cultures, Organizations and Societies* 7 (2): 159–198.

Smith, Robert W., and Joseph N. Tatarewicz. 1985. Replacing a Technology: The Large Space Telescope and CCDs. *Proceedings of the IEEE* 73 (7): 1221–1235.

Smith, Graham P., David J. Sand, Eiichi Egami, Daniel Stern, and Peter Eisenhardt. 2006. Optical and Infrared Nondetection of the z = 10 Galaxy Behind Abell 1835. *Astrophysical Journal* 636: 575–581.

Sohn-Rethel, Alfred. 1990. *Das Ideal des Kaputten*. Bremen: Edition Bettina Wassmann.

Weatherley, S.J., S.J. Warren, and T.S.R. Babbedge. 2004. Reanalysis of the Spectrum of the z = 10 Galaxy. *Astronomy and Astrophysics* 428: L29–L32.

Wimsatt, William. 2012. [1981]. Robustness, Reliability and Overdetermination. In *Characterizing the Robustness of Science*, ed. Léna Soler, Emiliano Trizio, Thomas Nickles, and William C. Wimsatt, 61–87. Dordrecht: Springer.

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.

Götz Hoeppe is Associate Professor of Anthropology at the University of Waterloo, Canada. He has written the books *Why the Sky Is Blue: Discovering the Color of Life* (Princeton University Press), which received the 2010 Louis J. Battan Author's Award of the American Meteorological Society, and *Conversations on the Beach: Fishermen's Knowledge, Metaphor and Environmental Change in South India* (Berghahn Books). He works on collaborative uses of digital scientific data, practices of knowledge-making and epistemic cultures in astronomy and the climate sciences.

Evaluating Data Journeys: Climategate, Synthetic Data and the Benchmarking of Methods for Climate Data Processing

Wendy S. Parker

Abstract This chapter concerns the benchmarking of methods used to process data in climate science. It explores the nature and value of benchmarking in this context by examining an ongoing initiative – the International Surface Temperature Initiative (ISTI) – that is developing a public databank of temperature observations as well as a system for benchmarking the methods that databank users employ to further process the data. Interestingly, the benchmarking system will make use of "synthetic data" generated with the help of computer simulation models. It is argued here that the benchmarking system has crucial scientific and gatekeeping roles to play in the context of ISTI. It is further suggested that, once we appreciate how synthetic data are to be produced and used by ISTI, we uncover yet another variety of what Paul Edwards (A vast machine: computer models, climate data, and the politics of global warming. MIT Press, Cambridge, MA, 2010) has described as "model-data symbiosis" in the practice of climate science.

1 Introduction

In November 2009, email exchanges among climate scientists were taken without authorization from servers at the U.K.'s Climatic Research Unit and made public on the Internet. Dubbed "Climategate" in blogs and popular media, the contents of the emails gave rise to allegations of fraud and scientific misconduct on the part of climate scientists and called attention to an ongoing struggle between climate scientists and climate contrarians over data access. Several independent reviews exonerated climate scientists of the charges of fraud and misconduct but did fault them in one significant respect: for being insufficiently open and transparent in their dealings with contrarian requests for information, including Freedom of Information requests for raw data used to estimate changes in global mean surface temperature over land (see e.g. Russell et al. 2010).

W. S. Parker (✉)
Department of Philosophy, Centre for Humanities Engaging Science and Society (CHESS) & Institute for Data Science (IDAS), Durham University, Durham, UK
e-mail: wendy.parker@durham.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_10

The International Surface Temperature Initiative (ISTI) was launched in 2010, in the wake of the Climategate episode, and seeks to promote transparency and openness in the process of producing temperature change estimates (Thorne et al. 2011). Spearheaded by leading climate data experts in the UK and around the world, ISTI is working to construct a comprehensive, publicly-accessible global databank of historical surface meteorological observations taken over land, providing data at monthly, daily and even sub-daily resolutions. This is a substantial undertaking.[1] It involves not only obtaining observational data from numerous sources around the world, but also getting the data and any available metadata into a common format and then merging the data records with the aim of maximizing station coverage and data quality while minimizing duplication. Release of the first version of the merged data, focused on monthly mean temperatures, occurred in June 2014 (Rennie et al. 2014), with an updated release in October 2015. These releases included data from over 30,000 observing stations worldwide, several times the number typically used in estimating global surface temperature changes over land.

In addition, ISTI intends to develop a set of benchmarking tests for users who generate "data products" from the databank (see also Tempini, this volume a, b on "derivative datasets"). These products include reconstructions of the evolution of global and regional temperature over time, from which trends and other changes are often calculated. Arriving at such data products requires the application of quality control and "homogenization" algorithms to data in the databank. *Homogenization* is a process that aims to remove jumps and trends in station time series that are due to non-climatic factors, e.g. because an instrument is replaced with a new one, a building is constructed nearby, or the timing of observations changes. In the envisioned benchmarking tests, users would apply their algorithms to synthetic data that contain deliberately-introduced artefacts (known as "inhomogeneities") that are not known to the users in advance. The idea is to test how well the different homogenization methodologies work by checking their performance on data that are like real climate data in many important respects, but for which the "true" underlying climate signal is known (Willett et al. 2014). ISTI hopes to host all data products developed using the databank on its website, along with information about benchmarking performance for the generating methodologies (Thorne et al. 2011).

This chapter discusses and reflects upon the data journeys envisioned by ISTI, with special attention to the accompanying benchmarking scheme. As outlined further in Sect. 2, these journeys include the traveling of temperature data from a source or holder, through a processing and merging procedure by ISTI, followed by subsequent quality control and homogenization processes undertaken by third parties, which deliver "data products". We will see that, given methodological decisions along the way, only some data will make the full journey. Section 3 turns to ISTI's envisioned benchmarking scheme, explaining how its synthetic data are to be produced with the help of simulation models that serve as analogues to the real

[1] It is also largely unfunded. Progress has been somewhat slower than desired, in part because participating researchers are largely volunteering their time (with in-kind support from some of their institutions).

world. The benchmarking scheme and its synthetic data are, in a sense, "external to" the envisioned data journeys, but it is argued that they are far from ancillary components of the ISTI project. On the contrary, benchmarking has crucial roles to play, not only in advancing the scientific goals of the project but also by serving an important gatekeeping function in the complex and politicized context of climate change research. Section 4 contends that the proposed use of synthetic data in ISTI's benchmarking scheme constitutes a distinctive variety of what Paul Edwards (2010) has called "model-data symbiosis" in the practice of climate science. Finally, Sect. 5 offers some concluding remarks.

2 Data Journeys Envisioned by ISTI

Today, there are thousands of land-based weather stations around the world making regular observations of temperature, pressure, humidity and other weather conditions, often overseen by national meteorological services. It was not always so, of course. Regular observations of temperature began at a few sites in Western Europe in the seventeenth century (Camuffo and Bertolin 2012), but it was not until the mid-nineteenth century that coordinated networks of land-based observing stations began to emerge; they expanded rapidly in the twentieth century (Fleming 1998, Ch. 3). In recent decades, there have been major efforts to locate and bring together records of these past surface observations in support of climate change research (e.g. Menne et al. 2012). These ongoing efforts require international cooperation and involve significant "data rescue" activities, including imaging and digitizing of paper records.

ISTI's envisioned journeys for surface temperature data – from individual records held by sources to data products of use in regional and global climate change research – are conceptualized in terms of six stages (Thorne et al. 2011). Paper records from observing stations, as well as digital images of those records, are what ISTI call "Stage 0" data. Many of the data obtained by ISTI in constructing their databank, however, are Stage 1 data: "digitized data, in their native format, provided by the contributor" (Rennie et al. 2014, 78). In the simplest case, Stage 1 data might have been produced from Stage 0 data by typing into a computer file what is shown on a paper record.[2] In other cases, Stage 1 data already reflect substantial processing by the contributor. For instance, many of the Stage 1 data obtained by ISTI had already been subjected to quality control and homogenization algorithms by their contributors; though "raw" data are preferable for the databank, these are not what some sources are willing or able to provide, whether for practical or proprietary reasons.

[2] That person might have translated or transformed the original data record into a preferred format of her own, so it seems that the "native format" here should be understood as whatever format the contributor to ISTI provides.

At Stage 2, data are converted by ISTI from their native format – units, temporal resolution, etc. – to a common format that also includes some metadata. The conversion to a common format sometimes involves averaging, e.g. in order to convert hourly data to daily or monthly average values. The metadata at Stage 2 indicate not only such things as the station's ID, latitude, longitude and elevation, but also whether the data have undergone quality control or homogenization by the contributor, how a daily or monthly average value was calculated from observations (if this was necessary), and the mode of transmission from contributor to ISTI (ibid., 79). The documentation accompanying the first release of ISTI data indicates that some 58 source collections were converted to Stage 2 data (see Table 1 for a snapshot). Many of these data collections were obtained from national meteorological services, universities and research stations.

At Stage 3, the data sources are prioritized and then subjected to a merge algorithm, with the aim of maximizing station coverage and data quality while minimizing duplication. In the merge performed for monthly data, ISTI chose to give higher priority to sources "that have better data provenance, extensive metadata, come from a national weather or hydrological service, or have long and consistent periods of record" (Rennie et al. 2014, 82). The highest priority source – in ISTI's case the Global Historical Climatology Network – Daily (GCHN-D) dataset, which contains on the order of a billion observational records (Durre et al. 2010) – becomes the starting point for building the merged dataset.

The merge algorithm then works through the remaining data sources according to their priority. Each record provided by a source is a candidate station. The algorithm first compares the record to a list of stations with known issues in their data or metadata; this list was generated using another algorithm that looks for signs of problems, such as an undocumented shift in units, or flipping the sign of the station's longitude, etc. If the record/candidate station is not withheld ("blacklisted") following this comparison, the merge algorithm continues, trying to determine whether the candidate station is unique or matches an existing station. This is a non-trivial task, given that different data sources can use different names for the same station, can represent latitude and longitude with different precision, etc. ISTI describes the merge algorithm as employing a "quasi-probabilistic approach" that "attempts to mimic the decisions an expert analyst would make manually" (Rennie et al. 2014, 81). It involves comparing features of the metadata of station records, and in some cases of the temperature data themselves, and then assigning scores on a set of metrics. Depending on whether those scores pass particular thresholds, the station records are either withheld, added to the dataset as new stations, or merged with records for existing stations (see Fig. 1). The merge algorithm is made available on the ISTI website, and ISTI emphasizes that users can change the threshold settings to produce alternative merged datasets, as ISTI did themselves (see Rennie et al. 2014, Table 12).

In ISTI's analysis, their "databank" project encompasses the journeys of data from Stage 0 to Stage 3. The final two stages of the envisioned journeys are left to users of the databank; since the databank is publicly available, in principle these users might be anyone. At Stage 4, quality control procedures are applied to Stage 3

Table 1 Partial list of sources of temperature data that were converted to Stage 2 data

Name	Source	Time scale	Raw/QC/ homogenized	TMAX	TMIN	TAVG
Antarctica	SCAR Reader Project	Monthly	Raw	N	N	Y
Antarctica (AWS)	Antarctic Meteorological Research Center	Daily	Raw	Y	Y	N
Antarctica (Palmer Station)	Antarctic Meteorological Research Center	Daily	Raw	Y	Y	Y
Antarctica (South Pole Station)	Antarctic Meteorological Research Center	Monthly	Raw	Y	Y	Y
Arctic	IARC/Univ of Alaska Fairbanks	Monthly	Homogenized	N	N	Y
Argentina	National Institute of Agricultural Technology (INTA)	Daily	Raw	Y	Y	N
Australia	Australia Bureau of Meteorology	Daily	Homogenized	Y	Y	Y
Brazil	INPE, Nat. Institute for Space Research	Daily	Raw	Y	Y	N
Brazil-In met	INMET	Daily	Raw	Y	Y	N
Canada	Environment Canada	Monthly	Homogenized	Y	Y	Y
Canada	Environment Canada	Monthly	Raw	Y	Y	Y
Central Asia	NSIDC	Monthly	Homogenized	Y	Y	Y
Channel Islands	States of Jersey Met	Daily	Raw	Y	Y	N
Colonial Era Archives	Griffith	Monthly	Raw	Y	Y	N
CRUTEM4	UKMO	Monthly	Homogenized	N	N	Y
East Africa	Univ. of Alabama Huntsville	Monthly	Raw	Y	Y	Y
Ecuador	Inst. Nacional De Met E Hidrologia	Daily	Raw	Y	Y	N
Europe/N. Africa	European Climate Assessment (Daily, Non-Blended)	Daily	Raw	Y	Y	Y

Source: Rennie et al. (2014, Table 1)

data. It turns out that the GCHN-D data, which form the starting point for constructing the ISTI monthly merged dataset, have already been subjected to quality control by the U.S. National Center for Environmental Information (NCEI).[3] The procedure there involves 19 automated tests designed to detect duplicate data, climatological outliers and spatial, temporal and internal inconsistencies; a small number of problematic data (well under 1%) are consequently excluded (Durre et al. 2010).

[3] This was formerly called the National Climatic Data Center (NCDC).

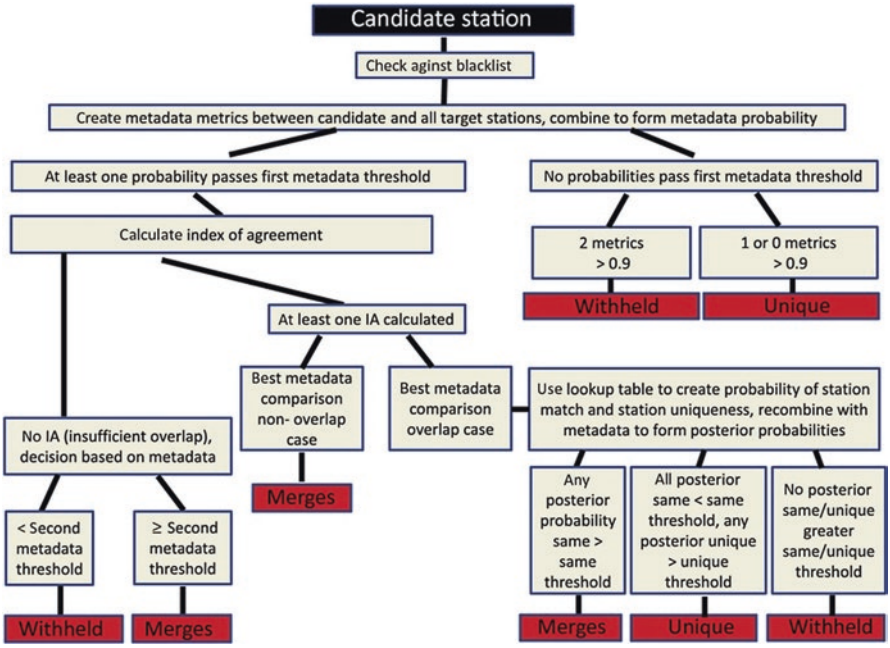


Fig. 1 Workflow for ISTI merge algorithm. (Source: Rennie et al. 2014, Fig. 5)

Many other sources in the ISTI databank, however, have not been subjected to quality control (as their metadata communicates), and it is up to users to address this.

Stage 5 data have, in addition, been homogenized. That is, the data at Stage 5 have been subjected to further processing to try to remove jumps and trends in station time series due to non-climatic factors. When station metadata are available (e.g. reporting a shift in instrument location), this can aid homogenization, but often such metadata are not available. Many homogenization methods thus are statistical methods that compare station records to those of neighbouring stations or of reference stations, identifying and correcting for inhomogeneities based on expected relationships among the records (see e.g. Costa and Soares 2009; Venema et al. 2012). There is substantial uncertainty about how best to identify and correct for inhomogeneities; statistical methods for doing so, for instance, can plausibly employ any of a number of approaches and assumptions. Table 2 summarizes features of several different homogenization algorithms. Even without going into the technical details, one can see that there are differences in what data are compared to (comparison), in how data are searched for potential inhomogeneities (search), and in the form of tests used to identify the presence of inhomogeneities (criterion); there are also differences in how corrections are applied to data once an inhomogeneity has been detected (not shown in Table 2). Attempting to correct for inhomogeneities is particularly important when data will be used to quantify changes in

Table 2 Homogenization algorithms differ in a number of respects

	Comparison		Detection		References
Method	Comparison	Time step	Search	Criterion	
MASH	Multiple references	Annual, parallel monthly	Exhaustive	Statistical test (MLR)	Szentimrey (2007, 2008)
PRODIGE	Pairwise, human synthesis	Annual, parallel monthly	DP	Penalized likelihood	Caussinus and Mestre (2004)
USHCN	Pairwise, automatic synthesis	Serial monthly	HBS	Statistical test (MLR)	Menne et al. (2009)
AnClim	Reference series	Annual, parallel monthly	HBS, moving window	Statistical test	Štepanek et al. (2009)
Craddock	Pairwise, human synthesis	Serial monthly	Visual	Visual	Craddock (1979) and Brunetti et al. (2006)
RhtestV2	Reference series or absolute	Serial monthly	Stepwise	Statistical test (modified Fisher)	Wang (2008)
SNHT	Reference series	Annual	HBS	Statistical test (MLR)	Alexandersson and Moberg (1997)
Climatol	Reference series	Parallel monthly	HBS, moving window	Statistical test	Guijarro (2011)
ACMANT	Reference series	Annual, joint seasonal	DP	Penalized likelihood	Domonkos et al. (2011)

Source: Venema et al. (2012, Table 1)

climate, since trends in the data introduced by non-climatic factors can be of similar size to the changes expected due to increased greenhouse gas emissions.

In contrast to the "data" of Stages 0–3, ISTI refers to Stage 4 and 5 results as "data products" (Thorne et al. 2011). It may be tempting to think that this shift in terminology reflects a substantive change, with later-stage data being, for instance, somehow more heavily processed. This is not really the case, however. As noted above, even some Stage 1 data held by ISTI have been subjected to quality control and homogenization by their sources (see Table 1 above). Thus, while Stage 4 and Stage 5 data will in fact reflect some additional processing by users, similar processing efforts will have already been made with respect to some of the data at earlier stages. ISTI's distinction between "data" and "data products" primarily marks the boundary of ISTI's control; results generated by third parties using ISTI's databank are "data products".

3 Evaluating Data Journeys: Benchmarking and Its Importance

ISTI scientists hope that users of the databank will develop multiple, independent data products for a given region and period. They hope, for instance, that a variety of reconstructions of global and regional temperature evolution over the twentieth century will be developed, where users apply their own preferred methods for quality control and homogenization to Stage 3 data. Such independent estimates, it is thought, could help to shed light on the extent to which there is uncertainty about temperature trends and other quantities commonly derived from such reconstructions: "Multiple products are the only conceivable way to get even a simple estimate of the structural (methodological choices) uncertainty; we need to attack the problem from many different a priori assumptions to create an ensemble of estimates" (Thorne et al. 2011, ES44). Although there are various climate data products already in existence, "quality assurance information is sparse, documentation quality is mixed, and different source data choices and methods can make meaningful intercomparison hard" (ibid). One reason that quality assurance information is sparse is that it is difficult to produce such information in a reliable way. Climate scientists do not have access to the true evolution of regional and global temperatures, nor to some known-to-be-accurate estimates, against which data products can be evaluated.

Benchmarking exercises are now emerging as one approach to learning about the reliability of methodologies used in generating climate data products – that is, in evaluating particular parts of climate data journeys. In very general terms, a benchmark can be understood as "a test or set of tests used to compare the performance of alternative tools or techniques" (Sim et al. 2003). The most ambitious benchmarking exercise to date in climate science is the COST-HOME (European Cooperation in Science and Technology – Advances in Homogenization Methods of Climate Series) project. COST-HOME developed a benchmark dataset and published it online, allowing anyone to attempt to homogenize it and submit data products for evaluation (see Venema et al. 2012). The COST-HOME benchmark dataset included three different types of data, but most contributors focused on the "surrogate data" portion, which was considered the "most realistic" of the three types (ibid., 92). These surrogate data, which represented conditions at a number of small networks of observing stations, were produced with the help of statistical methods, such that they reproduced important statistical features of real homogenized data, such as their "distribution, power spectrum and cross spectra"; several known types of inhomogeneities and other "data disturbances" were then added, and the task for participants was to recover the homogenous surrogate data (ibid.). Importantly, those homogenous data were not disclosed to participants until after a deadline for submission of data products. Twenty-five submissions were received, based on 13 different homogenization methods, including some manual methods (ibid.). These were evaluated on a variety of metrics that measure similarities between the submitted data product and the homogeneous surrogate data (i.e. "truth").

ISTI envisions a benchmarking scheme that is similar to that of COST-HOME in some respects. Participants submitting data products for evaluation will not know in advance the "true" underlying data to which inhomogeneities were added. In addition, the benchmarking exercise will be open to all. In fact, ISTI "strongly advocates" that anyone producing Stage 5 data products from the databank take part in benchmarking exercises (Willett et al. 2014). But there are also some differences. Rather than data for small networks of stations, ISTI plans to construct global benchmark datasets, representing what they refer to as "analog inhomogeneous worlds" (ibid.; Thorne et al. 2011), i.e. analogues to the inhomogeneous data collected in the real world. In addition, the construction of these benchmarks will begin not from homogenized real data, but from computer simulations from global climate models.[4] These simulation results, which include values of temperature on a regular grid, will be interpolated to a set of 30,000+ stations analogous to those in the databank (Willett et al. 2014). Inhomogeneities will then be added to these "analog-clean worlds", to produce "synthetic data". The inhomogeneities are intended to be "physically plausible representations of known causes of inhomogeneity (e.g. station moves, instrument malfunctions or changes, screen/shield changes, changes to observing practice over time, and local environment changes)" (ibid., 192). See Fig. 2 for a depiction of some of the ways in which the benchmarking exercise mirrors the analysis of the "real" ISTI databank data.

ISTI highlight several positive features of their envisioned simulation-based approach to the generation of benchmarking datasets. Time series of temperature values from a climate model will be free from inhomogeneities, so the "true" climate signal will be known. In addition, the data will include "globally consistent variability", including coherent variability associated with events like El Nino – Southern Oscillation (ENSO). Moreover, it will be possible to generate inhomogeneous worlds with different levels of background climate change, since climate models can be run under a variety of scenarios in which greenhouse gas concentrations are rising rapidly, held constant, etc.; at least some information then can be obtained about how the skill of different homogenization algorithms varies, if at all, with the level of background climate change.

ISTI proposes to provide ten inhomogeneous worlds/synthetic datasets in a given benchmarking cycle, each based on a different simulation, with the cycle of analysis and evaluation repeating roughly every 3 years (ibid.). The aim is for these different worlds to incorporate inhomogeneities with a range of frequencies and magnitudes, seasonality, and geographical pervasiveness (e.g. when a whole network changes observing practices at once). Participants would submit their homogenized benchmark data for evaluation by ISTI. The results of this assessment as well as "truth"

[4] These climate models incorporate both basic physical theory (from fluid dynamics, thermodynamics, etc.) and some simplified/idealized representations of small-scale processes; the latter are necessary in part because limited computational power constrains the resolution at which the climate system can be represented. The knowledge on which the models are based, including the theoretical knowledge, is of course empirical, but the climate models are not data-driven models obtained by fitting curves to observations.

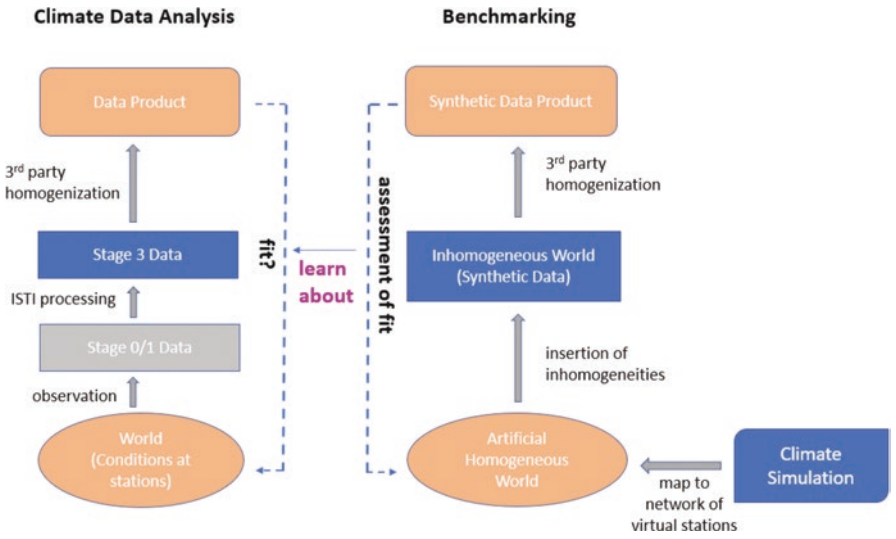


Fig. 2 Envisioned benchmarking of homogenization algorithms. ISTI's analogue worlds allow for testing of homogenization algorithms in cases where "truth" is known. The aim is to learn about how these algorithms are likely to perform on real data where similar inhomogeneities are present but truth cannot be known

for the ten cases – i.e. the clean analog worlds produced by sampling/interpolating simulation results – would subsequently be unveiled. The cycle would then repeat.

ISTI's envisioned benchmarking system is intended to support three important scientific goals of the ISTI project: quantification of the potential structural uncertainty of a given climate data product; objective intercomparison of such data products; and advancing homogenization algorithm development (Willett et al. 2014, 192). These are discussed here in reverse order.

The benchmarking scheme aims to support homogenization algorithm development by helping developers to learn more about the strengths and weaknesses of their algorithms – which sorts of inhomogeneities they are good at detecting and correcting, which they are not, etc. In further support of this goal, ISTI plans to provide some "open benchmarks" for which "truth" is also immediately available, so that participants can conduct some preliminary tests of their algorithms before submitting to the evaluation exercise. But the task of homogenizing data for which "truth" is not known to algorithm developers remains very important, since for these cases developers cannot optimize their algorithms to specific features of known inhomogeneities in the data; such optimization can make an algorithm a good performer on that particular dataset, even though it might perform poorly on datasets with somewhat different (but still plausible) inhomogeneity profiles.

It is important to recognize that, insofar as what is learned via ISTI's benchmarking exercises leads to changes in homogenization algorithms, data journeys of the *future* that involve the application of those algorithms will be somewhat different too.

Reconstructions of the evolution of global and regional surface temperatures since pre-industrial times will be produced again and again as new observations are made and additional past data are rescued and digitized; with homogenization algorithms that are changed in light of past benchmarking exercises, those reconstructions will be somewhat different than they otherwise would have been. Thus, while the sort of benchmarking exercises envisioned by ISTI can be considered "external" to data journeys involving real data, they can influence those journeys by prompting adjustments to homogenization algorithms whose application constitutes part of the journey.

Second, the benchmarking scheme supports the goal of objective and meaningful intercomparison of climate data products, such as reconstructions of global temperature change over the twentieth century. As noted earlier, for some types of data product there already have been multiple products developed by different scientific groups, but it is often difficult to compare the quality of these products, in part because they are constructed from somewhat different source data and in part because there can be no appeal to "truth" to settle the matter. In the benchmarking exercise, participants will all be starting from the same synthetic dataset; differences in their performance will be attributable to differences in their processing methodologies. Moreover, performance on the synthetic data will be objectively assessable, since for these data "truth" is known. Learning about such performance can be useful not only for homogenization algorithm developers (as just noted above), but also for users of climate data products. For instance, if such evaluation reveals that some homogenization algorithms are particularly good at correcting for some types of errors that are, for a user's intended application, particularly important to avoid, users can choose to work with data products generated with those homogenization algorithms. (In effect, users would then be selecting data products on inductive risk grounds, informed by what is learned via benchmarking activities.) This is just one important way in which the ISTI project can support climate-related research, including research intended to inform societal decision making (often called "climate services").

Finally, and relatedly, the benchmarking exercise supports ISTI's goal of providing information about uncertainties associated with climate data products, in particular uncertainties stemming from the process of homogenization. One of the potential benefits of an open-access observational databank is that multiple, independent groups can use the databank to construct data products for the same regions and periods; since there are uncertainties about how best to carry out that construction process, especially in the homogenization step, and since different groups will make somewhat different methodological choices in the face of that uncertainty (see Sect. 2 and Table 2), the products generated by the different groups can, in principle, sample current scientific uncertainty about past conditions in a particular region/period. This is analogous to the way in which a set of forecasts from different weather prediction models can, in principle, sample current scientific uncertainty about tomorrow's weather conditions. But just as there may be weather prediction models that have strong biases in particular regions – and whose forecasts for those regions we thus wouldn't want to take at face value – so can there be homogenization methods that have particular strengths and weaknesses that (if known) should

affect how we interpret their results. By helping to reveal those strengths and weaknesses, the benchmarking exercise can aid the interpretation of the set of data products generated, including whether their face-value spread should be considered a lower bound on current uncertainty.

Closely related to this is another important, beneficial function that the benchmarking scheme can serve, though it is not often emphasized by ISTI: a gate keeping function. When it comes to the generation of data products using the databank, ISTI explicitly encourages "contributions from non-traditional participants" (Thorne et al. 2011, ES44). They recognize the possibility of "useful insights from people tackling the problem by thinking "outside the box"" (ibid.). But while this is indeed a potential benefit of an open-access databank, there is also the risk that users with insufficient expertise, political motivations, and so on will decide to generate their own data products, e.g. their own reconstructions of global temperature change over the twentieth century. Such data products may, either unintentionally or intentionally, give a highly misleading picture of the evolution of past climate conditions. For example, suppose that a homogenization algorithm effectively guaranteed that temperature reconstructions would show very little twentieth century warming, almost regardless of the data; the worry arises that such a reconstruction would be touted in sceptical blogs, newspapers, etc. and would add further confusion to public discussion of climate change. If those generating the reconstruction were to participate in ISTI's benchmarking exercises, however, it might be revealed that their methodologies were highly flawed, in the sense that they did not recover anything like the "truth" in the benchmark cases. The benchmarking system thus could provide "a way of separating the wheat from the chaff" (Stott and Thorne 2010, 159) when it comes to data products generated from the ISTI databank. Of course, anyone might refuse to participate in ISTI's benchmarking exercises, but this refusal could itself constitute reasonable grounds for questioning the reliability of data products that differ markedly from those produced by others.

Thus, far from being an ancillary component of the ISTI project, synthetic data have crucial roles to play alongside "real" climate data when it comes to learning about past climate change; without synthetic data, and the accompanying benchmarking scheme, some of the primary scientific goals of the ISTI project would be in jeopardy. This does not mean, of course, that there are no limits to what benchmarking can achieve. The kinds of benchmarking exercises envisioned by ISTI can only gauge the performance of homogenization algorithms with respect to the particular inhomogeneities inserted into the synthetic data; even if an algorithm were to consistently and perfectly recover the "truth" in benchmarking exercises, this would be no guarantee that it performs similarly well on real climate data, since there is no guarantee that the inhomogeneities in the latter are fully encompassed by the inhomogeneity types present in the benchmark data. There may be types of inhomogeneities in actual climate data that go beyond those that current scientists have good reason to believe are sometimes present. Moreover, though the use of synthetic data generated with the help of simulation models has the attractive features discussed above, it is also true that simulation results (and synthetic station data interpolated from them) may lack some spatial and temporal characteristics of

real climate data, due to limitations of the climate models used (e.g. their omissions, simplifications, etc.). The ISTI benchmarking team suggests checking empirically whether synthetic data display key statistical properties of real climate data (e.g. levels of correlation among data for nearby stations, station autocorrelation, etc.), using real data that are thought to be of relatively high quality (Willett et al. 2014, 191).

4 Another Variety of Model-Data Symbiosis

In his insightful analysis of the development of modern meteorology and climate science, Paul Edwards (1999, 2010) argues that we find in these domains a kind of symbiosis between models and data – a mutually beneficial but mutually dependent relationship. Computer models of the atmosphere and climate system, he points out, are *data-laden* to a certain extent: in addition to equations from fundamental physical theory, they require various "semi-empirical parameters" that are derived (in a loose sense) from observations. At the same time, weather and climate data are often *model-filtered*. Here he has in mind several kinds of models.

Most striking is the use of computer simulation models in a process known as "data assimilation". A weather forecast from a computer simulation model provides a first-guess estimate of the atmospheric state, which is then updated in light of available observations to arrive at a revised, best-guess estimate of the state; this best-guess estimate then serves as the initial conditions for the next set of forecasts from the weather model. The same sort of technique has been used retrospectively in climate science, to generate long-term gridded datasets from gappy, irregular historical observations. These "reanalysis" datasets complement the kinds of climate data products described in previous sections of this paper (Parker 2016). When it comes to those data products, Edwards notes that what might be called "intermediate models" – which include models of instrument behaviour, techniques for quality control and many other methods (1999, 450) – are essential to their production; he explicitly notes their use in the process of homogenization.

ISTI's benchmarking scheme employing synthetic data illustrates yet another variety of model-data symbiosis in climate science, once again involving computer simulation models. Here, however, simulation models are used not to *fill in gaps in datasets* (as they in effect are used in data assimilation) but rather to help *evaluate the quality of datasets/data products*, by helping to assess the strengths and weaknesses of some of the methods used in the production of those datasets/data products. An understanding of the quality is in turn important for using the datasets effectively for various purposes, including for the evaluation of computer simulation models themselves. Indeed, one of ISTI's stated motivations for constructing an open-access observational databank that includes not just monthly but daily data, is that sub-monthly data are needed for studies of changes in climate extremes, like floods and heatwaves, as well as for evaluating today's climate models' ability to

simulate such extremes. Thus, we have climate models assisting in the evaluation of climate data products, so that those climate data products in turn can assist in the evaluation of climate models – a mutually beneficial, but mutually dependent relationship.

5 Concluding Remarks

ISTI is a major effort to promote transparency and openness in the management of surface temperature data, one which has the potential not only to help circumvent the kinds of skirmishes over access to climate data that have occurred in the recent past but also to provide better insight into the uncertainties associated with existing estimates of changes in temperature since pre-industrial times. Its success in the latter, however, depends not only on users actually generating data products that reflect a range of different methodological choices, but also on there being a means of ensuring that these products are of sufficient quality. While still under development, an ingenious benchmarking scheme, involving tests of data processing algorithms on synthetic data, is meant to serve as one important way of gauging the quality of user-generated data products. Far from being an ancillary component of the ISTI project, the benchmarking system has crucial roles to play, not only in advancing the scientific goals of the project but also by serving an important gatekeeping function in the complex and politicized context of climate change research.

The use of synthetic data in benchmarking efforts like that envisioned by ISTI also illustrates a distinctive variety of Edwards' model-data symbiosis in climate science. While he calls attention to cases in which computer simulation models have been used to help fill in gaps in observational data, the envisioned use of synthetic data in benchmarking exercises would involve simulation models aiding the process of evaluating climate datasets, including their attendant uncertainties. These datasets in turn are to be used for, among other purposes, evaluating climate models themselves. Once again, we find climate models and climate data standing in a mutually beneficial but mutually dependent relationship.

References

Alexandersson, H., and A. Moberg. 1997. Homogenization of Swedish temperature data.1, Homogeneity test for linear trends. *International Journal of Climatology* 17: 25–34.

Brunetti, M., M. Maugeri, F. Monti, and T. Nanni. 2006. Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series. *International Journal of Climatology* 26: 345–381.

Camuffo, Dario, and Chiara Bertolin. 2012. The Earliest Temperature Observations in the world: The Medici Network (1654–1670). *Climatic Change* 111 (2): 335–363. https://doi.org/10.1007/s10584-011-0142-5.

Caussinus, H., and O. Mestre. 2004. Detection and correction of artificial shifts in climate series. *Applied Statistics* 53: 405–425.

Costa, Ana Cristina, and Amílcar Soares. 2009. Homogenization of Climate Data: Review and New Perspectives Using Geostatistics. *Mathematical Geosciences* 41 (3): 291–305. https://doi.org/10.1007/s11004-008-9203-3.

Craddock, J.M. 1979. Methods of comparing annual rainfall records for climatic purposes. *Weather* 34: 332–346.

Domonkos, P., R. Poza, and D. Efthymiadis. 2011. Newest developments of ACMANT. *Advances in Science and Research* 6: 7–11. https://doi.org/10.5194/asr6-7-2011,

Durre, Imke, Matthew J. Menne, Byron E. Gleason, Tamara G. Houston, and Russell S. Vose. 2010. Comprehensive Automated Quality Assurance of Daily Surface Observations. *Journal of Applied Meteorology and Climatology* L49: 1615–1633. https://doi.org/10.1175/2010JAMC2375.1.

Edwards, Paul N. 1999. Global Climate Science, Uncertainty and Politics: Data-Laden Models, Model-Filtered Data. *Science as Culture* 8 (4): 437–472. https://doi.org/10.1080/09505439909526558.

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

Fleming, James R. 1998. *Historical Perspectives on Climate Change*. New York: Oxford University Press.

Guijarro, J.A. 2011. *User's guide to climatol. An R contributed package for homogenization of climatological series*. State Meteorological Agency, Balearic Islands Office, Spain: Report. available at: http://webs.ono.com/climatol/climatol.html.

Menne, Matthew J., Imke Durre, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston. 2012. An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology* 29: 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1.

Menne, M.J., C.N. Williams Jr. and R.S. Vose. 2009. The U.S. historical climatology network monthly temperature data, version 2. *B. Am. Meteorol. Soc*. 90: 993–1007. https://doi.org/10.1175/2008BAMS2613.1.

Parker, Wendy S. 2016. Reanalyses and Observations: What's the Difference? *Bulletin of the American Meteorological Society* 97 (9): 1565–1572. https://doi.org/10.1175/BAMS-D-14-00226.1.

Rennie, Jared J., Jay H. Lawrimore, Byron E. Gleason, et al. 2014. The International Surface Temperature Initiative Global Land Surface Databank: Monthly Temperature Data Release Description and Methods. *Geoscience Data Journal* 1: 75–102. https://doi.org/10.1002/gdj3.8.

Russell, Muir, et al. 2010. *The Independent Climate Change Email Review.* Available at http://www.cce-review.org/index.php. Accessed 10 Sept. 2017.

Sim, Susan E., Steve Easterbrook, and Richard C. Holt. 2003. Using Benchmarking to Advance Research: A Challenge to Software Engineering. *Proceedings of the Twenty-fifth International Conference on Software Engineering*, pp. 74–83. doi: https://doi.org/10.1109/ICSE.2003.1201189.

Štěpánek, P., P. Zahradníček, and P. Skalák. 2009. Data Quality Control and Homogenization of the Air Temperature and Precipitation Series in the Czech Republic in the Period 1961–2007. *Advances in Science and Research* 3: 23–26.

Stott, Peter, and Peter Thorne. 2010. How best to log local temperatures? *Nature* 465: 158–159.

Szentimrey, T. 2007. Manual of homogenization software MASHv3.02. *Hungarian Meteorological Service* 65.

———. 2008. *Development of MASH homogenization procedure for daily data*. Proceedings of the fifth seminar for homogenization and quality control in climatological databases, Budapest, Hungary, 2006. WCDMP-No. 71, pp. 123–130.

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..

Thorne, Peter, et al. 2011. Guiding the Creation of a Comprehensive Surface Temperature Resource for Twenty-First-Century Climate Science. *Bulletin of the American Meteorological Society* 92: ES40–ES47. https://doi.org/10.1175/2011BAMS3124.1.

Venema, Victor K.C., et al. 2012. Benchmarking Homogenization Algorithms for Monthly Data. *Climate of the Past* 8: 89–115. https://doi.org/10.5194/cp-8-89-2012.

Wang, X.L.L. 2008. Accounting for autocorrelation in detecting meanshifts in climate data series using the penalized maximal t or F test. *Journal of Applied Meteorology and Climatology*. 47: 2423–2444.

Willett, Kate, et al. 2014. A Framework for Benchmarking of Homogenisation Algorithm Performance on the Global Scale. *Geoscientific Instrumentation Methods and Data Systems* 3: 187–200. https://doi.org/10.5194/gi-3-187-2014.

Wendy S. Parker is Associate Professor of Philosophy at Durham University, where she also codirects the Centre for Humanities Engaging Science and Society (CHESS) and the Institute for Data Science (IDAS). She received her PhD in History and Philosophy of Science from the University of Pittsburgh in 2003. Her research focuses on the methodology and epistemology of contemporary science – especially questions related to modelling, evidence, explanation and values – with a particular focus on climate science and meteorology. Her papers have been published in a range of philosophical and scientific journals.

The Babel of Drugs: On the Consequences of Evidential Pluralism in Pharmaceutical Regulation and Regulatory Data Journeys

Niccolò Tempini and David Teira

Abstract Throughout the last century, pharmaceutical regulators all over the world have used various methods to test medical treatments. From 1962 until 2016, the Randomized Clinical Trial (RCT) was the reference test for most regulatory agencies. Today, the standards are about to change, and in this chapter we draw on the idea of the data journey to illuminate the trade-offs involved. The 21st Century Cures Act (21CCA) allows for the use of Electronic Health Records (EHRs) for the assessment of different treatment indications for already approved drugs. This might arguably shorten the testing period, bringing treatments to patients faster. Yet, EHR are not generated for testing purposes and no amount of standardization and curation can fully make up for their potential flaws as evidence of safety and efficacy. The more noise in the data, the more mistakes regulators are likely to make in granting market access to new drugs. In this paper we will discuss the different dimensions of this journey: the different sources and levels of curation involved, the speed at which they can travel, and the level of risk of regulatory error involved as compared with the RCT standard. We are going to defend that what counts as evidence, at the end of the journey, depends on the risk definition and threshold regulators work with.

1 Introduction

Since the early 1900s, the US Food and Drug Administration (FDA) has been using laboratory and clinical experiments to test toxicity and safety of pharmaceutical compounds before approving their release to the market (Carpenter 2010). In other words, the FDA sets the threshold of risk that patients can take when they decide

N. Tempini
Department of Sociology, Philosophy and Anthropology & Exeter Centre for the Study of the Life Sciences (Egenis), Exeter, UK

Alan Turing Institute, London, UK
e-mail: n.tempini@exeter.ac.uk

D. Teira (✉)
Department of Logic, History and Philosophy of Science, UNED, Madrid, Spain
e-mail: dteira@fsof.uned.es

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_11

between treatment options. However, for the last 50 years, only data from highly standardized experiments, Randomized Clinical Trials (RCTs), has counted as legitimate regulatory evidence for market approval. This regime is changing. The 21st Century Cures Act (21CAA) invites the FDA to consider new evidentiary standards in assessing treatments, including data from Electronic Health Records (EHR).[1] For more than 50 years pharmaceutical regulators dealt with evidence mainly originating in single-purpose drug tests. The 21CCA allows them to use data (1) generated with goals different than testing and, (2) *repackaged* and *re-purposed* to assess the safety and efficacy of a treatment. *Travelling data* (Leonelli 2016) enter the field of pharmaceutical regulation.[2] This chapter tries to understand *what the use of different kinds of data for pharmaceutical regulation means for the assessment and comparison of risks linked to different drugs.*

The FDA is a unique institution. Its gatekeeping power is based on scientific evidence: successful tests are a pre-condition for market access. It is the most influential regulatory agency, setting the regulatory paradigm and benchmarks that others all over the world follow. Any shift has great magnitude, as FDA decisions shape a global market for prescription drugs: in 2016, worldwide sales have been estimated in 768$bn (EvaluatePharma 2017).

Pharmaceutical regulators have to strike a balance between very powerful and conflicting institutional principles: access to worthy new compounds should be granted quickly, while guaranteeing strict thresholds of safety and efficacy. They also need to face conflicts of interest in the industry (e.g., patent versus generic manufacturers) and also among patients (e.g., depending on their risk aversion in trying new drugs). Regulatory tests have enforced the epistemic standard that arbitrates these conflicting principles: if a promising drug fails, the sponsor will lose the investment and patients their hope, but the result will be accepted.

The 21CCA introduces evidentiary pluralism in drug testing: instead of a single source of *regulatory objectivity* (Cambrosio et al. 2006), the FDA will define different standards about what counts as evidence of safety and efficacy in a treatment for each research design allowed, including how to evaluate EHR that can now be put in motion and journey towards yet new uses (Leonelli, this volume, Chap. 1). The implications are not clear. Some welcome the initiative as necessary for bringing

[1] Signed by President Obama on December 13, 2016, the 21CCA is law aimed at expediting "the discovery, development, and delivery of new treatments and cures and maintain America's global status as the leader in biomedical innovation." Section 2061 invites the FDA "to issue guidance that addresses using alternative statistical methods in clinical trials and in the development and review of drugs." Section 2062, on which this papers focus, claims: "To support approval of a drug for a new indication, the FDA must evaluate the use of evidence from clinical experience (in place of evidence from clinical trials) and establish a streamlined data review program". For the full text, see: https://www.congress.gov/bill/114th-congress/house-bill/6 (Accessed on August 24, 2107)

[2] Sabina Leonelli theorizes the travelling of data as the achievement of purposeful strategies aimed at enabling the reuse of scientific data in situations and for uses that can be different from those originating the same data (Leonelli, this volume). To travel, it is paramount that data keep the capacity to hold evidential value in new stages of the inquiry (Leonelli 2016). She calls these strategies of data repurposing *packaging*.

new treatments to the market, others denounce it as paving the way for pharmaceutical fraud.[3] We contribute to this ongoing debate with an analysis of the epistemic and political implications of the use of EHR for drug assessment: what can we expect from regulatory decisions based on these data?

EHR are systems digitalizing medical records in standardized formats, to gather information regarding a) patients, as obtained during their visits to medical facilities (e.g. clinical interview, anamnesis and assessment, or diagnosis in an emergency room); b) the complementary evidence generated by those visits (e.g., medical imagery, test results); and c) data gathered by measurement devices that patients wear and use while away from the point of care (e.g., glucose sensors inserted under the skin). EHR data are generated in the context of routine activities that are shaped not only by standards of care but also statistics, administration and billing. They are not a record of scientific observation and intervention performed in isolation. They are a product of hybrid accounting practices, a record of clinical care just as much as of auditable administration (cfr. Ramsden, this volume, Chap. 17; Power 1997). It is very difficult to clean the data from the traces of interactions that are not of interest, and as such the re-use of EHR requires complex arrangements and specialised expertise (Tempini and Leonelli 2018).

To use EHR for regulatory activity requires different standards of practice and evidence than those involved in the evaluation of RCT results. Drawing on the Guidance documents so far issued by the FDA and on our own fieldwork in EHR reuse,[4] in Sect. 4 we argue that the successful use of EHR in tests depends on adequate data management. In order to control for bias, experts need information about potential confounders. This should be included in the travel package (cfr. Leonelli 2016) of the EHR, or be otherwise available. Against a popular belief, we shall here rehearse an old statistical argument about how Big Data, on its own, is not going to correct for such biases (also Boyd and Crawford 2012). The implication is that the evidential standards of the new regulatory pluralism will be different: *what counts as evidence will depend on the risk threshold one works with*. Risk involved in using drugs approved through new testing standards might be passed downstream, to patients and their carers. They will have to decide whether to take the risks involved in taking drugs tested with inferior evidentiary standards.

In Sect. 2, we will defend the claim that the pre-21CCA regulatory regime hinged on two value judgments: the FDA should (a) behave as a strongly paternalist regulator and (b) adopt the RCT as sole source of evidence for safety and efficacy. In Sect.

[3] For a glimpse of this debate, see (Avorn and Kesselheim 2015) and the responses in the same issue, plus (Kesselheim and Avorn 2017)

[4] This paper draws on the qualitative fieldwork of one of us (Tempini) on two leading EHR-reuse infrastructures: the Secure Anonymised Information Linkage databank (SAIL) and the Medical and Environmental Data Mashup Infrastructure (MEDMI). Visiting and interviewing developers, managers, data scientists and clinical researchers between September 2015 and January 2017, he approached these infrastructures with a view to document the associations between organisational forms and processes, activities of infrastructure development, data practices, and scientific research practice that underlie the debate on EHRs.

3, we show how the relaxation of standards of evidence (b) has also relaxed regulatory paternalism, giving patients more access to treatments approved on different sources of evidence. Regulating with travelling EHR data would imply a further step away from paternalism: we still don't know how good this evidence is as a source of decisions about treatments. In Sects. 4 and 5, we review some of the challenges involved in standardizing evidence from EHR. We advise paternalistic caution arguing that even the most libertarian patients would want to know how reliable a testing standard is, in order to make an informed decision about treatment options.

2 Regulation with Non-travelling Data

For more than half a century, the international paradigm in drug regulation was set by the 1962 amendment to the FDA Act: a pharmaceutical company seeking approval for the commercialization of a new treatment should submit "adequate and well-controlled clinical studies" for evidence of efficacy and safety. The definition of a well-controlled study would not be clarified until 1970, when it was defined as two well-controlled clinical trials.

Testing treatments with RCTs is a long process. RCT data are gathered according to a research protocol in which statistical considerations are paramount. Treatment effects should be estimated in trials that have a given statistical *power*: the patients' sample size will determine the probability of making a type I error (accepting into the market inferior treatments). Once the administration of the treatment and the measurement intervals are pre-established in the trial protocol, the duration of the trial will mostly depend on the amount of time it takes to enroll the predesignated number of patients and consequently execute the protocol. According to (DiMasi et al. 2016), the average time from the start of clinical testing to marketing approval is 96.8 months. Administering a treatment to a patient may take weeks or months until the time comes to measure the target outcome. Gathering the data for enough completed treatment protocols may take years, as enrolment is difficult and time-consuming and depends on the condition of interest and the eligibility criteria for the participants. Even well-funded, high-impact research fields suffer from slow trials: for instance, only 3–5% of cancer patients enrol in trials (Bell and Balneaves 2015).

From the standpoint of many patients, the wait is too long. Although some of them might have early access to the drug through trial participation (e.g., via Right to Try Laws: Carrieri et al. 2018), everybody else wait till market approval to benefit from the treatment. Even the "luckiest" few, those patients who benefit from an effective drug within the trial, might have to wait for years after the protocol completion before they can access the drug again in the market. Also the industry argues that the process is too long, although for different reasons. A company will reap most profits from a compound during patent time, and this starts counting before the RCTs even start. The longer the testing and approval process, the less patent time to exploit commercially.

Why then did regulatory authorities choose this route? The main reason is normative: the 1962 Act gave the FDA a *paternalistic* gatekeeping power on pharmaceutical markets in order to protect patients from repeats of past pharmaceutical catastrophes (e.g., Thalidomide). The safety and efficacy of a product should be assessed *ex ante,* before market release. RCTs were chosen as the sole regulatory yardstick thanks to the advocacy of American pharmacologists, who defended their superiority to grasp treatment effects on the basis of a sample of patients (Marks 1997; Podolsky 2015). Other sources of evidence about treatments (e.g., case studies), until then in use to assess their effects by doctors, were discarded in regulation.[5]

The 1962 Act hinged then on two value judgments (Teira Forthcoming). First, the 1962 FDA ACT established a strongly paternalist regulatory body. Physicians and patients were deprived of treatments lacking safety and efficacy without their consent and for their own good. Second, the RCT was selected as the gold standard for determining whether a treatment was safe and efficacious. Any concerns are subordinated to the greatest good the regulator should protect: the safety and security of the pharmaceutical consumer.

With this normative justification, *the regulator exclusively evaluates non-travelling data*. Trial data are indeed designed for one-use only: testing the safety and efficacy of treatments. RCTs are not experiments to learn in which the experimenter is free to try as many things as she may see fit, in order to find out how a treatment works. RCTs are experiments to prove (Teira 2013), in which the whole test design serves the purpose to convince the regulators that a treatment is safe and effective. They are a paradigmatic example of hypothesis-driven research. RCT data are rarely re-purposed for other ends and their 'travel equipment' is consequently basic: datasets store the outcomes for the different variables measured, in a format suitable for statistical analysis. These data are seldom portable onwards, to the clinics where patients receive care after the trial.[6] The situation of inquiry within which trial data are used does not usually change.

Yet, trial data move. Trials are often distributed over a number of different sites and for this reason their organization requires to carefully consider issues related to metadata and the standardization of practices. In order to speed up the testing pro-

[5] Case reports provided evidence about safety and efficacy in the first decades of the twentieth century, but they were gradually displaced by statistical trials from the 1930s onwards. The rise of Evidence Hierarchies in the 1990s consecrated the principle that statistically designed experiments could deal with the biological variability of treatment effects in patient populations. However, case reports have since attracted efforts of standardization, as Rachel Ankeny shows in her chapter this volume, aimed at reclaiming their evidential value. The developments on the regulatory use of EHR we are focusing on here can be seen, in Ankeny's terms, as part of those "emerging efforts to develop deeper understandings of appropriate, effective, and rigorous ways of using observation-based methodologies in the biomedical sciences". In this chapter we will point out some of the limitations of these initiatives as well.

[6] As a matter of fact, experienced researchers have found difficult to simply handle them: according to (Gøtzsche 2013), the clinical documentation for just three drugs he tried to access in 2010 from the Swedish regulatory agency was compiled in 70 meters of binders, about half a million printed pages.

cess, trials are conducted in multiple clinical facilities, where patients are admitted and treated in accordance to a shared protocol. For the first three decades after the 1962 Act, these facilities were standard health institutions in which the trial participants were recruited among the accruing patients. From the 1990s onwards, the industry has sponsored the rise of Contract Research Organizations (CRO) that find patients wherever they are and enroll them in the trial protocol on a dedicated site, not always a conventional medical facility. By 2005, only 25% of all pharmaceutical research was conducted in academic medical centers (as opposed to 80% before 1990) (Fisher 2009).

The mobility of data is monitored by regulatory bodies with careful audit rules (Helgesson 2010). For almost 20 years now, the FDA has developed guidance documents establishing the ALCOA principles of data quality to be observed in either electronic or paper records (e.g., CDER 2018). Data should be Attributable, Legible, Contemporaneous, Original and Accurate. The records should document who created or changed a record; they should be readable (to third parties); they must contain a time stamp of its generation; they should be the first place where the data are recorded; and they should be faithful to the actual measurement. The major problem with trial data is that their mobility stops as soon as they reach the sponsoring company headquarters: according to a recent study, an astonishing 45,2% of the outcomes of the approximately 25.927 RCTs registered at ClinicalTrials.gov by major trial sponsors have not been published (Powell-Smith and Goldacre 2016). There have been prominent campaigns advocating for a legal mandate to register all the conducted trials and release the raw outcomes (e.g., AllTrials.net) and the European Union is about to implement a systematic policy in that regard.[7]

Yet, even if trial data were routinely released to the public, it would be mostly for replication and validation of the sponsor analyses. As of today, there are no systematic plans of curating these data into databases for general research purposes.[8]

3 Regulation with Travelling Data

The 1962 FDA Act established a paternalistic pharmaceutical regulator with a single standard of evidence for testing safety and efficacy. *But if the FDA approaches pharmaceutical regulation with different value judgments, we may let other kinds of data to travel and be used as evidence in regulation*. Already in the 1970s, libertarian critics of the FDA made this possibility explicit (Wardell and Lasagna 1975). If patients were allowed access to experimental treatments (under the prescription of a qualified physician and an informed consent form) regulatory agencies would 'simply' need to collect adverse event reports as promptly as possible. They could then proceed, when necessary, to withdraw unsafe treatments. In this anti-paternalist

[7] Through the Clinical Trial Regulation EU No. 536/2014 to be implemented in 2019.

[8] See, for instance, http://www.alltrials.net/find-out-more/all-trials/ (Accessed on August 26, 2017)

approach, physicians and patients are free to explore treatment options. Pharmaceutical regulators exploit the data users generate with whatever statistical tools available. Adverse event data from any source should travel to the regulator's desk.

However, pharmaceutical regulation is not only about *ends*: it is also a matter of *means*. In the 1970s, such a reporting system would have been probably paper-based and relatively slow in processing and acting upon information. Contergan, the German brand name of the sleeping pill sold in the US as Thalidomide, was withdrawn from the German market 'only' a couple of months after its adverse effects were noticed in a medical journal. But at that point, 4000 children had already been born with severe deformations (Gaudillière and Hess 2012, pp. 1–2). Even a libertarian regulator could be averse to the possibility of a pharmaceutical catastrophe with too many patients harmed for delayed reporting, detection and reaction.

Both ends and means have shifted throughout the last five decades. First of all, *regulatory paternalism has been gradually relaxed*, mostly after the participants' revolts during the antiretroviral AZT trials in the 1980s (Epstein 1996). AIDS patients advocated for their freedom to decide which treatment to take, against trial designs that imposed placebos on some of them. In response to their demands, the FDA introduced an early access system based on quicker trials with surrogate endpoints: instead of following the treatment until its final outcome, the trials tracked a variable that predicted this outcome, shortening the testing process. However, this prediction may fail. Critics of the pharmaceutical industry have denounced that treatments tested in trials with surrogate outcomes have a different level of safety and efficacy than compounds tried in standard RCTs (Gonzalez-Moreno et al. 2015; Pease et al. 2017). In other words, the FDA offers different levels of patient protection according to the testing standard it chooses. Nonetheless, patients (with or without the support of the pharmaceutical industry) have continued to advocate their right to try experimental treatments, even when there is no solid RCT evidence to support them.[9] Although the FDA remains the gatekeeper to the pharmaceutical marketplace, its paternalism has been implicitly softened with the relaxation of its testing standards.

As to the means, during the last decade, the rise of computing and digital networks enabled the diffusion of the *electronic health record* (EHR): according to the regulator, EHR systems are "electronic platforms that contain individual electronic health records for patients and are maintained by health care organizations and institutions." (FDA 2016, p. 4). With EHR clinical data can start travelling more easily, but the landscape is fragmented. There are multiple sources for EHR and many different ways to exploit them. In the first place, there are hospitals and all sorts of medical institutions (from physician offices to multi-speciality practices), but also insurance claims databases and registries. The multitude of vendors providing EHR systems has made the achievement of data interoperability and comparability a long-term issue requiring sustained standardization efforts. Recently, relative advancements in standardization united with cheap availability of enormous com-

[9] See, for instance, the recent FDA decision to approve eteplirsen (Exondys 51), against the recommendation of its own scientific board but accepting the demands of patients with Duchenne muscular dystrophy (Kesselheim and Avorn 2016).

puting capabilities have made it possible for some infrastructures to achieve a scale of data integration that could only dreamed of only a decade ago.

E.g., Kaiser Permanente is a US based integrated managed care consortium with 11.7 million health plan members as of October 2017 (Wikipedia, March 1, 2018). It is now constructing a virtual data warehouse, with a view to study the effectiveness and safety of the treatments prescribed. Kaiser Permanente is just one of the sources feeding the Sentinel initiative (FDA 2018), by which the FDA is monitoring the safety of medical products already in the market, drawing on normalized and validated records from a group of data partners. As of 2017, Sentinel was accessing data from 193 million individuals. At an international level, the Observational Health Data Sciences and Informatics (OHDSI) is a collaboration between researches in 12 countries based on a Common Data Model that specifies how to encode and store clinical data. As of 2016, there were 52 databases, with a total of 682 million patient records (Hripcsak et al. 2016).

Yet, as of today, these are all pioneering initiatives: database interoperability and standardization is not the norm (Fleming et al. 2014; Ford et al. 2009; Lyons et al. 2009). EHR are extensively used in healthcare management, both for administrative and clinical purposes. The use of EHR for other purposes is mostly derivative. Scientific concerns have not been top priority in EHR design and practice. The generation and maintenance of EHR data has instead been shaped by the situated requirements of healthcare, local information infrastructure and institutional routines and reporting policies. It is thus difficult to render different sets of EHRs comparable (Demir and Murtagh 2013). This requires intensive "cleaning", curation and external validation. Furthermore, there are serious privacy issues: EHR contain personal information and there are a number of legal and procedural principles that should be observed in their handling. Most EHR are not ready-made to travel onwards for scientific reuse.

The travelling of EHR data thus needs to be achieved through methodological, technological and organizational solutions. An increasingly frequent approach has been the creation of secure analytical environments, where researchers can transform datasets to suit their research needs (see Tempini, this volume, Chap. 13). Data transform operations are carried out through a combination of automated pipelines and human judgement and intervention. A deep knowledge of the idiosyncrasies of each dataset is paramount and some data infrastructures have dedicated data analysts to provide just such expertise (Tempini and Leonelli 2018).

New developments in respect to both ends and means of regulation set the foundations for the 21st Century Cures Act (21CCA). Epistemic and methodological novelties come together in section 2062 of the 21CCA, which opens up the possibility of using electronic health records to assess new indications for already approved treatments. It mandates the FDA "to use of evidence from clinical experience (in place of evidence from clinical trials)" and "establish a streamlined data review program" in order to support approval of a drug for new indications.

Drug repositioning is indeed a booming field (Institute of Medicine 2014). Once drugs are in the market, physicians are free to prescribe them as they see fit. Pharmaceutical companies cannot promote off-label prescription, since regulatory protection against any adverse effect liability extends as far as the indications

recorded in the treatment label – those tested with an RCT. Nonetheless off-label use is sometimes successful, at least *prima facie*. The 21CCA intends to capitalize on the wealth of information on off-label use captured in EHR systems to faster evaluate alternative indications. Assuming that any *safety* issue neglected in the original trial would have already emerged in the market, the 21CCA focuses on an alternative approach to *efficacy* testing: the clinical data may be sourced from many different contexts and the statistical techniques for the analysis of treatment effects should go beyond RCT hypothesis-testing.

The 21CCA has been a controversial bill: to name just a few of the lobbying groups involved, pharmaceutical, device and biotech companies reported more than $192 million in lobbying expenses; more than two dozen patient groups reported spending $6.4 million in disclosures that named the bill as one of their issues (Lupkin and Findlay 2016). Not all these groups were focusing on the testing standards for drug approval: the legislation is tied to a huge raise in funding for the National Institutes of Health, enough for many stakeholders in the biomedical community to support it. Yet, the 21CCA has initiated a paradigm shift in drug testing that, according to very qualified critics, will not promote "a 21st century of cures, but a return to the 19th century of frauds" (Gonsalves et al. 2016). If the new evidentiary standards for drug approval admit inferior drugs into the market, in the absence of counter-measures many patients may return to an era in which they could not tell apart good and bad treatments.

However, preferences about testing standards depend on value judgments. Many patients might want to be protected (to a given degree) by a paternalist regulator. The degree of protection they should expect depends on the testing standard for safety and efficacy. Of course, patients might be willing to take more or less risk depending on the situation, and conditions such as access to expert counselling and high quality information, ability to process complex information, and the range of options afforded by each one's insurance plan. Our assumption is that, faced with the increasing complexity involved in evaluating treatment options, most decision takers would welcome the availability of estimates of a testing standard's reliability. As we are going to discuss in the following section, the use of travelling data (via EHR) for drug testing poses precisely this question. Whereas so far EHR have been packaged without paying any attention to regulatory needs, the 21CCA opens up the possibility of converting EHR for regulatory purposes. If so, we may ask when do EHR provide reliable evidence for assessing new drug indications? How shall we measure and share the risks involved in a regulatory decision based on EHR?

4 How Far Can EHR Data Travel?

Perhaps it is too early for a conclusive answer. As we are writing (March 2018), the US Office of the National Coordinator for Health Information Technology is opening to public discussion how to articulate the 21CCA "trusted exchange framework", a first step towards achieving a flow of interoperable health information

across different networks in the country. The FDA should still issue guidance documents in order to implement the new testing standards promoted by the 21CAA. It will take years until we see the full consequences of the incorporation of travelling data into drug regulatory testing.

Yet, the debate about how to use EHRs for regulatory purposes does not start from scratch. There are already a number of FDA guidance documents about the use of EHR in both clinical trials and epidemiological studies. E.g., the FDA guidance on the *Use of Electronic Health Record Data in Clinical Investigations*, issued in July 2018, refers to the use of EHR in standard regulatory trials (CDER 2018). In accordance with the ALCOA principles mentioned above, the main goal of the document is to guarantee the auditability of every data record presented for regulatory use. More relevant for our purposes is the *Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data*, an FDA guidance issued in May 2013 (CDER and CBER 2013). It implements the same approach to data audit, but adds some significant methodological caveats:

> Investigators should demonstrate a complete understanding of the electronic healthcare data source and its appropriateness to address specific hypotheses. Because existing electronic healthcare data systems were generated for purposes other than drug safety investigations, it is important that investigators understand their potential limitations and make provisions to use the data systems appropriately. (FDA 2013, p. 13)

The implicit principle is that RCT (non-travelling) data are the evidentiary benchmark for assessing the appropriateness of EHR for drug safety investigation. If this is the case, the regulator may expect that the packaging of the EHR should include enough information for the investigator to perform the assessment of their *limitations* (as compared to RCTs), and/or human expertise to be otherwise available. But the Guidance assumes no standardized curation defining a suitable EHR. It rather leaves in the investigator's hands, the internal and external evaluation of the EHR. As to the internal assessment, the investigator should evaluate best strategies for data coding (a key step of data repurposing): "Safety outcomes that cannot be identified using International Classification of Diseases (ICD) codes cannot be appropriately studied using data sources that rely solely on ICD codes in claims data" (FDA 2013, p. 14). As to the external assessment, using again an example from the Guidance, administrative claims data generated to support payment for care should be used taking into account the payor's policies governing the approval and denial of such payments, in order not to introduce a selection bias in the analysis (e.g., patients who should have been included for clinical reasons do not leave a record if payment is denied and treatment discontinued). Investigators will need to dispose of a wealth of informal knowledge about the practices and institutional shifts that shape clinical reporting. As other auditing practices, the assessment of EHR remains opaque as to the definition of its own core matter, and is shaped by economic constraints and attitudes towards cost-benefit trade-offs (see Power 1997).

The Guidance lists, without any pretence of completeness, a number of dimensions in the EHR that investigators should consider when assessing their databases.

In order to grasp the general principle behind this assessment, we need to understand first how the comparative benchmark works. RCTs are experiments designed to generate data that would allow for a clear test of a pharmacological hypothesis: for a given population, is the new treatment for condition X equal or better than an already established alternative? The design should exclude whatever potential confounders may interfere in the outcome variables measured. RCTs should compare like with like: the circumstances of the patients in every arm of the trial should be the same except for the interventions under study, so that if any difference is observed in the outcome, we may safely attribute it to the treatment administered. Sameness between trial arms is constructed by adopting control measures for a list of potential confounders (e.g., since the expectations of patients on the treatments they receive may play a role in the outcome, they should be administered in a way that no patient can discern which of the treatments in the test they are receiving – *blinding*). Throughout the last five decades, trialists have accumulated a good understanding of the different sources of bias in their experimental setups and have organized checklists to score the reliability of a test (Higgins et al. 2011).

In order to match this level of experimental control of the treatment effect, observational studies should measure potential confounders associated with the outcome and conduct an adjusted statistical analysis that accounts for differences in the distribution of these factors between intervention and control groups. Exceptionally, the size of the treatment effect might be so large as to swamp all the potential confounders (Glasziou et al. 2007). But most RCTs do not observe very large effects – and when they do, they are not necessarily reliable (Nagendran et al. 2016). The regulator using EHR data must expect that they are packaged for travel with enough information about potential confounders as to grant a solid assessment. It is then necessary that experts construct one such list of items for EHR studies, scoring the degree of control on the treatment effects allowed by a given EHR dataset. They will need to ensure that procedures are in place to operationalize expert knowledge of the specific datasets in ways that are accountable to the regulator. For these data to travel, researchers will need to explicitly account for the uniquely contextual features of each data 'assemblage' used in a study.

As of now, best experiences in EHR re-use (exemplified in infrastructures like SAIL in the United Kingdom – Ford et al. 2009; Lyons et al. 2009) suggest that dedicated data analysts are in the best position to develop deep knowledge of the potential sources of bias in the EHR they specialise in. Specialised data analysts flank researchers in the selection, modelling, extraction and analysis of the dataset while at the same time relying on the clinical expertise of the researchers (Fleming et al. 2014). Knowing directly about the quality of the data and getting past the initial selection of variables from a list of available sources is of paramount importance (see Tempini and Leonelli 2018). Issues such as missing, unknown or uncertain values are endemic in EHR datasets. Data collection practices in the health care system are greatly variable. Even a high quality curated EHR cohort can sometimes offer only limited coverage for important confounders. For example, in the creation of an electronic cohort of children living in Wales (the Welsh Electronic Cohort for Children), data about maternal smoking could be missing for up to 50% and contrib-

uted to the redefinition of the sample set (in this case, the cohort came to comprise children born in Wales, because children who moved to Wales after birth had comparatively poorer data). Source databases could disagree on the sex of a child, requiring researchers to harmonize even 'basic' data. More generally, missing data can be detected a) at the level of the individual records; b) at the institutional level (values can be missing from all the records contributed by one organization); or c) at the infrastructural level (all records from a particular EHR software vendor). US and UK systems are fragmented into multiple infrastructures marketed by competing vendors, though industry concentration is increasing. EHR data to be made available for re-use and travel are sometimes selected by vendor, again with an uncertain effect on sampling.

There is a growing literature on the EHR study biases (Pivovarov et al. 2014; Rusanov et al. 2014; Vawdrey and Hripcsak 2013). A key concern is with the *event-based nature* of EHR data (see Jorm 2015): data are collected in the occasion of patient encounters with the healthcare system. The timing and reason for these encounters have not been pre-emptively stipulated by a study protocol and are instead associated to patient needs. Data about healthier patients are thus scarcer, and this can have implications for sample selection. Shifting reporting policies (administrative, accounting and fiscal frameworks) and other circumstances of health care coverage can mean certain phenomena are under- or over-reported (Dixon et al. 2015; Fleming et al. 2014). In addition, algorithms used for curation and modelling need to be validated: coding can be simplistic and/or overlapping, often requiring researchers to create custom code-lists and control for duplication. A complex ecosystem of practices, solutions and institutions is necessary to make scientific reuse of EHR possible (Hripcsak and Albers 2013). In 2014, a review of the state of EHR implementations in the US found that only a small proportion of systems meet "meaningful use guidelines"; while most systems met basic standards for data collection, only by 40–60% of systems satisfied criteria for the sharing of data between points of care and with public health agencies (Adler-Milstein et al. 2014). Underperforming systems are not randomly distributed and have a higher share in small and rural hospitals.

How should we think about these caveats? A quick rejoinder would contest the status of RCT data as benchmark in this comparison: many philosophers of science have defended evidentiary pluralism regarding medical causality, contesting the gold standard status of RCTs – for a review, see (Reiss 2017). Although, a priori, RCTs allow a high degree of causal control on an intervention, the theoretical assumptions behind this superiority may not hold empirically and, depending on the context, observational studies might be equally defensible. In other words, biases may equally harm RCTs and observational studies – see (Senn 2013) for a discussion. Thanks to EHR, observational studies may reach a sample size that no RCT can match and, with adequate data mining processes, true treatment effects may be detected.

Following (Senn 2008), it is worth recalling here that observational studies can improve only to a limited extent solely thanks to the addition of data. In assessing the reliability of a statistical estimator (e.g., of a treatment effect), we depend on two magnitudes. On the one hand, we have the underlying biases in the measurement

process, arising from the methodological limitations in the study we discussed above: e.g., not properly blinded patients may distort the treatment outcome. On the other hand, there is the *standard error* of the measurement process, arising from the sheer variability between the subjects measured: not every patient reacts in the same way to the treatment and we need to find a reasonable average. The standard error is (roughly) inversely proportional to the number of subjects in the study. Here comes the power of *big* data: the bigger the number of EHR, the lower the standard error. But even if the standard error tends to zero as the sample size grows, the bias will remain constant. The only reliable approach to controlling for biases is in the design of the study, and here is where RCTs dominate. This seems to be the position of the FDA research staff as of the end of 2016:

> EHR and claims data are not collected or organized with the goal of supporting research, nor have they typically been optimized for such purposes, and the accuracy and reliability of data gathered by many personal devices and health-related apps are unknown. (Sherman et al. 2016)

5 "Delivering the Proof in the Policy of Truth"

There is thus no *a priori* reason to expect that EHRs can be as reliable as conventional RCTs for regulatory purposes. If so, the 21CCA is set to push further the methodological relaxation of the FDA's regulatory paternalism. The FDA will still act as a gatekeeper, but it will allow into the pharmaceutical marketplace drugs with as many different levels of safety and efficacy as the testing standards that are used. Just as trials with surrogate outcomes turned out to be often less reliable than old-fashioned RCTs, the assessment of new indications for already approved drugs with EHRs may introduce a new safety and efficacy threshold, and one that lowers the levels of protection for pharmaceutical consumers. The incommensurability between FDA approvals based on RCT vs EHR evidence generates a risk for clinicians and patients taking a therapeutic decision between heterogeneous options. As we put forth in the introduction, *what counts as data depends on the risk threshold one works with.* The FDA, we argue, is lowering its risk threshold, and this is allowing different kinds of data to travel and be used as evidence. Clinicians and patients will then have to set their own threshold in turn.

The residual risk involved by EHR-based regulation (i.e. risk that is not controlled by FDA regulatory activity) is thus passed downstream to patients. Each patient will take decisions based on different risk thresholds and standards of evidence. Their decision will also depend on standards of care patients are able to access and the specialists they will consult with. Will patients accept drugs based on different kinds of evidence? All evidence points to a positive answer, especially if we consider the influence that pharmaceutical marketing, once deployed to promote newly approved uses, can exert on the entire cultural frame in which health is understood and evaluated (Dumit 2012). The trend, Dumit shows, is for more and the most profitable drugs to succeed.

It remains to be seen whether and how international regulatory agencies other than the FDA will revise their position with respect to the use of EHRs in regulation. At stake, we have competing forces in pharmaceutical markets. The 21CCA is supposedly addressing the crisis of innovation in the pharmaceutical industry and bringing new treatments to patients. In adopting evidentiary pluralism, the 21CCA implicitly sanctions a popular hypothesis on the causes of the crisis: it is partly due to the high amount of treatments that are lost to inadequate testing standards. Using different sources of evidence, regulatory agencies will be able to minimize these treatment losses. From a political standpoint, the question with this regulatory shift is whose interests it serves. If all evidentiary standards were equally reliable, the interests of the industry and of many patients might be aligned. But if we are right in our diagnosis, and we are left with uncertainty as to the comparability of RCT vs EHR tests, patients may have a dilemma. In the best case scenario, new tests would bring more cures to the market, but some of these may be ineffective or even harmful. Is it worth having more treatment options available even if not all of them are equally reliable?

These shifts point negotiate the core of liberal democratic polities: the move of the FDA further away from regulatory paternalism marks a retreat of the State from protecting its citizens from harm (through legal and bureaucratic devices such as regulatory activity).[10] A drug regulation framework stipulates what is the acceptable evidence of risk magnitude and risk structure (e.g., bias vs standard error) for granting approval of experimental treatments. Thus the 1938 and 1962 acts created an *ex ante* protection from harm. Before then, and since the last 150 years, only tort law was available (and still is) (Gibbs and Mackler 1987) – an *ex post* reparation for the injury caused by a compound. Following Agamben (1998), we can interpret both tort law and the FDA regulatory frameworks as core institutions of the liberal democratic polity: they protect one citizen from harm inflicted by another.

The protective power of the FDA mark of approval is complex. On the one hand, the prohibition of releasing and administering drugs that have not been tested is an example of how regulation anticipates potential harm and protects from it. On the other hand, when a drug is tested according to stipulated testing regimes and considered safe enough after evaluation of the supporting evidence, its approval by the FDA is voucher for the limited liability of the manufacturer for the harms that a pharmaceutical might still cause.[11] Harm eventually inflicted by the off-label use of

[10] The political metaphysics of the 21CCA shift in pharmaceutical regulation are stark. According to Agamben (1998), the *dispositif* of law is intended to demarcate a field of social relations that are protected from harm. Harm inflicted within the field will be punished through sanctions. However, a field cannot be demarcated without creating an opposing field outside protection. Key to the cultural and legal evolution of Western polities at least since Roman law has thus been the existence of a field of social relations that is outside the scope of law, and most importantly the continuous negotiation and redefinition of its boundaries. This is the field of *nuda vita.* Agamben argues that the existence of such a field is a necessary precondition of the social contract – complete protection of all the living would immobilize society. Note that here we are not taking a strange turn: this is the theme of the crisis of innovation, elaborated from a different perspective.

[11] The policy defines a case logic whereby the harm inflicted despite fulfilment of due process is not to be sanctioned in the same way that harm otherwise inflicted is culpable. Any changes in drug

a drug is consequently more easily sanctioned (and pharmaceutical companies do not promote such uses for this reason).

There are of course exceptions where the manufacturer can still be liable after drug approval. This can be for negligence (defect in design, testing, manufacturing or labelling); strict liability (injury caused by avoidable reasons: does not apply if "the product is properly prepared and accompanied by proper directions and warning"); and breach of contract (does not usually apply to drugs). Key here is the assumption by the legislator that drugs have *unavoidable risks*: perfect knowledge about them is impossible. To receive satisfaction, the plaintiff should then argue that the risks were avoidable for either improper design, manufacturing or labelling.

However, for the last five decades, the existence of FDA approval has ruled out improper design or testing as a litigation pathway. The courts rarely failed manufacturers for the harm caused by properly produced and labelled FDA-approved drugs. The FDA regulation has been until now authoritatively demarcating what epistemic risks (as implied by each accepted test design) will be treated as *unavoidable* and therefore not culpable.

A framework less centred on safety such as that the 21CCA is introducing, we argue, increases patient choice while shifting some of the harm involved in taking drugs from *ex ante* to *ex post* protection devices. Whether the 21CCA will then open a new space of litigation (thus undermining the evidentiary power of FDA approvals), or if instead it will simply mean that more harms will be unsanctionable (if American courts continue the practice of accepting FDA approval as evidence of test quality), it remains to be seen. With potentially inferior testing standards regulating access to market, it becomes possible that some harm is inflicted because of failure of the inferior safety standard and which could have been avoided by a superior standard. Until the reliability of the new standards is fully grasped, patients will have to suffer the eventual consequences of lesser State protection.

In the while, we expect that individual risk aversion will shape market outcomes: some patients (and their caretakers and doctors) will welcome uncertain but more abundant treatment choices; others will not. The attitudes of US citizens towards pharmaceutical risks changed throughout the twentieth century to support increasingly strict safety regulations, at least if we judge it by Congress decisions (Carpenter 2010). Is there a public demand for more cures offsetting this previous risk aversion? And is it a well-formed demand or does it rather reflect the marketing pressure of pharmaceutical lobbies, as critics contend?

In sum, the 21CCA paves the way for the regulatory use of EHR. We have argued that, before these data start their *journey* to the regulator's desk, it is crucial that we debate how to *package* EHR in order to make the best use of the information they provide. In designing the journey, one crucial point is how to convey the information about its potential limitations for regulatory use in a standardized format. Only

---

testing regulation are thus concerned with redefining the boundary between one type of harm risk (not culpable) and the other (culpable). Paternalistic policies following scandals such as Thalidomide expanded the field of protected life. With the 21CCA's prospect of easier approval to more drugs through inferior testing standards, the boundary between the two harms is moving again but in the opposite direction.

with some degree of package standardization, we can estimate the reliability of EHR in making regulatory decision, how often they yield to error, as compared to other sources of evidence. And this is the sort of information that a robust public sphere needs to debate whether the sort of evidentiary pluralism promoted by the 21CCA is welcome. If the journey of EHR data becomes so long as to require clinicians and patients to evaluate the evidence in favour of a treatment option, it might be travelling that eventually can go too far.

Acknowledgements Tempini's work was supported by ERC grant award 335925 (DATA_SCIENCE), and by EPSRC grant EP/N510129/1.

References

Adler-Milstein, Julia, Catherine DesRoches, Michael Furukawa, et al. 2014. More Than Half of US Hospitals Have at Least a Basic EHR, But Stage 2 Criteria Remain Challenging for Most. *Health Aff (Millwood)* 33: 1664–1671. https://doi.org/10.1377/hlthaff.2014.0453.

Agamben, Giorgio. 1998. *Homo Sacer*. Stanford, CA: Stanford University Press.

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.

Avorn, Jerry., and Aaaron S. Kesselheim. 2015. The 21st Century Cures Act — Will It Take Us Back in Time? *New England Journal of Medicine* 372: 2473–2475. https://doi.org/10.1056/NEJMp1506964.

Bell, Jennifer A., and Lynda G. Balneaves. 2015. Cancer Patient Decision Making Related to Clinical Trial Participation: An Integrative Review with Implications for Patients' Relational Autonomy. *Support Care Cancer* 23: 1169–1196. https://doi.org/10.1007/s00520-014-2581-9.

Boyd, danah m., and Kate Crawford. 2012. Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15: 662–679.

Cambrosio, Alberto, Peter Keating, Thomas Schlich, and George Weisz. 2006. Regulatory Objectivity and the Generation and Management of Evidence in Medicine. *Social Science & Medicine* 63: 189–199. https://doi.org/10.1016/j.socscimed.2005.12.007.

Carpenter, Daniel P. 2010. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA*. Princeton: Princeton University Press.

Carrieri, Daniele, Fabio A. Peccatori, and Giovanni Boniolo. 2018. The Ethical Plausibility of the 'Right To Try' Laws. *Critical Reviews in Oncology/Hematology* 122: 64–71. https://doi.org/10.1016/j.critrevonc.2017.12.014.

Center for Drug Evaluation and Research (CDER). 2018. Use of Electronic Health Record Data in Clinical Investigations. *FDA*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-electronic-health-record-data-clinical-investigations-guidance-industry. Accessed 14 Sept 2019.

Center for Drug Evaluation and Research (CDER) & Center for Biologics Evaluation and Research (CBER). 2013. Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data. *FDA*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/best-practices-conducting-and-reporting-pharmacoepidemiologic-safety-studies-using-electronic. Accessed 14 Sept 2019.

Demir, Ipek, and Madeleine J. Murtagh. 2013. Data Sharing Across Biobanks: Epistemic Values, Data Mutability and Data Incommensurability. *New Genetics and Society* 32: 350–365. https://doi.org/10.1080/14636778.2013.846582.

DiMasi, Joseph A., Henry G. Grabowski, and Ronald W. Hansen. 2016. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *Journal of Health Economics* 47: 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

Dixon, Brian. E., P. Joseph. Gibson, Karen. F. Comer et al. 2015. *Measuring Population Health Using Electronic Health Records: Exploring Biases and Representativeness in a Community Health Information Exchange.* Paper presented at the 15th World Congress on Health and Biomedical Informatics, MEDINFO 2015.

Dumit, Joseph. 2012. *Drugs for life: how pharmaceutical companies define our health*. Durham, NC: Duke University Press.

Epstein, Steven. 1996. *Impure Science. Aids and the Politics of Knowledge*. Berkeley/Los Angeles: University of California Press.

EvaluatePharma. 2017. *World Preview 2017, Outlook to 2022*. Evaluate. http://info.evaluategroup.com/rs/607-YGS-364/images/WP17.pdf. Accessed 14 Sept 2019.

FDA. 2013. *Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data*. Washington: U.S. Department of Health and Human Services.

———. 2016. Use of Electronic Health Record Data in Clinical Investigations. In *Draft Guidance*. Washington: U. S. Department of Health and Human Services.

Fisher, Jill A. 2009. *Medical Research for Hire: The Political Economy of Pharmaceutical Clinical Trials*. New Brunswick, NJ: Rutgers University Press.

Fleming, Lora E., Andy Haines, Brian Golding, et al. 2014. Data Mashups: Potential Contribution to Decision Support on Climate Change and Health. *International Journal of Environmental Research and Public Health* 11: 1725–1746. https://doi.org/10.3390/ijerph110201725.

Food and Drug Administration (FDA). 2018. FDA's Sentinel Initiative. FDA. https://www.fda.gov/safety/fdas-sentinel-initiative. Accessed 14 Sept 2019.

Ford, David V., Kerina H. Jones, Jean-Philippe Verplancke, et al. 2009. The SAIL Databank: Building a National Architecture for E-Health Research and Evaluation. *BMC Health Services Research* 9: 157. https://doi.org/10.1186/1472-6963-9-157.

Gaudillière, Jean Paul, and Volker Hess. 2012. *Ways of Regulating Drugs in the 19th and 20th Centuries*. Basingstoke, UK: Palgrave Macmillan.

Gibbs, Jeffrey N., and Bruce F. Mackler. 1987. Food and Drug Administration Regulation and Products Liability: Strong Sword, Weak Shield. *Tort & Insurance Law Journal* 22: 194–243.

Glasziou, Paul, Iain Chalmers, Michael Rawlins, et al. 2007. When Are Randomised Trials Unnecessary? Picking Signal from Noise. *The BMJ* 334: 349–351. https://doi.org/10.1136/bmj.39070.527986.68.

Gonsalves, Gregg, Daniel P. Carpenter, and Joseph S. Ross (Producer). 2016. Lawmakers Must Ask Tough Questions About the 21st Century Cures Act. *The Hill*. http://thehill.com/blogs/congress-blog/healthcare/307020-lawmakers-must-ask-tough-questions-about-the-21st-century. Accessed 14 Sept 2019.

Gonzalez-Moreno, Maria, Cristian Saborido, and David Teira. 2015. Disease-Mongering Through Clinical Trials. *Studies in History and Philosophy of Science* 51: 11–18. https://doi.org/10.1016/j.shpsc.2015.02.007.

Gøtzsche, Peter C. 2013. *Deadly Medicines and Organised Crime: How Big Pharma Has Corrupted Healthcare*. London: Radcliffe Publishing.

Helgesson, Claes-Frederick. 2010. From Dirty Data to Credible Scientific Evidence: Some Practices Used to Clean Data in Large Randomised Clinical Trials. In *Medical Proofs, Social Experiments: Clinical Trials in Shifting Contexts*, ed. C. Will and T. Moreira, 49–64. Farnham: Ashgate.

Higgins, Julian, Douglas.G. Altman, Peter. Gøtzsche, et al. 2011. The Cochrane Collaboration's Tool for Assessing Risk of Bias in Randomised Trials. *The BMJ* 343. https://doi.org/10.1136/bmj.d5928.

Hripcsak, G., and D.J. Albers. 2013. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 20 (1): 117–121.

Hripcsak, George., Patrick B. Ryan, John D. Duke, et al. 2016. Characterizing Treatment Pathways at Scale Using the Ohdsi Network. *Proceedings of the National Academy of Sciences* 113: 7329–7336. https://doi.org/10.1073/pnas.1510502113.

Institute of Medicine. 2014. *Drug Repurposing and Repositioning: Workshop Summary*. Washington, DC: The National Academies Press.

Jorm, L. 2015. Routinely collected data as a strategic resource for research: Priorities for methods and workforce. *Public Health Research and Practice* 25 (4): e2541540.

Kesselheim, Aaron S., and Jerry Avorn. 2016. Approving a Problematic Muscular Dystrophy Drug: Implications for FDA Policy. *JAMA* 316: 2357–2358. https://doi.org/10.1001/jama.2016.16437.

———. 2017. New "21st Century Cures" Legislation: Speed and Ease vs Science. *JAMA* 317: 581–582. https://doi.org/10.1001/jama.2016.20640.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago/London: The University of Chicago Press.

Leonelli, Sabina. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Lupkin, Steven, and Sydney Findlay, (Producer). 2016. Grab Bag of Goodies In 21st Century Cures Act. *Kaiser Health News*. http://khn.org/news/grab-bag-of-goodies-in-21st-century-cures-act/. Accessed 14 Sept 2019.

Lyons, Ronan A., Kerina H. Jones, Gareth John, et al. 2009. The SAIL Databank: Linking Multiple Health and Social Care Datasets. *BMC Medical Informatics and Decision Making* 9: 3. https://doi.org/10.1186/1472-6947-9-3.

Marks, Harry M. 1997. *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900–1990*. Cambridge/New York: Cambridge University Press.

Nagendran, Myura, Tiago V. Pereira, Grace Kiew, et al. 2016. Very Large Treatment Effects in Randomised Trials as an Empirical Marker to Indicate Whether Subsequent Trials Are Necessary: Meta-Epidemiological Assessment. *The BMJ* 355: i5432. https://doi.org/10.1136/bmj.i5432.

Pease, Alison M., Harlan M. Krumholz, Nicholas S. Downing, et al. 2017. Postapproval Studies of Drugs Initially Approved by the FDA on the Basis of Limited Evidence: Systematic Review. *The BMJ* 357: j1680. https://doi.org/10.1136/bmj.j1680.

Pivovarov, Rimma., David J. Albers, Jorge L. Sepulveda, et al. 2014. Identifying and Mitigating Biases in EHR Laboratory Tests. *Journal of Biomedical Informatics* 0: 24–34. https://doi.org/10.1016/j.jbi.2014.03.016.

Podolsky, Scott H. 2015. *The Antibiotic Era: Reform, Resistance, and the Pursuit of a Rational Therapeutics*. Baltimore: Johns Hopkins University Press.

Powell-Smith, Anna, and Ben. Goldacre. 2016. The TrialsTracker: Automated Ongoing Monitoring of Failure to Share Clinical Trial Results by all Major Companies and Research Institutions [Version 1; Peer Review: 2 Approved]. *F1000Research* 5:2629.

Power, Michael. 1997. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

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.

Reiss, Julian. 2017. Causality and Causal Inference in Medicine. In *The Routledge Companion to Philosophy of Medicine*, ed. Miriam Solomon, Jeremy R. Simon, and Harold Kincaid, 58–70. London: Routledge.

Rusanov, Alexander, Nicole G. Weiskopf, Shuang Wang, et al. 2014. Hidden in Plain Sight: Bias Towards Sick Patients When Sampling Patients with Sufficient Electronic Health Record Data for Research. *BMC Medical Informatics and Decision Making* 14: 51. https://doi.org/10.1186/1472-6947-14-51.

Senn, Stephen. 2008. Lessons from TGN1412 and TARGET: Implications for Observational Studies and Meta-Analysis. *Pharmaceutical Statistics* 7: 294–301. https://doi.org/10.1002/pst.322.

———. 2013. Seven Myths of Randomisation in Clinical Trials. *Statistics in Medicine* 32: 1439–1450. https://doi.org/10.1002/sim.5713.

Sherman, Rachel E., Steven A. Anderson, Gerald J. Dal Pan, et al. 2016. Real-World Evidence — What Is It and What Can It Tell Us? *New England Journal of Medicine* 375: 2293–2297. https://doi.org/10.1056/NEJMsb1609216.

Teira, David. 2013. A Contractarian Solution to the Experimenter's Regress. *Philosophy of Science* 80: 709–720.

———. Forthcoming. On the Normative Foundations of Pharmaceutical Regulation. In *Philosophy of Pharmacology*, ed. A. LaCaze and B. Osimani. Dordrecht: Springer.

Tempini, Niccolò. this volume. 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, and Sabina Leonelli. 2018. Concealment and Discovery: The Role of Information Security in Biomedical Data Re-Use. *Social Studies of Science* 48: 663–690.

Vawdrey, David K., and George Hripcsak. 2013. Publication Bias in Clinical Trials of Electronic Health Records. *Journal of Biomedical Informatics* 46: 139–141. https://doi.org/10.1016/j.jbi.2012.08.007.

Wardell, William M., and Louis Lasagna. 1975. *Regulation and Drug Development*. Washington: American Enterprise Institute for Public Policy Research.

Wikipedia. 2018. Kaiser Permanente. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Kaiser_Permanente&oldid=828279882 Accessed 14 Sept 2019.

Niccolò Tempini is Senior Lecturer in Data Studies at the University of Exeter, Department of Sociology, Philosophy and Anthropology, and a Turing Fellow at the Alan Turing Institute. He is an interdisciplinary social scientist interested in questions of information, data, technology, organization, value and knowledge. He researches Big Data research and digital infrastructures, investigating the specific knowledge production economies, organization forms and data management innovations that these projects engender with a focus in their social and epistemic consequences. He studies the practices of data scientists, software developers, researchers and nonprofessionalised experts to understand how different forms of knowledge and value intersect with each other when different actors come to grips with new methods and new forms of data, information technology and organization. His research has been published in international journals across science and technology studies, information systems, sociology and philosophy (more information at www.tempini.info).

David Teira is Professor at the Department of Logic, History and Philosophy of Science at UNED, the Spanish Open University. He has worked on the uses of statistical methods in economics and medicine. In particular, he has studied how the different stakeholders in a clinical trial can agree on an experimental outcome, despite their conflicts of interests (http://www.uned.es/personal/dteira).

Part IV
Interlude

Most Often, What Is Transmitted Is Transformed

Theodore M. Porter

Abstract This short interlude prompts reflection on the transformations involved in data mobilization through a vivid discussion of the changing circumstances of the visualization of data about family histories of mental illness – and their interpretation in relation to questions around inheritability and underpinning biological causes – through graphs and tables produced between the mid-nineteenth century and the early twentieth century.

Henri Legrand du Saulle used the title phrase above to encapsulate his teacher B. A. Morel's doctrine of hereditary degeneration. Degenerative heredity was defined not by stable transmission of traits from generation to generation, but as a trajectory of decline leading often to extinction of the line (Legrand du Saulle 1873, 9). This theory was a hit with doctors, novelists, and other authorities on human heredity for about half a century. Its fall from favor cannot be attributed to any shortage of data. It was not easy, however, to reach agreement as to what the data meant. One notable collection of family records that came to be cited in support of Morel's theory had been published in 1859 by a Norwegian asylum doctor and researcher, Ludvig Dahl. His tables of mental illness were redrawn and republished half a century later by English biometricians, then relabeled as evidence of Mendelian degeneration for a German health exposition. In each case, Dahl's data was assigned new meanings. Often, when data travels, it will be transformed (Porter 2018, 131–142 and 179).

Dahl created a partly novel visual technology, the pedigree table, to convey his understanding of pathological inheritance. Although he admitted variability in the manifestation of hereditary elements for mental disease, he regarded a close resemblance between parent and child as the most compelling indication of inheritance. His book on the subject attracted immediate attention across northern Europe for its insights on the causes and transmission of mental illness. Although he wrote in a somewhat inaccessible language, Danish/Norwegian, he attracted knowledgeable commentators in French, English, and especially German. They did not need to be convinced that heredity was key to the perpetuation of insanity, and likely its most

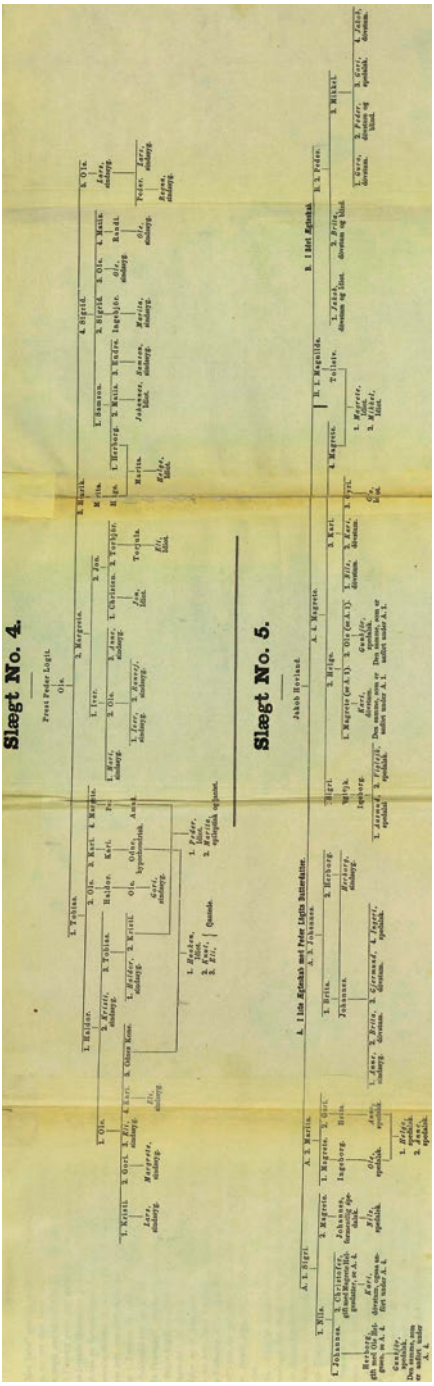
T. M. Porter (✉)
Department of History, University of California, Los Angeles, Los Angeles, CA, USA
e-mail: tporter@history.ucla.edu

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_12

fundamental cause. Cascades of annual asylum reports from Europe and North America included lists of the "presumed causes" of illness for newly-admitted patients, and heredity was consistently at or near the top. Dahl's book was distinctive for its fine-grained studies at a local level, typically of a single parish, which he compared with numbers from the much-admired Norwegian decennial census of insanity, first taken from 1825 to 1828, and with data from the new asylum of Gaustad in the hills above Christiania (Oslo). Records of the 1855 census gave him access to the unpublished names of individuals reported as insane, which he supplemented by surveying parishes of interest, talking with doctors, priests and families of these unfortunate souls. This information, the basis for his kinship tables, pertained specifically to the question of hereditary transmission.

Dahl, a data guy, resisted the temptations of dogmatism. Often, children had cases very similar to their parents, but not always. He had enough examples of disparate forms of illness within a single family to declare with some assurance that a hereditary *Anlæg* or factor could have diverse manifestations. Insanity he understood as an "acquired" condition, distinguished by its invisibility until late adolescence. Madness was easily distinguished from idiocy, or mental weakness, which was typically congenital and often appeared in association with bodily deformities, especially of the cranium. Yet he turned up many families manifesting both conditions. Idiocy, in turn, was not only linked by heredity to deaf-mutism, but often appeared alongside it in the same individual. Dahl also mentioned albinism and even leprosy, a relatively common and much-studied condition in Norway, as other afflictions that were often allied to idiocy. The proliferation of mental and physical defects seemed to be more common where there was intermarriage, especially if a hereditary factor was present in the family (Dahl 1859, 82–86). In a section on hereditary causes, Dahl printed eight pedigrees of kin groups showing a high level of inherited illness. The most extensive of these came from the parish of Kinservik on the Hardanger fjord, east of Bergen, where the inhabitants (he said) were especially attentive to the memory of their ancestry and where the priest zealously aided the research. Despite using two foldout pages, Dahl had to divide this extended family into two charts, kin groups 4 and 5. They revealed a variety of conditions that seemed to be joined together by heredity, including deaf-mutism, epilepsy, leprosy, blindness and albinism as well as insanity and idiocy (Dahl 1859, tables 4–5 and pp. 82–86). To this extent his tables resembled those of Morel's students, but Dahl found no directional tendency. His tables also documented intermarriage of close relatives, which, he speculated, may strengthen a hereditary tendency, but he wanted more data to be confident.

A German commentator and translator expressed puzzlement that an *Anlæg* (in German, *Anlage*) could be expressed in such heterogeneous forms, sometimes even without cousin marriages. Such instability of types of insanity was an old story, and not only as heredity. A patient admitted to an asylum with a diagnosis based on one set of symptoms might have to be assigned another when these manifestations changed. The boundary between madness and idiocy, in contrast, was mostly reliable, and neither of these could be confused with albinism or leprosy (von dem Busch 1861, 483–485).


Slægt No. 4.
Slægt No. 5.

Although the work was enthusiastically received, Dahl's tables did not at first inspire imitators. In 1877, the alienist William Ireland translated a few of them into English for his book on mental defect (Ireland 1877, tables at back). But it was not until the new century that pedigree tables emerged as the indispensable tool for documenting inherited defect. The most important site of their reappearance was in the *Treasury of Human Inheritance*, a reference work funded by Francis Galton's Eugenics Laboratory, now under the direction Karl Pearson. The conditions documented in the initial fascicules, issued in 1909, included some anatomical abnormalities that could be described very precisely. For conditions such as mental illness and tuberculosis, however, Pearson and his coworkers preferred to speak of diathesis or constitutional susceptibility, to be identified from readily-apparent symptoms. Pearson was, after all, a statistician, not a doctor, and he was in no position impose any system of classification on such a slippery subject. Also, since these maladies were not often identified before age 20 or 25, it might well be impossible to examine ancestors beyond a single generation. Dahl had relied on written records and family recollections to compile his kinship tables.

Pearson, who stressed the painstaking labor of checking and rechecking required to assemble even one solid table of this kind, treated "Dahl's case" as having met this high standard of quality. That meant they were fit to serve as a data resource, to be compared and analyzed in pursuit of scientific conclusions on the transmission of human defects. He indicated provenance but did not call attention to singularities, and he printed tables of multiple families by multiple researchers on the same page. There is no discussion of the sites of research, and individual names were omitted. By redrawing all pedigrees in a common format, he has made them almost interchangeable. Pearson's formidable erudition included a working knowledge of the Norwegian language, which he had studied in order to read Ibsen in the original, so it is quite possible that he had a hand in the excavation of Dahl's data. Yet there is not a word here about Dahl, his site, or his methods. Pearson's ambition was to create a database of interchangeable data, one that did not require researchers to go back into the sources. He also did not use this work to defend hereditary theories or to take shots at Mendelian reductions of complex traits and behaviors. Rather, he sold his numbers as independent of all theories (Pearson 1912 [these sheets first printed 1909] plate 10; Porter 2004).

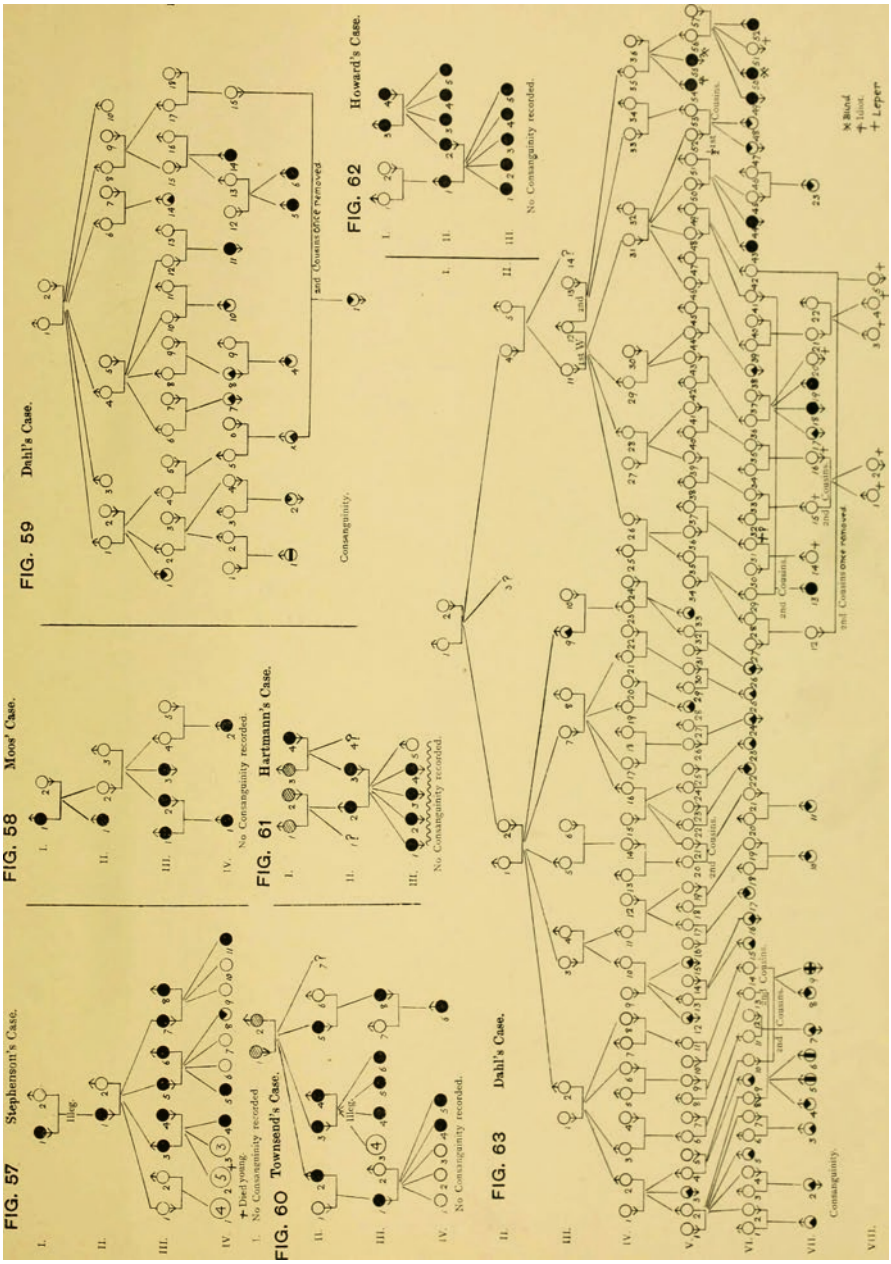


FIG. 57 Stephenson's Case.
FIG. 58 Moss' Case.
FIG. 59 Dahl's Case.
FIG. 60 Townsend's Case.
FIG. 61 Hartmann's Case.
FIG. 62 Howard's Case.
FIG. 63 Dahl's Case.
No Consanguinity recorded.
Consanguinity.
Blood
Idiot
Leper

In one important sense, his effort to create a neutral database of heredity was a success. Dahl's kinship tables from Hardanger, as redrawn for the *Treasury of Human Inheritance*, gained instant recognition, and were featured, for example at the 1911 International Hygiene Exposition in Dresden. The German organizers and editors, Max von Grüber and Ernst Rüdin, simply reprinted the redrawn Norwegian material alongside other graphs and tables of inherited mental and nervous conditions, without even translating loose English words on these charts. As data they took Pearson's tables to be authoritative (See Nikolow 2001 on the Dresden Exhibition).

The treatment in the catalogue of Dahl's case (*Fall von Dahl*), however, involves some perplexing little oddities, and was anything but atheoretical. Grüber and Rüdin put to the side what Dahl had understood as the most remarkable feature of this chart, the juxtaposition of so many conditions there. The German heading for the Hardanger table reduced its multifarious defects to just one, deaf-mutism (*Taubstummheit*). Clearly they were looking for a striking example of Mendelian neurological inheritance. This fixation on Mendelian inheritance, nonetheless, was perfectly compatible with a relentlessly statistical presentation. And this was not all. In the course of the work, they came to be tantalized by the hopes of demonstrating Morel's mechanism of hereditary degeneration.

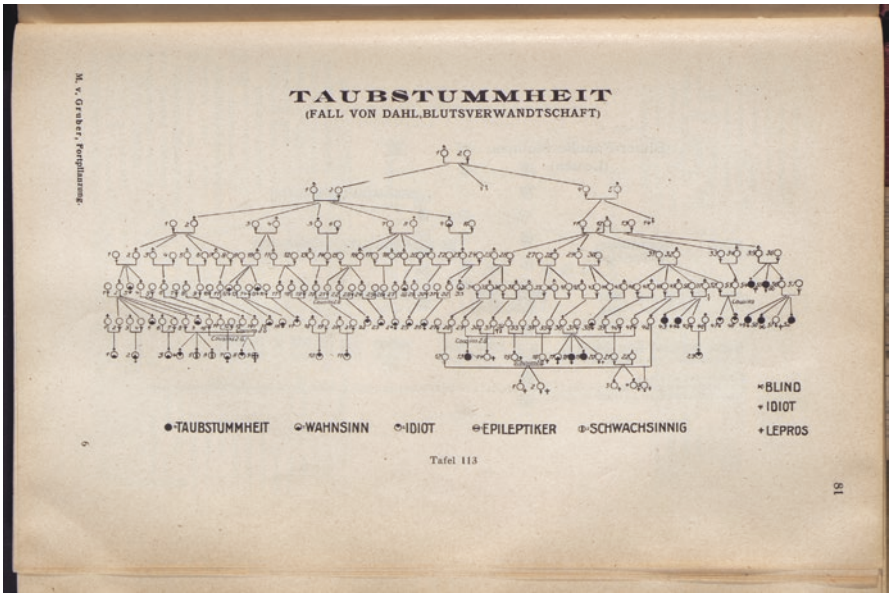
The printed record, consisting of two editions of the exhibition catalogue (both dated 1911), and a crowd of inconsistencies gives evidence of momentary thoughts and dreams, rushed into print and then disappearing into smoke. In the first catalogue, for example, the authors describe the crucial Table 113 as omitted just a few lines before it appears (von Grüber and Rüdin 1911a, 73). The pages instance hemophilia, congenital night blindness, and brachydactyly (shortened fingers) as known to be inherited independently and to segregate (*mendeln)*, and indeed to be governed by a single genetic unit (*Erbeinheit*) or gene (*Gene*). The catalogue next refers back to "Table 112 Dahl's case on deaf-mutism," here described displaying a remarkable, simultaneous appearance of deaf-mutism and insanity in distant relatives in the fifth generation. The crucial point here is that the third and fourth generations were "practically free" of these conditions. More mistakes: 112, though from Dahl, was a different table, also copied from Pearson's *Treasury*. It showed no such eruption of hereditary illness. Table 113, they now declare, referring to the important table they claimed earlier to have omitted, "is entirely similar." But their topic here was degeneration, whereas Table 113 concerned Mendelism.

They come finally to the most astonishing result of all, the reconciliation of Mendel and Morel. The catalogue text veers back to speculate that Dahl's kinship table of deaf-mutism might supply a concrete instance of hereditary degeneration. In the next line they tried out a fusion of theories, Mendelism and degeneration. "Supposing the information (*Angaben*) in the kinship tables is complete in this respect, it gives the impression that an abnormal gene or an abnormal combination of genes from the shared heritage of the progenitors has at last attained so great a

degree of degeneration that manifest derangements can occur." The degenerative force, they were suggesting, must have been intensified by family relationships and shared heredity – that is, cousin marriages. What else could explain the simultaneous appearance of a new irregularity in distinct lines of this kin group (von Grüber and Rüdin 1911a, 71–77, quotes 76–77)?

So many inconsistencies seem to reflect a momentary but irrepressible excitement regarding a putative demonstration from Dahl's data of Morel-style degeneration. In the "enlarged and completed" edition issued later the same year, the mistakes in the identification and numbering of tables were rectified. Rüdin, the psychiatrist, who presumably was responsible for this material, hints now at doubts as to the evidence for Morel-type degeneration by inserting a question mark: "Supposing the information… is complete (?)" Complete information on generations long dead may be depicted as a tree, but it does not grow on trees. Both editions, however, include an example of polydactyly (extra finger or toes) as an instance of the intensification of heredity, a tendency that Dahl, too, had endorsed. Certainly the authors did not rule out degeneration. This passage concludes by calling for more information (or data), that is more family trees of inherited illness (von Grüber and Rüdin 1911b, 75, 78, 81).

The investigation of madness and heredity was, by 1859, a recognized and even exemplary focus of data production. The hope that this data could be consolidated into databases of ever greater scale, to be analyzed in offices and exhibited in museums, burned brightly in those years, as it does in our own. But the detachment of data from the concrete conditions of its production is always risky. Data, as it moves, is most often thinned, and what is thinned is necessarily transformed.



References

Dahl, Ludvig. 1859. *Bidrag til Kundskab om de Sindssyge i Norge*. Christiania: Det Steenske Bogtrykkeri.

Ireland, William W. 1877. *Idiocy and Imbecility*. London: J & A Churchill.

Legrand du Saulle, Henri. 1873. *La folie héréditaire*. Paris: Adrien Delahaye.

Nikolow, Sybilla. 2001. Der statistische Blick auf Krankheit und Gesundheit. In *Infografiken, Medien, Normalisierung: Zur Kartografie politisch-sozialer Landschaften*, ed. Ute Gehrheit, Jürgen Link, and Ernst Schulte-Holtey, 223–241. Heidelberg: Synchron Wiss.-Verl. der Autoren.

Pearson, Karl, ed. 1912. *Treasury of Human Inheritance*, Vol. 1. London: Cambridge University Press, plate 10; this section was first printed in 1909.

Porter, Theodore M. 2004. *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton: Princeton University Press.

———. 2018. *Genetics in the Madhouse: The Unknown History of Human Heredity*. Princeton: Princeton University Press.

von dem Busch, Gerhard. 1861. Review of Dahl, *Bidrag*. *Allgemeine Zeitschrift für Psychiatrie* 18: 474–518.

von Grüber, Max, and Ernst Rüdin, eds. 1911a. *Fortpflanzung, Vererbung, Rassenhygiene: Katalog der Gruppe Rasenhygiene der Internationalen Hygiene-Ausstellung*, 1st ed. Munich: J. F. Lehmanns Verlag.

———, eds., 1911b. *Fortpflanzung, Vererbung, Rassenhygiene: Illustrierter Fuhrer durch die Gruppe Rasenhygiene der Internationalen Hygiene-Ausstellung*, zweite ergänzte und verbesserte Auflage. Munich: J. F. Lehmanns Verlag.

Theodore M. Porter is Distinguished Professor of History at the University of California, Los Angeles. Much of his research has involved the history of statistics, quantification, calculation and data, often in relation to the human sciences. His most recent book, *Genetics in the Madhouse: The Unknown History of Human Heredity*, appeared in Princeton University Press in 2018. His other books include *Karl Pearson: The Scientific Life in a Statistical Age* (Princeton University Press, 2004), *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton University Press, 1995) and *The Rise of Statistical Thinking* (Princeton University Press, 1986). He also credited with Dorothy Ross *The Cambridge History of Science, Volume 7: Modern Social Sciences* (Cambridge University Press, 2003).

Part V
Interpreting: Data Transformation, Analysis and Reuse

The Reuse of Digital Computer Data: Transformation, Recombination and Generation of *Data Mixes* in Big Data Science

Niccolò Tempini

Abstract This chapter is concerned with the relationship between the materiality of digital computer data and their reuse in scientific practice. It builds on the case study of a 'data mash-up' infrastructure for research with environmental, weather and population health data. I problematise the extent to which scientists reusing digital computer data heavily manipulate the sources through complex and situated calculative operations, as they attempt to re-situate data well beyond the epistemic community in which they originated, and adapt them to different theoretical frameworks, methods and evidential standards. The chapter interrogates the consequent relationship between *derivative* data and the data sources from which they originate. The deep relationality of *scientific computer data* is multi-layered and scaffolded, as it depends on relations between various kinds of data, computing technologies, assumptions, theoretical scaffoldings, hypotheses and other features of the situation at hand.

1 Introduction

This chapter is concerned with the relationship between the materiality of digital computer data and their reuse in scientific practice. It builds on the case study of the Medical and Environmental Data Mash-up Infrastructure, a project born at the interdisciplinary crossroads between environmental and weather sciences and population health research. Studying the practices of development and use and the operational characteristics of the infrastructure, I aim to show the extent to which scientists reusing digital computer data proceed to heavy manipulation of the *sources* through complex, intermediated and situated calculative operations. Consequently, this chapter interrogates the relationship between *derivative* data and

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

Alan Turing Institute, London, UK
e-mail: n.tempini@exeter.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_13

the data *sources* from which they originate. It argues that systematic transformation and recombination of both data source values and structures, involved in the reuse of computer data, lead to the creation of deeply derivative data that are best considered new digital and epistemic objects.[1]

This is important to advance our understanding of the journeys of computer data, and especially so since much novelty of big data innovations seems to hang on successfully repurposing a great variety of digital traces that must be available in great quantity. Indeed, any initial assessment of what is happening with the advent of huge digital infrastructures that warp our understanding of notions such as scale, size, speed and boundary of information begs the question, *in what way does digital materiality make a difference for scientific practice, and what is the purchase of an account of data practices that is specific about digital data?* The chapter builds on empirical material gathered through participant observation, first-person involvement in data science exercises, and insights from literature in information science, media studies and the philosophy of technology and of science. The aim is to offer an original angle for data and data reuse theorisation, one that more deeply considers the specific characteristics of digital technologies while attending to the epistemic practices of human actors at the same time.

The topic has started to surface in the philosophy and sociology of science literature interested in digital data, but has not raised sufficient attention. Thanks to an increasing interest in empirically attending to scientific practices, philosophers of science and STS scholars have started to ask questions of definition, character and materiality of data that were once absent from the debate. Discussions relating questions of materiality to the epistemic and social role of data have necessarily featured in the debate (Rheinberger 2010), and feature in this volume accordingly (Halfmann this volume; Wylie this volume). For instance, starting from the study of data practices in archaeology, Wylie argued that the materiality of an object is crucial in shaping the ways it can serve as data (Chapman and Wylie 2016; Wylie 2017). Observing how scientists can return several times to the same object in order both to challenge and to reaffirm hypotheses, and to discover new lines of interpretation, she shows that over time the "same" object can be mobilised to serve completely different lines of argument. Objects can take new roles because their specific materiality can confer to them a persistent, residual character that is not fully exhausted by their mobilisation in previous lines of inquiry.

Focusing on data sharing through online databases and its impact in the practices and culture of biology, Leonelli (2016) develops a *relational* definition of data from a pragmatist perspective. She holds that in the first place, what counts as data depends on *situated* evaluations. Data can be any object that can be used in support

[1] In this volume, Parker discusses the case of "data products" in climate science: data that are manipulated by third parties from data sources. She highlights how different methods for manipulating data sources create completely different data products that retain a "potential structural uncertainty". She also highlights how data products have a role of social intermediation: they are mobilised on a new ground (the heated political arena of climate change), outside the institutional boundaries within which data sources are used (Parker this volume).

of evidential claims at specific moments of the scientific inquiry. A number of conditions shape an object's potential to be used as data, which include material issues. Key to ensuring reuse are what she calls "packaging strategies": the activities aimed at preparing the data for de-contextualisation, transfer, and re-contextualisation in the new situation of use. Leonelli points out that packaging frequently intervenes on the material characteristics of the data and "often change format, medium and shape of the data" (Leonelli 2016:76); consequently, "biological data are anything but stable objects" (ibid.). An example are sequencing data: these can come in different formats, which might or not be compatible with the machinery employed downstream in the journey. Data formats change as "data start their journeys across screens, printouts and databases around the world" (2016:84).[2] She argues that the identity of data can be traced throughout and despite these material discontinuities if one focuses on the association between "researchers' perceptions of what counts as data and the type and stage of inquiry in which such perceptions emerge" (2016:77).[3]

1.1 Scientific Data vs Computer Data

The argument of this chapter starts from a juxtaposition between the meaning of data as in *scientific* data, and data as in *computer* data. This is to demonstrate, as I have already anticipated, that the use of big data in science *depends on successful strategies of computation and transformation of digital data qua computational objects*. The data journey is underpinned by a rather continuous and tightly interlocked chain of custody granted by technical operations on digital equipment. These are complex manipulations that selectively transform symbolic values at the level of specific fields or portions of the semantic content, while leaving other components

[2] It is relevant to point out the standpoint of Leonelli's analysis. Focusing on the practices of scientists, she observes that scientists work with all kinds of object with no stable or predilected feature to be discerned. From this perspective she elaborates a 'general' philosophy of science framework that aims to apply both to practices with digital objects, as with any other object used by scientists as data.

[3] The theory of data travel is grounded with two further conditions. First is that to assign, to two materially different objects, same identity as data should be a criterion of *epistemic function continuity*. If despite (or rather thanks to) the changes to their "format, media and shape," data objects keep an identity as objects that can be used for knowledge claims, the travelling continues. The specific function will change depending on the situation of use, but continuity has to be of the 'dataness' of the object: whether something can endure these shifts and still be used by somebody as data. The second condition immediately follows from the first. It deals with the problem of how to account for the relationship between "type" (the semantically unique) and "token" (the material instantiation), when data are translated multiple times over various formats and media. Leonelli questions altogether the usefulness of this distinction for understanding data journeys: even the 'same' data change meaning with a change in situation (can be interpreted differently in different situations), so we will often lack a strong grounding for an identity of the 'original' in the first place.

untouched. They allow the repurposing of data sources that were not designed for travel and reuse.[4]

In thinking about the relationship between the scientific use of big data and the computational transformations that they undergo, I want to take the opportunity to open the data object blackbox. As it has been duly noted (Leonelli 2016), changes in format and media can often disrupt the reuse process; but from a computer data perspective, these might amount to as little as a change in file headers (a form of machine-readable metadata). The data reuse operations I am interested in run deeper, to the heart of a digital object, and can completely undo its semantic fabric.

To avoid any hesitation in linking specific material characteristics of data objects to their epistemic roles, it is useful to recognise the specific angle that philosophers and science studies scholars often take on the category of data, and the theoretical assumptions and goals that inform it. According to Leonelli's account of data, the status of an object as data depends on situated evaluations by actors relative to goals, expectations, resources and background theories. Consequently, in this chapter I will use the term *scientific data* to refer to objects that are held to satisfy the following key requirements:

– The object has *epistemic value* because a social actor considers it to be usable to stake a claim about the world (Leonelli stresses this value is *evidential* – 2016).
– *Scientific practice determines its data status*: does it satisfy the needs of a specific situation of inquiry?
– Relational objecthood: the object can change materially yet retain data status if above conditions are granted – *it continues to be usable in scientific inquiry.*

The data word has a number of other uses. Mind-numbing advances in computational and networking technologies have left no domain of social life untouched. In studies primarily concerned with the impact of computing technologies on social process and culture the term data is often used to refer to the digital records stored on a computing machine,[5] the existence of which is a precondition to the everyday operation of digital systems. In this perhaps most common use of the word, data are digital objects at the centre of *socio-technical practices of computation*. Accordingly, in this chapter I use the term *computer data* to refer to digital objects that are held to satisfy the following key requirements:

– The digital object is an object described through binary numbers and which can be accordingly manipulated through mathematical functions, as commonly embedded through software in programmable computer machines.

[4] A case in point are routine data generated through encounters at the points of care within the health system (see Tempini and Teira this volume), but also, as the case I present illustrates, weather and environmental data. As Parker's chapter in this volume (Parker this volume) also shows, this kind of operations are often carried out by 3rd parties to the original data producers.

[5] Hui (2012) makes a somewhat similar point, while juxtaposing data as "given" to data as "trasmittable information."

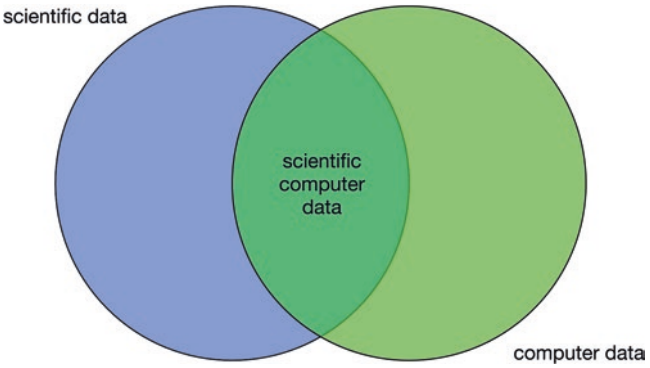


Fig. 1 Scientific data, computer data, and scientific computer data

- The object has *cognitive value* because it can be used to access or generate new information through computation, and is a required resource for the functioning of digital systems.
- *Socio-technical practice of computation determines its data status:* is it computable? Is it integrated in a *technological milieu* such that it is interacted with in a way that is socio-technically meaningful?
- Relational objecthood: the object can change materially at the level of its symbolic constitution yet retain data status if above conditions are granted (it continues to be usable in socio-technical practices of computation).

Note how the first of these requirements is a specific material condition. Note also how the two terms of scientific vs computer data are broadly parallel, often but not necessarily overlapping, and the second is narrower than the first. By and large, existing accounts of data have so far neglected this juxtaposition, working instead from the single standpoint offered by either of the two meanings.[6] Others did worse and conflated them.[7] I would like instead to stay as long as possible in the uncomfortable zone where objects could be (both, either, or neither) computer and/or scientific data (Fig. 1).

A host of research questions arise as we try to hold this juxtaposition alive. For instance, one may ask: *Computational socio-technical practices can generate new objects easily from existing digital material, but will they be epistemic objects? How*

[6] Hence confusion ensues with everyday use of both meanings of data. People can complain 'my data are lost!' after a virus wiped out indiscriminate portions of their disk or file system; but they can also look at charts on their screen and cry 'these data are rubbish!'

[7] See for instance, Mayer-Schönberger and Cukier (2013), conflating the existence of a record, and the record's power to evidence. This confusion is best exemplified by expressions such as "let the data speak", suggesting records truth-tell if only humans remove the encumbrances. However, literature has overwhelmingly justified why we need a definition of scientific data that is different from that of computer data, by focusing on the conditions that take data to evidence (e.g. Gitelman 2013; Leonelli 2016; Tempini 2015).

can certain objects have epistemic value in scientific practice when their digital sources are not directly useful in science?

In the next section, I will introduce the empirical material that can help us in relating *computer data* to the problem of *scientific data* reuse. This is a case study of a *data linkage* infrastructure and the practices associated to its development and use in research. The term of art, data linkage, is used in public health research to speak about the combined re-use of datasets of different origins. Records of a patient's interactions with a hospital or a GP practice can be combined with records from other institutions and sites of data generation (e.g. genetic profiles, environmental and weather data, and socio-economic data among others), to investigate multi-sited relations between phenomena. More about what data linkage is and its current relevance is offered in the beginning of the next section. In the following section, I discuss a framework to understand the digital apparatus affording big data practices in science, first by analysing key characteristics of computational technology, then its operations on computer data. Elaborating on the case material in light of the framework, the concluding section will argue that working from a perspective that is specific on digital objects in science is worth the effort and makes an original contribution to the fields of philosophy and social studies of science.

2 Unpacking Digital Data Reuse in Data Linkage Practice

Data linkage can open new spaces of research, allowing to investigate questions that would be otherwise very difficult to pursue and for which no pre-existing data source, taken alone, can provide enough information. Itself, the term already stresses how in some situations *data can be productively used only if they are put in some kind of relation with other data.* In particular we are concerned with an additive process here: linkage tries to *make data usable for more purposes*.

2.1 Introduction to MEDMI

Accordingly, data linkage infrastructures are projects aimed at enabling the re-use of certain datasets well beyond their original use cases. The case study I present in this chapter is the one I conducted of the Medical and Environmental Data Mash-up Infrastructure (MEDMI).[8] It is an infrastructure and data repository developed to

[8] The following empirical narrative is built on an extensive qualitative case study that I have conducted in 2015–2017 on several infrastructures for the reuse of heterogenous data sources in biomedical research. I approached these infrastructures with a general view to document the associations between organisational forms and processes, infrastructure development, specific data science and data reuse practices, and scientific research concerns, standards and outcomes. Data

foster interdisciplinary research on the links between weather, environment and human health. MEDMI brings together four leading UK research organisations: University of Exeter Medical School, Met Office, Public Health England and the London School of Hygiene and Tropical Medicine. MEDMI aims to develop at once new data linkage methods, technology and demonstrative research. This requires the fulfilment of a few interdependent goals.

First, MEDMI *sources and hosts datasets* that are relevant for the kinds of research it purports to foster. Human health data were sourced from governmental health surveillance databases or GP practice software providers, which are third parties to the project, while environmental and weather data are mainly provided by the data owner the Met Office, who is project partner. MEDMI has datasets of gridded weather variables values (NCIC), surface station observed and derived parameters (MIDAS), and automatically-collected air quality data (AURN) and ozone data from the UK DEFRA[9]; health data include, among others, datasets about observed cases of infections caused by seasonal pathogens (Second Generation Health Surveillance System – SGSS), but on a more restricted basis researchers have had access to mortality data from the Office for National Statistics, and GP practice data shared by one of the major software vendors in the UK (TPP). Several other health datasets owned by individual researchers have also been linked to MEDMI data for specific research projects. The task of making these datasets available includes their curation and harmonisation (more later).

Second, MEDMI researchers *develop data linkage methods and infrastructures* needed to make the combined re-use of these datasets possible. The linkage methods were devised through a collective interdisciplinary effort involving mathematicians, statisticians, weather and environmental scientists, informaticians, and health researchers. Results are a distributed and optimised data storage architecture and a library of highly configurable tools, developed in Python programming language scripts, that allow the researcher to connect to the hosting server and start to probe the depths and shape of the datasets. How these tools *interface* the researcher with the data is key for this investigation into data materiality and use, as we will see.

Third, MEDMI aims to *demonstrate the research* that new infrastructures for data linkage can make possible. The emphasis on demonstration highlights how the value of research thus conducted was not to be defined solely by the knowledge they contributed, but also and especially because of the way they exemplify, and let others

collection included both primary data (in the form of noted observations, interviews, and screenshots), and secondary data (mainly in the form of documents, spreadsheets and presentations) and was executed in the occasion of site visits, participation in meetings, and computer-mediated data gathering. I conducted a total of 24 interviews with MEDMI researchers at all levels, all focused on documenting data reuse and linkage practices and the experiences and challenges associated to them, visiting teams in Truro, Exeter, Colindale, Swansea and London in the UK. Recorded observations included auto-ethnographic notes that I performed by using first-hand the MEDMI data linkage infrastructure, in training sessions hosted at the UK MET Office and from my own home through SSH remote terminal connection.

[9] Department for Environment, Food and Rural Affairs.

imagine, a new way of doing research with and through the infrastructure. Three larger demonstrative projects were part of the initial project plan. At a later stage and close to the expiration of grant funding, the re-allocation of some financial resources allowed to also sponsor some "pilot projects" of shorter duration. A highly heterogeneous set of projects tested the infrastructure – some examples will be mentioned shortly – and provided feedback about the new research tools and linkage infrastructure.

2.2 Data Relations and Epistemic Relations

The overarching premise of MEDMI is that researchers can use combined weather, environment and human health data to understand the effects of climatic and environmental change on human health. In order to do so, they need to access heterogeneous data that originated in different epistemic communities in response to various research questions, standards and assumptions. To make conjoint use of different data in new situations of inquiry, researchers need to define some parameters to be the *invariants* that can act as shared reference point, the contact points or pivots, as it were, that allow juxtaposed datasets to be analysed consistently. For instance, Leonelli and Tempini (2018) examined how location is constructed and used as invariant parameter by finding ways to commensurate between very different definitions of space (e.g. grids, postcodes, catchment areas, ground observations – see also Shavit and Griesemer 2009). The interdisciplinary questions of the kind that MEDMI researchers study hypothesize relations between phenomena (e.g. 'a pathogen responding to weather fluctuations will cause occurrence of health cases with variable incidence') that require these kinds of *data linkage through invariants* in order to be investigated.

In one such project, MEDMI researchers aimed at investigating pathogen seasonality – and more specifically the relation between certain cases of human infections (e.g. food poisoning), the pathogen populations, and weather variation. A hypothesis of this kind implies a complex causal chain, as researchers try to understand the relative weight of different components of climate (e.g., rainfall vs temperature) on the growth of various strains of pathogens, and finally the relation of fluctuating pathogen populations to the number and timing of the observed cases of infection. It requires also to try and account for external confounders such as, for instance, vaccination campaigns. To do this, researchers used national health surveillance data, provided with some location information (in this case, lab postcodes), and weather data on a number of parameters and for a time range of up to 25 years (Djennad et al. 2017). Since the spatial coordinates for a food poisoning event had to be based on the location of the testing lab (and the specific rationalisation of space embedded in its postcode) researchers needed to decide how to spatially partition weather dataset (originally modelled on grid space). Consequently, they would decide what portions would be capturing information about weather events that are deemed to be relevant for explaining swings in pathogen populations.

However, and before examining the more traditionally recognisable scientific work, the focus of this chapter requires us to examine in detail how the data in MEDMI are operationally prepared, accessed, and worked with – in other words, the computational strategies and operations that are put in place in order to mediate and enable the re-use of these scientific data. What the empirical material shows is that a crucial feature of data linkage practices is that linkage is not an operation that stops at the surface level of the dataset. Instead, data linkage practices open up the datasets from the 'inside', to select among available source data, transform the data into different constructs, and compile a derivative dataset (the 'linked data') that is an exportable product of this data processing activity. In MEDMI there is no such thing as 'prêt-à-porter' linked data. The sheer size of datasets would make this practically cumbersome when at all meaningful. Instead, the way the dataset is interacted with is as a navigable space, of which no comprehensive 'view' is possible, but one that the user can probe via the terminal interface. Despite these interface constraints, datasets are not a monolith object, of which only pre-determined chunks can be exported.

Editing datasets in order to link them with one another is an activity made complex by the fact that the different components of source datasets are structurally related to one another. Understanding the repercussions of each applied change is crucial. When stored in industry-standard *relational databases*, data values are organised in tables. The structure of a table, organised in rows and columns, reflects a statement about how groups of data values *relate* to one another: in the case of a food poisoning pathogen, a basic set comprising the time of the scientific observation, the place of the observation and the object of the observation are some of the values that are related. Each of them *complements* the information that the other provides. The relations between data values thus encoded by the database structure make part of the informational context in which every data value is embedded and evaluated. Metadata are thus themselves data; the designation of metadata simply reflects assumptions as to what data values are seen as central in a particular – data values that are seen more as context are the meta-. The existence of a structure of epistemic relations between various data fields stored in a database makes it very sensitive to 'lift' certain values from a table without the others following as well, or to 'manipulate' them. And yet, databases' granular[10] structure is powerful precisely because it can be easily changed, its components can be unbundled, modified and reassembled in new tables. Researchers will hope them to reflect those putative relations between phenomena that can be statistically analysed further.

Hence, far from a digital equivalent of a well-ordered library to upload and download packaged volumes of data, the full reuse of MEDMI data is made possible only once the researcher: is granted remote access to the server; has selected a few

[10] In this chapter, I define granular a complex object including parts that are in homogeneous and commensurable. In this volume, Cambrosio and colleagues also use the term granularity to talk about differences of resolution in knowledge about cancer (Cambrosio et al. this volume): *"while knowledge at the level of a gene, as captured in guidelines and regulatory documents might be relatively stable and/or robust, the same does not necessarily apply to gene variants."*

of the available datasets, has then further selected subsets of data from the datasets (i.e. specific columns in each table, and specific spatial and time ranges); has set the parameters for the computation of derivative data values and eventually transformed some into the set *linkage denominators* (e.g. calculating equivalences between different spatial or temporal resolutions, or other quantitatively measured dimensions; but also establishing commensurability between different qualitative or non-numerical denominators); has linked the data (by retrieving the matching records from different database storage locations, computing them and storing the results into a new working table); and has eventually exported the new dataset into a standard format file to further model and analyse them with statistical packages and other tools of choice.

This is a process which researchers can repeat multiple times, if needed, to tweak parametric choices. But it eventually leads to the production of a new heterogeneous composite which I will call the *data mix,* which can be exported in new CSV files.[11] For the researcher performing the linkage, the data mix is a new epistemic object that joins together, in a stable form, information about different phenomena that was previously unavailable, latent or separate, and that will be further analysed with computational technology. The data mix – in the words of a population health researcher: *"something that can be used again and again"* – will be taken to the researcher's computational environment of choice. With the help of various software packages (e.g., R, Stata, MatLab, etc.) it will be further modelled, analysed, used as evidence for evaluating knowledge claims about the world, and eventually further transformed into the material for publications: tables of aggregate values, diagrams, etc.

All MEDMI researchers thus navigate the datasets and evaluate between various possibilities of configuration and recombination of the data sources. This interaction between the human actor, the computational infrastructure and the available digital computer data is a necessary step without which reuse of the data in scientific practice is not possible. The infrastructure is a flexible *virtual analytical environment*[12] that is used to explore and understand the properties of datasets, as well as to *construct, generate and export new data mixes.*

From an infrastructure architecture perspective, the data mix construction workflow I built around computational interactions with two classes of software objects: "imports", to be used by infrastructure developers for importing source datasets and performing data management and preparatory curation; and "datasets", which are used to construct the linked data from the imports, by selection, manipulation and extraction of the data, in the way I have just described. For more detail, Box 1 gives a simple example of a data linkage commands sequence that can be executed in order to prepare the data mix needed to analyse the relationship between nettle pollen and humidity.

[11] *Comma-separated values*, a standard format for spreadsheet like tabulations.

[12] I use 'virtual' here as 'a space of prefigured combinatorial possibilities,' that shape the potential operations to be done with the data – a space of potentialities that are *not spontaneous* upon occurrence.

Box 1: Basics of Data Linkage

This data linkage exercise was part of the MEDMI researcher training sessions I took part in at the UK MET Office.[13] The computer commands reported below are executed in live Python environment (Python is a programming language very popular in data science practice). In order to be able to input these commands, a researcher needs only a conventional computer connected to the Internet. She has successfully used an operating system shell (a command-line interface for entering computer commands) to securely connect via remote terminal to the MEDMI servers, which host the source data and execute the data linkage computations. Once connected, the system assigns her with a working folder, hosted remotely. This is a space to store the files resulting from data linkage operations. By inputting sequences of custom commands, she can thus proceed to select, manipulate, generate and extract the data of interest.

The following commands are an example of selection of environmental and weather data (pollen and humidity measurements). Their juxtaposition with one another (linkage) is made possible by the selection of common spatial and temporal denominators and the consequent computation of the source data according to the new denominators.

d1 = Dataset({'Source reference': 'midas.pollen_drnl_ob.urtica', 'Time range': ['2014-8-1', '2014-9-1']})1']})

The researcher selects measurements for nettle pollen from August 2014 which were originally imported from the MIDAS dataset, and notes it as d1.

d1b = Dataset({'Source reference': 'midas.weather_hrly_ob.rltv_hum'})

The researcher selects humidity data from another dataset originating from MIDAS and notes it as d1b.

d1b.process({'Method': 'sp_mean', 'Radius': 100000})
d1b.process({'Method': 'tp_mean'})

The hourly humidity needs to be averaged. The first command will average humidity spatially, selecting all data points falling in a radius of 100 km *around the site of nettle pollen measurement.* The second command will average measurements for the selected time range.

d2.link(d2b)

The two datasets are linked, by executing extraction processes and the transformations as they have been set up by previous commands.

(continued)

[13] I am indebted to Christophe Sarran, MEDMI developer and MET Office scientist, for welcoming me to the MEDMI training sessions and allowing me to reproduce and explain some of the steps involved.

Box 1 (continued)
d2.save_csv('exercise2')
The linked data are exported to a CSV file, and the file can be transferred to other packages and machines.

Myriad other combinations of extraction and transformation requirements can be set up. The parameters can be changed at will by the researchers to further explore the correlation of interest, and similarly can be exported multiple times.

2.3 The Computational Logistics of Digital Data Mixing

To make the data linking process possible, several 'staging' operations need to take place to load the data in the infrastructure and make them computable by the research software. Here developers consider an entire set of concerns that I call the *computational logistics* of working with very large datasets. In spite of its apparent straightforwardness, MEDMI easily tested the limits of the MET Office's supercomputer, one of the most powerful in the UK. Environmental and weather data alone include more than 9.5 billion values over more than 400 parameters, and when initial versions of the linkage software were run computations could take months to complete. Technological architectures can intermediate interaction with data in such extremely different ways from one another, that some approaches can simply make the work impossible, while others reduce costs to irrelevance and make for 'seamless' experiences.

Computational logistics are shaped by how digital data are structured and stored, and how programs access and operate on them. They are determined by the relation between computer data and the computational software that process them. A programmer can conceive of a number of different approaches to data structure, without an end user at the interface level knowing any difference about the rules that computer software must consequently follow to access them. Similarly the programmer can conceive of a great number of algorithms for accessing, processing and storing data according to the same operational specification; each algorithm can execute a different sequence of operations, while all produce, once processing is complete, to the interface results that are all the same from a symbolic point of view. Different combinations of choices for data structure and algorithm sequence, respectively, can have completely different implications for hardware usage patterns and costs.

Hence, if data are structured and stored in ways that favour the most likely styles of retrieval and processing, data re-use will be faster and more reliable from the point of view of scientific research activity and its shifting, situated demands. Developers aim to integrate expectations, demands and models of the scientist's workflow in the design specifications they implement. Consequently, in MEDMI

imported source datasets undergo a number of deeply restructuring operations, to the point that the dataset 'as a file' or single object disappears. The data are broken down in many fragments according to a few structuring principles (e.g., by time range – date of observation), for each fragment to then be pooled, by the same token, together with heterogeneous fragments originating from other source datasets. This pooling is not a linkage in itself, but by pooling data together that are most likely to be computed and linked together, the structure is intended to prefigure a set of 'styles' and 'choices' of data linkage, in a form of *expectant organizing* that is coded in the infrastructure.

Accordingly, the MEDMI software workflow was optimized to this database structure. Linkage steps had to be broken down in piecemeal operations that would retrieve, compute and store data efficiently. Sub-steps should be integrated in sequences so as to enforce a specific order of execution, that is optimised for the retrieval and storage computational logistics that the data structure best affords: as Box 1 exemplifies, MEDMI infrastructure requires the user to fully specify the linkage requirements *before* the processes of data retrieval and computational transformation start. Early MEDMI prototypes allowed a more piecemeal configuration of linkage parameters and computation of linked data. While this would arguably allow researchers more flexibility, data processing times inflated beyond feasible. Refinement of data structure and processing sequences according to computational logistics requirements allowed to shrink completion times.

With the development of infrastructure, linkage is thus part under way. Yet, the interface user (and the philosopher or social scientist that takes the same standpoint) is unaware of it. For the user not to know how the data are fragmented and pooled 'underneath,' the interface software layer virtualises each dataset – describing it as a whole so that it can be 'navigated' seamlessly. Guessing the logistical state of the data from the interface is quite like trying to guess the catch under the waterline with a fishing rod.

Computational logistic strategies reconfigure the way different data structures and technologies relate to each another, and are greatly relevant for our understanding of digital epistemic practices. As *computational infrastructure data, data are structured differently* from how they are structured in the upstream context of origination and the downstream context of reuse. Data are here structured according to considerations of (1) their provenance; (2) the pluripotential, prefigured uses they will be put to in the creation of new linked data datasets, and the related assumptions about the epistemic relations between phenomena that the researchers will seek to investigate analysing the dataset;[14] and (3) constraints on feasible and efficient computation. The three dimensions are interdependent.

[14] In an interesting parallel with Hoeppe's chapter on digital data in astronomy: the digital then is not only what facilitates a certain culture and practice of accountability, but is also a regime of interaction and communication that has logistics and economics shaping that culture in turn (cfr. Hoeppe this volume). Karaka's chapter on data acquisition in high-energy physics also shows the importance of what I call computational logistics in enabling generation and mobilisation of digital data (Karaca this volume).

2.4 "You Need to Say Exactly What You Want": *Data Mixing and Boundaries of Practice*

It is very important to take stock of the breadth of operations that the infrastructure supports, and their epistemological relevance. MEDMI researchers use the infrastructure to prepare a derivative dataset that suits the context of further scientific inquiry and research questions, working hypotheses and assumptions among others. For this, the library of Python modules affords operations such as moving, separating and joining subsets of columns and rows from available tables; and various calculations that generate new variables, which include coding or translating values, interpolations and other estimations, and sampling.

Data management and the kinds of data manipulations involved in shuffling data between relational tables have often been considered a sort of backstage operation of no epistemic relevance. Yet such a range of computational operations on source datasets challenges us to see the entire spectrum of activities so far described as part of scientific data reuse activity, and the data infrastructure developer as a scientist. As we have seen, through careful consideration of different epistemic strategies and their purchase for further data reuse developers optimize data sources and computational infrastructure.

Data linkage operations also have deep implications for the sophisticated analyses that will follow and are as such performed by researchers fully within the context of an active scientific inquiry. They depend on the specific research question that is pursued and the background theoretical scaffolding. Even simple transformations (e.g., the computation of time and spatial arithmetic means) can deeply affect the structure of relations between data fields in a relational table. Results of statistical analyses of the derivative vs source data are differently able to lend evidential support to hypotheses under testing.[15] Once a derivative dataset is created and

[15] Common operations to transform data to the desired level of spatial and temporal resolution and definition are arithmetic mean and minimum and maximum values. These operations can be applied to both space and time values and involve very important trade-offs. An infrastructure developer provided a telling example with the problem of repurposing wind magnitude and direction data captured at a specific time and place:

> You can get a complex mean, which is a mean of the vectors, as opposed to a mean of the magnitude. […] If you have two vectors of the same magnitude in opposite directions then the mean will be zero. While obviously if you just take a mean of the magnitude it will just be the magnitude. […] If it's an atmospheric dispersion question, if you want wind combined with pollen, then you want the mean of the vectors because you want to know where the pollen is going. If you want wind as an exposure value for somebody then the person is exposed to the mean of the magnitudes. If it's windy in every direction, as far as the individual is concerned their exposure is not going to reduce to zero. So, while for pollen, the pollen grain will be moved this way when the wind is in this direction, and it will come back if the wind comes back. So that comes as if it's a wind of zero. So, it really means that the user really needs to think through, 'Actually what is it [that] I want?'

That operations of data processing including estimate and interpolation of new or missing variables have great relevance for consequent analyses should be beyond doubt. In a seminal paper,

exported, it has long departed from the sources that it was built from, but despite its 'newness', it is considered the working material that can be used in further stages of statistical analysis. As it should be clear by now, no reuse of 'as is' MEDMI source data is likely to be ever made.

Negotiation of the epistemic assumptions that the data linkage technology was a central concern. It is precisely for the appreciation of the deep implications of data linkage operations that MEDMI developers opted for a conservative approach as they set which data linkage choices and parameters should be pre-empted by default. They chose to provide researchers with very granular control on data linkage configurations.

The first approach MEDMI scientists took was to default enough linkage parameters so as to build and make available a huge database of already linked datasets. In this approach, a researcher would have needed to perform fewer operations in order to retrieve the data of interest (for instance, selecting subsets of data by specifying time and spatial ranges) and extract the mix. There would be few 'moving parts' to be configured by the researchers, and greater logistical efficiency: datasets could be pre-linked so that many calculations could be performed in advance, and accordingly optimised for faster navigation and retrieval. This approach was abandoned after 1 year of development as the team grew uncomfortable about the amount of assumptions now embedded in hundreds of defaulting parametric choices, and how these choices could remain opaque to end users.[16]

To avoid grafting too many assumptions in the data, the current approach offers instead a different trade-off in the support of scientific inquiries: a steeper learning curve for a more flexible data reuse infrastructure. Importantly, even an approach that postpones many manipulations to a latter stage requires a combined data structure and computational optimization strategy of computational logistics. The datasets were then re-factored once again to reduce some computational tasks from 2 weeks of computer time to less than 1 day, and the emphasis moved on programming more powerful data linkage technologies.

---

leading statistician Meng (1994) clarified the downstream implications of this kind of generative pre-processing: *"imputation is not (merely) a computational tool but rather a mode of inference, which allows hierarchical and sequential input of assessment and information"* (539). Meng introduced the notion of uncongeniality to highlight how assumptions and frameworks informing the data processor can be at odds with those of the end analyst and, most problematically, difficult to scrutinize (Xie and Meng 2016).

[16] An informant explained the compromise:

> That unfortunately meant that we'd potentially have had to go through each of the 400–500 parameters that are in the environmental datasets and determine what are the sensible defaults. We found first of all users were not going into the code to use the code [i.e. to understand the defaults], simply because they are not used to that, I think, in the health sector. In particular, coding is not a huge skill. We also found that how the data was being processed, these defaults were not transparent enough. So users were still not really understanding what was happening to the data before it was being released to them. So the new approach will get rid of all that and we would simply say, 'All of these data are available. These tools are available to process the data. *You need to say exactly what you want*.' [emphasis mine]

3 Discussion: The Relationality of Scientific Computer Data

My main argument is that attending to *computer data* and the practices aimed at turning them into data that can be used as evidence in scientific investigations (*scientific data*) is key to fully understand the conditions shaping digital data reuse and big data innovations in the sciences.[17] The MEDMI case indeed shows how the specific materiality of computer data is implicated in their epistemic journey. In this section I outline a way in which we can further think about digital materiality and computer data productively with respect to our interest in scientific data practices.

3.1 Computer Data as Socio-Technical Relational Objects

Philosophers of technology following Simondon see digital objects as technical objects (Hui 2017; Feenberg 2017), a form of standardised and 'concretised' social practice, whose significance and social role depends on the ways in which it is embedded in the fabric of society and the life-world. Importantly, they understand *computer data as relational.* The ways in which digital objects interact with other technical objects and forms of social activity shape the ways in which these objects are defined.[18] Because of digital objects' extreme level of physical abstraction (inaccessible to us in any direct way, we require several layers of computing technology to interact with them), they are an excellent example of a *socio-technical relational object:* digital data exist, and are interacted with, only through a milieu of other socio-technical elements forming a computational system. Understanding computer intermediation is thus a key step to understand the ways in which a social actor relates to computer data.

3.2 Computer Data as Programmable, Granular and Composite

As new media theorist Manovich reminds us (2001), digital objects are ultimately described in numbers. This makes digital objects *programmable,* amenable to computational manipulation (through any mathematical function that can be successfully scripted as algorithm) at the very lowest level of representation (Borgmann

[17] This has been a key point in my research (e.g., Kallinikos and Tempini 2014; Tempini 2015, 2017).

[18] Ultimately, Hui argues (2017) reading Heidegger (1962), all objects are. The situations of human activity are understood as shaped in time through the nexus relating beings with one another (Dreyfus 1991). Context is a web of constitutive relations.

1999). This also means that digital objects are inherently open and interactive (Manovich 2001): their symbolic nature makes it possible to selectively scan, and interact with, at the level of their constituent components.[19] Components can thus make up a larger object but remain identifiable within it. For instance, one can apply piecemeal changes of an individual data field in a large table. By the same token, I can now create a copy of this Word file, rename it, then open the copy, change a few words in a specific point, leave the rest untouched and save the file. The way in which a selective intervention – swapping characters for one another – can be carried out *within* the document or at other levels of abstraction (such as at the *boundary* of the file object in the case of a format conversion), is specifically "afforded" (Gibson 2013; Faraj and Azad 2012) by a situated socio-technical assemblage in which computing technologies take centre stage.[20]

As we have seen, in MEDMI data linkage practice specific data values (components of the dataset object) within the same table are indeed discriminated from one another and differently manipulated. And at the same time, digital technology also supports developers to carry out computational logistics manipulations at a comprehensive level of abstraction (at the level of a plurality, data pool or set). Therefore, we should highlight two key relational features of what digital data offer to the data scientist. First, computable data sets are *granular* (granular is a complex object including parts that are in some respect homogeneous and commensurable – granules are kin to one another). A dictionary definition defines granularity as "the scale or level of detail in a set of data" (Oxford Dictionary of English 2018; also Aaltonen and Tempini 2014; Dourish 2014; Kallinikos et al. 2013). Second, computable data sets are *composite* (composite is a complex object including parts that are in some respect heterogeneous and incommensurable – composites are made of alterities). It is because a dataset is granular and composite that we can say that the socio-technical relationality of MEDMI data applies at the level of individual values – it is not only the computer file object as a whole that is relational, but also its components.

[19] On this backdrop, Kallinikos et al. (2010) identify as the key attributes of digital objects: editability, interactivity, openness and distributedness. Datasets can be reordered, navigated and made sense of in myriad of ways, and through multiple tools and interfaces. Often, their specific design or their size imply that they are distributed and not accessible in their entirety in an individual site at a given moment – this is often the case with distributed infrastructures.

[20] Aaltonen and Tempini (2014), focusing on *data pools* and big data practices in a commercial setting, highlight how big data work is often articulated at a different scale than that of the individual record, where the sets of data that are relevant for a specific purpose do not necessarily have fixed boundaries. They suggest to be key characteristics of the elusive data pool objects: *comprehensiveness* (data work can survey the entirety of a big data collection), *granularity* (data work can parse through highly granular, individually irrelevant, data points) and *unboundedness* (data work can span beyond clearly perceived boundaries of use). A different use of granularity in relation to data is in Dourish (2014).

3.3 Socio-Technical Relations and Epistemic Relations

Data linkage technology, with its capacity to translate, calculate, juxtapose and recombine large quantities of data about select observational variables thus allows researchers to explore relations between data found *within* the same dataset or *across* different datasets. By mixing data over common definition and resolution of space and time, and by computing means, vectors or other derivative values, data linkage juxtapositions enable the observation of relations between data that are latent within or across datasets. These data relations, of the (here oversimplified) sort of 'warming weather patterns correlate with incidence of food poisoning infections' can then be used to test working hypotheses of the climate change consequences on pathogen seasonality.

Here it is key to appreciate the crucial effect that the recording of scientific information about observed phenomena over symbolic notation has on the possibilities of reuse of the data in computationally transformed and mixed form. Of the huge material diversity of the scientific data that the philosophy of science discusses (including, for instance, biological specimens; artefacts; systematic collections; photographic slides; printed maps; graphs; networks; texts; numerical tables; sequences), the material-agnosticity of symbols makes them the form of data that, in order to generate new meaning, is easiest to aggregate in sets, to mix at the granular level of the individual data token or datum, and to enable the computer to intervene *inside* the dataset object along the ways I have been describing so far. For this reason symbolic data are enjoying the vastest possibilities of reuse in data linkage, analytics, and other big data science applications. As an incomparably vast array of methods for symbol manipulation is then available for implementation over digital means.

From the same infrastructure of methods and calculative procedures, an infinite variety of outcome data mixes is possible, each of which can have different epistemic performance (from each other and from the data sources), depending on the characteristics of the situation at hand. Data mixes of the sort I have been describing can now be a central development in the sciences because they are mixes of symbols. Of course, digital objects are, strictly speaking, entirely symbolic and so, change is always bound to be symbolic at its most fundamental layer of description (for instance, changes in file formats that can make it more difficult to feed a file to software). But here I am trying to work with a distinction between symbolic change 'at the boundary' and what I call selective and granular change, which is change of a select part of the composite object that in turn changes the kinds of epistemic relations that the object can entertain. Many data manipulations such as the estimation of a wind vector from multiple sources are aimed at refining and enabling a certain epistemic performance of the data in a statistical analysis.

3.4 The Scaffolded Relationality of Scientific Computer Data

We must now return to the distinction set out in the introduction, and observe how the status of objects that are at once *scientific data* and *computer data* is thus relational at several levels. Thinking about scientific computer data as embedded in a computational system allows us to think about these objects as characterised by a *scaffolded relationality* dependent on both the socio-technical relationality of computer data and the epistemic relationality of research data. The relational openness of digital objects is dependent on a technical milieu (Hui 2017; Feenberg 2017) whereby computing technologies shape the levels of detail and abstraction, and the operations, through which interaction with data objects takes place. Key operations aimed at assessing, exploring, refining, developing and operationalising their epistemic value can only be applied through computing technologies that, ultimately, are developed according to principles of *computer data* use and manipulation. As computer data's ineliminable 'other', it is key to hold into account the computing *technologies* and computational *operations* through which data practices unfold. Manovich (2001) argues that paying attention to computational operations allows us not to reduce computer technology to 'tool' or 'medium' – a common shortcoming in the philosophy and social studies of science. Dourish (2014) points out that the word database has often been used inconsistently and often with the effect of erasing differences in concept and implementation that have implications for data practices.

It is important here to understand that the two different kinds of relationality of scientific computer data are closely *interdependent*, and this can be explained by looking at the way in which, in data linkage research, the exploration of epistemic relationships between data points recording certain events *is grounded on the capabilities of relational databases and the computational data work they afford*. As I have already pointed out, scientists linking different datasets in order to explore relations between environmental and public health phenomena work by choosing a parameter that can act as a common invariant (see also Leonelli and Tempini 2018).

At a computational level, relational databases revolutionised the way computer data are stored and accessed because of the way in which they allowed to generate new *relations between data* (Hui 2017; Manovich 2001).[21] Dourish (2014) highlights two main ways. First, relational databases are structured through tables, whereby relations between values are expressed as a row conjoins data points distributed over the different columns in a plurality that is more than the sum of its parts. A row recording, over different columns, my demographic details (name, address, gender, age, …) implies that a phenomenic relation exists in the world that holds these values together – this relation is meant to map to myself. Second, the methods to query relational databases with allow the data scientist to explore further relations that link different tables to one another. Here a common point of

[21] Dourish (2014) postulates three key relational database operations (edit data values; insert new row-relation; delete row-relation).

invariance is required between them: if my demographic data were split over two tables (name, address; and name, gender, age) the 'name' data can be used to draw relations across the two tables and between all data points involved, parsing all the demographic data points about myself together again.

MEDMI data linkage practices closely track these two ways of exploring relations between data: first, researchers assess different dataset sources and explore the potential for juxtaposition and the elicitation of latent relations between heterogeneous data; second, they complete the linkage by transforming and pulling data into the new tables of the derivative dataset. Crucially, data linkage practices move from the more flexible and precarious arrangement for exploring epistemic relations between data (screening and exploring source datasets and their metadata) to the more inflexible and stabilised socio-technical arrangement: a unified dataset table, where data values are interrelated through their distribution in rows and columns, that can be more easily exported and analysed with statistical software of choice. The way of relational databases of relating data with one another is closely mapped by the way data linkage researchers are working with data sources and prepare them for reuse.

We can thus recast the scientific practice of data linkage in a new light, if we understand the way in which the relational database and associated computing technology are a key enabling factor enabling methodological strategies based on the construction of invariant parameters. As I have argued throughout, to do this we need to pay attention to computational operations aimed at *constructing new computer data relations and storing them in new dataset objects,* and how these relations are linked to the epistemic relations of interest at a specific stage of the scientific inquiry. We must also be asking what kinds of relations between phenomena are the researchers investigating, and in what ways are the data deemed to speak to them.

The digital dataset, I have argued, is a kind of data object that must be closely studied. I paid special attention to the role of a set of lower level operations that explore and manipulate the composite structure of a dataset. Operations such as those that change the format, code, arrangement and value of symbolic content of digital data in ways that alter the set of uses that the object can undergo in the social settings of scientific practice are *key epistemic object transformations*, that can be now linked to questions of data identity, functional continuity, and data travelling and packaging.

3.5 Computational Data Journeys

The data mixing practices that we have observed in the case of MEDMI in particular stress how in certain situations of scientific inquiry, *data can be productively reused only if they related with other data, and this relation is stabilised in a new relational dataset object.* What does this say to the concern of this volume in the travelling of

data?[22] As it has become clear through this chapter, in MEDMI there is no data that travel in any straightforward sense. Travel evokes a principle of continuity, but neither material nor functional continuities are at play here. Datasets are systematically disassembled in several different ways, transformed and mixed with others. Data about wind measurements needs to be transformed into data about mean wind direction, in the example of pollen dispersion research (see endnote xiii). Source datasets and derivative mixes have very different uses from one another. The mix must fit assumptions, frameworks, methods and research questions of the investigation at hand in a way that the neither of the sources does. New digital mixes have new identity from both a material and epistemic function point of view.

Yet, there is a lot of data movement in the closed confines of a MEDMI's virtual analytical environment, which acts as a sort of template builder, a 'system of infinite dataset generation' somehow recalling what Borges' Library of Babel could make of books. With its collection of computational scripts and methods this digital "library of predefined choices" (Manovich 2001) virtually (and partially) prefigures the data mixes that users produce. Database structures, together with the algorithms that enable to access and edit them, shape digital data reuse by prefiguring and concatenating operations of certain kinds.

We can thus understand the relationship between data sources and derivative mixes *only if we account for what computational technology does* and the computational strategies and methods that it embeds. The operations that are carried out on the data (e.g. comparison, averaging, estimation) are prescribed in the 'memory space of the algorithm technology,[23] and the programmability of digital computational machines allows concatenations of simple operations to be inscribed as steps and combined in complex automated sequences. As obvious as this may all seem, it stresses that computational technologies should not be described as 'tool' or 'media', as they often are, and should rather be approached as complex procedural systems. Computational technology and data should thus be studied together. Digital data are neither a static object not an undefinedly dynamic one, but certainly one that is in a permanently dynamic relationship with the computational technology that access and process them.

Lacking an object that traverses the infrastructure without dissolution and reassembly, it is at this relationship between data and computer that we shall return to explain how digital data 'journey.'[24] Indeed, the gap separating the source dataset and its derivative (which, as I observed, undermines an intuitive interpretation of the journey) can be filled only by *taking the procedural continuity of algorithmic computations as the missing link* in the chain of data travel steps. This is the anchor that materially connects two dataset objects through a traceable path of calculations. A traceable path of computational instructions allows to account for the metamorphosis

[22] Other chapters also discuss relations entertained between data (Morgan this volume), and the dataset as a context that holds these relations together (Griesemer this volume). In her afterword, Longino mentions relations between data (and operations of recording and selection) as a key focal point to overcome a naïve opposition between 'naturalistic' vs 'interpretive' approaches to data (Longino this volume).

[23] Of course, complex software often use structured storage in turn, to support execution.

[24] Needless to say, what I have called so far the digital data journey in MEDMI is just a sub-section of a longer journey.

of datasets, as operations are standardised and remain available for scrutiny and reproduction. The specification of concatenated computational operations allows the chain of custody of the data's power to evidence to be continued despite the literal symbolic transformation and manipulation occurring within the dataset objects. Despite the lack of a data object that can be ostensibly referred to, this intermediate step of the data journey is standing, for a relatively short time, on the shifting ground of computational processes' own determination. Traceable computational procedures here help to secure evidence's chain of custody (cfr. Wylie 2017, this volume), and to recompose the journey. This data journey thus *moves between data and computational technology*, linking together a source dataset object, the intermediate set of computational transformations executed by technology, and a derivative dataset object.

4 Conclusion

Manovich (2001) provocatively argues the *mix* to be the key cultural form of new media and the DJ its artist, highlighting their post-industrial, post-modern roots. I took this as an opportunity to think of new data science methods as *data mixing* and of the data mix as a quintessential object of big data innovation. The metaphor choice of the *data mix* strongly resonates with MEDMI actors' own use of the *data mashup* category,[25] but has better theoretical grounding.

In this chapter I argued that the technology-intermediated practices of manipulation of computer data relations are key to epistemic practices concerned with developing new data objects. In these new data objects, scientists isolate and point to specific epistemic relations, bestowing the content of the dataset with the status of scientific data relative to specific situations of inquiry (Leonelli 2016). I argued that the computational processes underpinning these practices bear the chain of custody that enables derivative data to be used as a source of scientific evidence. I claimed that, ultimately, digital materiality bears a difference for scientific practice that is worth understanding. The intention was, all along, to invite philosophers and social scholars of science to study digital technology more closely.

Key strengths of the framework for the study of scientific computer data that I have been proposing include:

- Understanding computer data allows to problematise their relational objecthood with questions on the computability of data, the relationality between data and computational systems, and the epistemic consequences of technological intermediation.
- Understanding computer datasets as granular and composite, amenable to discrete intervention, highlights how scientists achieve their reuse through complex chains of operations of disassembly, transformation and re-assembly; and puts into focus the relationship between the dataset and the data point by highlighting how data components, such as a string or data point, are usually not mobilised as individual tokens, but rather, together with others and as a set.

[25] As they observe, the mashup terminology has roots in jazz (Fleming et al. 2014). It comes to data science through systems engineering (Daniel and Matera 2014).

- Understanding digital objects as technical relational objects allows to pay special attention to the role of computing technology as a key intermediary and step of the data journey; and to understand of the epistemological implications of computational logistics, optimisation choices and alternative computational strategies and infrastructure architectures – steps in the data journey that are epistemically relevant yet fall between clearer stages of scientific data origin and reuse.
- Studying the kinds of operations that computational technology carries out uncovers the key importance of computer systems' focus on the creation and organization of relations between computer data that feed in scientific practice, where they are evaluated; and it demonstrates the link between the creation of new *computer data relations* in computer systems and their potential role in support of evidential claims about the world, relative to hypotheses and assumptions about relations between phenomena of interest.
- Through this 'cascade' of observations about what makes *computer* vs *scientific* data, we can then grasp that the intense relationality of *scientific computer data* is multi-layered and scaffolded, as it depends on relations between various kinds of data, computing technologies, assumptions, theoretical scaffoldings, hypotheses and other features of the situation at hand.

Acknowledgements I would like to thank the audience of the 4th Exeter Data Studies workshop *Varieties of Data Journeys: Data Processing and Movements Within and Across Practices*, Nov 2017. An early draft benefited from various comments. Similarly, I thank the audiences of the *Scales of Justice: Calculating Norms and Probability* workshop, Institute for Advanced Legal Studies, London, June 2018; and the *Data Science and Social Research Conference,* Milan, February 2019. In particular, thanks to Hyo Yoon Kang, Rachel Ankeny, Mary Morgan, Gabriele Gramelsberger and Judit Varga. Special thanks to Sabina Leonelli and this book's inspiring authors for the opportunity to share our work together. This work was supported by ERC grant award 335925 (DATA_SCIENCE), and by EPSRC grant EP/N510129/1.

References

Aaltonen, Aleksi, and Niccolò Tempini. 2014. Everything Counts in Large Amounts: A Critical Realist Case Study on Data-Based Production. *Journal of Information Technology* 29: 97–110. https://doi.org/10.1057/jit.2013.29.

Borgmann, Albert. 1999. *Holding on to Reality: The Nature of Information at the Turn of the Millennium*. Chicago: The University of Chicago Press.

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.

Chapman, Robert, and Alison Wylie. 2016. *Evidential Reasoning in Archaeology*. Bloomsbury Publishing.

Daniel, Florian, and Maristella Matera. 2014. *Mashups: Concepts, Models and Architectures (Data-Centric Systems and Applications)*. New York: Springer.

Djennad, Abdelmajid, Gordon Nichols, Gianni Loiacono, Lora Fleming, Anthony Kessel, Sari Kovats, Iain Lake, et al. 2017. The Seasonality and Effects of Temperature and Rainfall on Campylobacter Infections. *International Journal for Population Data Science* 1. https://doi.org/10.23889/ijpds.v1i1.51.

Dourish, Paul. 2014. No SQL: The Shifting Materialities of Database Technology : Computational Culture. *Computational Culture*.

Dreyfus, Hubert L. 1991. *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. London: MIT Press.

Faraj, Samer, and Bijan Azad. 2012. The Materiality of Technology: An Affordance Perspective. In *Materiality and Organizing: Social Interaction in a Technological World*, ed. Paul M. Leonardi, Bonnie A. Nardi, and Jannis Kallinikos, 237–258. Oxford: Oxford University Press.

Feenberg, Andrew Lewis. 2017. Concretizing Simondon and Constructivism: A Recursive Contribution to the Theory of Concretization. *Science, Technology, & Human Values* 42: 62–85. https://doi.org/10.1177/0162243916661763.

Fleming, Lora E., Andy Haines, Brian Golding, Anthony Kessel, Anna Cichowska, Clive E. Sabel, Michael H. Depledge, et al. 2014. Data Mashups: Potential Contribution to Decision Support on Climate Change and Health. *International Journal of Environmental Research and Public Health* 11: 1725–1746. https://doi.org/10.3390/ijerph110201725.

Fleming, Lora, Niccolò Tempini, Harriet Gordon-Brown, Gordon L. Nichols, Christophe Sarran, Paolo Vineis, Giovanni Leonardi, et al. 2017. Big Data in Environment and Human Health. In *Oxford Research Encyclopedia of Environmental Science*, Vol. 1. Oxford University Press

Gibson, James J. 2013. *The Ecological Approach to Visual Perception*. Psychology Press.

Gitelman, Lisa, ed. 2013. *Raw Data is an Oxymoron*. Cambridge, MA: The 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.

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.

Heidegger, Martin. 1962. *Being and Time*. Oxford: Blackwell.

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.

Hui, Yuk. 2012. What is a Digital Object? *Metaphilosophy* 43: 380–395. https://doi.org/10.1111/j.1467-9973.2012.01761.x.

———. 2017. *On the Existence of Digital Objects*. Minneapolis: University of Minnesota Press.

Kallinikos, Jannis, and Niccolò Tempini. 2014. Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation. *Information Systems Research* 25: 817–833. https://doi.org/10.1287/isre.2014.0544.

Kallinikos, Jannis, Aleksi Ville Aaltonen, and Attila Marton. 2010. A Theory of Digital Objects. *First Monday* 15 (6). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3033/2564.

———. 2013. The Ambivalent Ontology of Digital Artifacts. *MIS Quarterly* 37: 357–370.

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.

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

Leonelli, Sabina, and Niccolò Tempini. 2018. Where Health and Environment Meet: The Use of Invariant Parameters in Big Data Analysis. *Synthese*: 1–20. https://doi.org/10.1007/s11229-018-1844-2.

Longino, Helen E. this volume. Afterword: Data in Transit. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Manovich, Lev. 2001. *The Language of New Media*. Cambridge, MA: MIT Press.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution that will Transform How We Live, Work and Think*. London: John Murray.

Meng, Xiao-Li. 1994. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science* 9: 538–558. https://doi.org/10.1214/ss/1177010269.

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.

Oxford Dictionary of English. 2018. *Granularity*. Definition of Granularity in English by Oxford Dictionaries. Oxford Dictionaries | English.

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.

Rheinberger, Hans-Jörg. 2010. *An Epistemology of the Concrete: Twentieth-Century Histories of Life*. Duke University Press.

Shavit, Ayelet, and James Griesemer. 2009. There and Back Again, or the Problem of Locality in Biodiversity Surveys. *Philosophy of Science* 76: 273–294. https://doi.org/10.1086/649805.

Tempini, Niccolò. 2015. Governing PatientsLikeMe: Information Production and Research Through an Open, Distributed and Data-Based Social Media Network. *The Information Society* 31: 193–211. https://doi.org/10.1080/01972243.2015.998108.

———. 2017. Till Data Do Us Part: Understanding Data-Based Value Creation in Data-Intensive Infrastructures. *Information and Organization* 27: 191–210. https://doi.org/10.1016/j.infoandorg.2017.08.001.

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.

Wylie, Alison. 2017. How Archaeological Evidence Bites Back: Strategies for Putting Old Data to Work in New Ways. *Science, Technology, & Human Values* 42: 203–225. https://doi.org/10.1177/0162243916671200.

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.

Xie, Xianchao, and Xiao-Li Meng. 2016. Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial? *Statistica Sinica*. https://doi.org/10.5705/ss.2014.067.

Niccolò Tempini is Senior Lecturer in Data Studies at the University of Exeter, Department of Sociology, Philosophy and Anthropology, and a Turing Fellow at the Alan Turing Institute. He is an interdisciplinary social scientist interested in questions of information, data, technology, organization, value and knowledge. He researches Big Data research and digital infrastructures, investigating the specific knowledge production economies, organization forms and data management innovations that these projects engender with a focus in their social and epistemic consequences. He studies the practices of data scientists, software developers, researchers and nonprofessionalised experts to understand how different forms of knowledge and value intersect with each other when different actors come to grips with new methods and new forms of data, information technology and organization. His research has been published in international journals across science and technology studies, information systems, sociology and philosophy (more information at www.tempini.info).

Data, Meta Data and Pattern Data: How Franz Boas Mobilized Anthropometric Data, 1890 and Beyond

Staffan Müller-Wille

Abstract Between 1890 and 1911, the German-American anthropologist Franz Boas conducted a whole suite of anthropometric studies, which all in all generated data from body measurements carried out on about 27,000 individuals. To this day, this data is being re-analyzed by researchers with a range of disciplinary interests. In my chapter, I will take a close look at a small subset of the original datasheets Boas used in his surveys, and how he and other scientists processed the data in later publications. My analysis will reveal that the extraordinary potential for travel and re-use of Boas's data crucially depended on the way in which he designed his surveys. Alongside recording standard anthropometric variables, Boas collected genealogical and geographical information on the individuals measured, which allowed him to flexibly classify data in a variety of ways. It is this richness in structure, or "pattern data," that explains why the data from Boas's anthropometric projects remain valuable for researchers from a variety of disciplines to this very day.

1 Introduction

In June 2003, in a special section of the journal *American Anthropologist* entitled "Did Boas Get It Right or Wrong?," a debate played out between two teams of researchers under the journal's rubric "Exchange Across Differences". The subject of the debate was a large-scale anthropometric study carried out by the German-American anthropologist Franz Boas (1858–1942) on a cohort of European immigrants to the US and their American-born children (Gravlee et al. 2003b; Sparks and Jantz 2003). That Boas's study, by then more than 90 years old, should still spark debate after such a long time is not surprising. Boas had found statistical evidence for slight but significant changes in physical traits such as head-form among descendants of immigrants pointing to changes in "type". This finding formed one of the empirical cornerstones of his sustained critique of racial typologies (Boas 1911, p. 53–58), and this critique, in turn, has been framing debates among anthropologists

S. Müller-Wille (✉)
Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK
e-mail: sewm3@cam.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_14

and population geneticists about the biological and political meaningfulness of the concept of race to this very day (Jackson and Depew 2017). What is more surprising is that each of the two opposing teams of researchers reached their divergent assessments by an independent reanalysis of the very same data that Boas had collected in his original study.

What are the conditions that make such re-use of data across time and changing disciplinary contexts possible? Or to ask the same question in the terms favored by this volume: What enabled Boas's data to journey from their original site of production in early twentieth-century New York, across the changing landscape of twentieth-century physical anthropology and human population genetics, and into the electronic databases of twenty-first century researchers? Historians and philosophers of science as well as STS scholars have emphasized in recent years the key role that metadata play in enabling data to travel. But metadata is a deceivingly simple concept; it is usually understood to refer to information that helps evaluating and analyzing data by providing information regarding the circumstances of their production (Leonelli 2014, 4–5; for a more detailed discussion, see Leonelli 2016, ch. 4). Complexities arise, however, from the fact that, in any given study, it is not obvious what counts as relevant metadata, and what standards should be followed in annotating them. While there exist regimes of data-production that can rely on notions of metadata that have remained stable for centuries – e.g. in bibliography or taxonomy – ongoing research often involves improvised and shifting sets of metadata (Edwards et al. 2011). What counts as data, and what counts as metadata – or what counts as product, and what counts as circumstance of a given experiment or observation – is hardly down to an analytical distinction, but depends on the theoretical perspective of, and questions being asked by, researchers.

In this chapter, I am going to explore these complexities by taking a close look at the data collected by Boas in his statistical studies of physical variation among different human "races" and "tribes," and how these data were reused not only by Boas himself, but also by later researchers. Boas conducted a whole suite of anthropometric studies between 1890 and 1911, which all in all generated data from body measurements carried out on about 27,000 individuals. These anthropometric campaigns where funded by various organizations, including the British Association for the Advancement of Science, the Bureau of American Ethnology and the US Immigration Commission, and peaked twice: once, in 1891 and 1892, when Boas and about 50 field observers collected data on c. 12,000 persons of Native American origin; and a second time in 1909, when Boas took measurements on c. 10,000 immigrants to the United States and their children (for a succinct overview and assessment of Boas's anthropometric surveys, see Jantz 2003).

I am going to approach this case study, first, by analyzing early programmatic statements by Boas that cast light on his statistical outlook on human diversity, which placed emphasis on individuals, not types. In the second section, I will zoom in on a sample of the data sheets that Boas used in his surveys in order to provide a detailed reconstruction of how the original data was coproduced by Boas, his field observers, and their informants. The final section will then look at how Boas, but also various anthropologists in the twentieth century, used this data to draw out a

variety of general conclusions about the evolution of human populations. Sustained data journeys in human population studies, I will conclude, are not only made possible by "fixing" data once and for all in some durable numerical and tabular form, but more importantly by including qualitative "pattern data" in the study. "Pattern data" sit uneasily within the distinction of data and metadata, referring to structures within a population such as genealogical or geographic origin that can both be seen as data about the study subjects and describing the circumstances under which data are produced. They play a crucial role, however, in mobilizing data for re-use by allowing for the flexible re-arrangement of data in order to employ new statistical methods or address new questions.

2 Boas's Statistical Outlook

George Stocking has assigned Boas the role of a founding father of the "modern anthropological culture concept" characterized by "historicity, plurality, behavioral determinism, integration, and relativism" (Stocking Jr 1983, 230). At the same time, Stocking has portrayed Boas's work in physical anthropology as instrumental in the "passing of a romantic conception of race – of the ideas of racial 'essence,' of racial 'genius,' of racial 'soul,' of race as a supra-individual organic identity." In particular it was Boas's statistical approach that was, as Stocking put it, "subversive of traditional racial assumptions" (Ibid., 192–94; see also Xie 1988). And this "critique of racial formalism", as he dubbed it, was not just theoretical. Boas, as we will see in the next two sections of this chapter, was an ardent and up-to-date practitioner of physical anthropology and biometry, highly aware of the intricate problems of the "personal equation" involved in anthropometric measurement, innovative in the design of anthropometric surveys, and creating new mathematical and visual tools for studying statistical correlations. But in order to understand his statistical approach, it is useful to leave anthropometry aside and turn to some early programmatic statements in which Boas advocated the use of statistical methods for the study of culture.

In 1887, Boas became involved in a debate about museum displays (Jacknis 1985; Jenkins 1994). Otis Tufton Mason, curator of ethnology at the Smithsonian Institution, had suggested to arrange ethnological displays at the United States National Museum according to a classification of the objects displayed; exemplars of different varieties of artifacts, he maintained, should be arranged in series, each representing a stage in the evolution of its kind; the rationale on which this presentation rested was borrowed from evolutionary biology. As Boas quoted Mason (without specifying his source):

> [Human inventions] may be divided into families, genera, and species. They may be studied in their several ontogenies (that is we may watch the unfolding of each individual thing from its raw material to its finished production). They may be regarded as the products of specific evolution out of natural objects serving human wants and up to the most delicate machine performing the same function. They may be modified by their relationship, one to

> another, in sets, outfits, apparatus, just as the insect and flower are co-ordinately transformed. They observe the law of change under environment and geographical distribution. (Boas 1887a, 485)

The alternative Boas proposed was to arrange collections "according to tribes, in order to teach the peculiar style of each group." The reasons he adduced for this position were epistemological:

> In regarding the technological phenomenon as a biological specimen, and trying to classify it, [Mason] introduces the rigid abstractions species, genus, and family into ethnology, the true meaning of which it took so long to understand. It is only since the development of the evolutional [sic] theory that it became clear that the object of study is the individual, not abstractions from the individual under observation. We have to study each ethnological specimen individually in its history and in its medium […]. Our objection to Mason's idea is, that classification is not explanation. (Ibid. 485)

This seems to be a strange way of reasoning: first of all, "studying each ethnological specimen individually in its history and in its medium" would, taken literally, be an endless task, and both Mason as well as other participants in the debate pointed out the practical difficulties that an arrangement by tribes would imply (Dall 1887, 587; Powell 1887, 612–13). Secondly, an arrangement according to tribes seems to involve as much classification as that proposed by Mason. What, one can ask, defines a "tribe," especially since tribal identity is highly fluid over time? Also this criticism was raised in the debate, accompanied by the remarkable observation that "a museum collected to represent the tribes of America … to be properly representative, would have to be collected as the census of the native inhabitants of India has been taken, all in one day, by an army of collectors" (Powell 1887, 612). To this criticism, Boas only had a short, categorical reply: "Such groups [i.e. tribes, and groups of tribes] are not at all intended to be classifications" (Boas 1887b, 614).

Boas's studies of native myths along the North-Pacific coast carried out between 1888 and 1895 can serve as an example to elucidate what he had in mind with this strange assertion. In these studies, Boas broke down the myths into constituent "elements" and recorded their distribution within a group of geographically contiguous "tribes". "We can in this manner," as Boas explained in a paper summarizing the results of his mythological studies, "trace what we might call a dwindling down of an elaborate cyclus [sic] of myths to mere adventures, or even to incidents of adventures, and we can follow the process step by step." In more detail, he described this method as follows:

> If we have a full collection of the tales and myths of all the tribes of a certain region, and then tabulate the number of incidents which all the collections from each tribe have in common with any selected tribe, the number of common incidents will be larger the more intimate the relation of the two tribes and the nearer they live together. This is what we observe in a tabulation of the material collected at the North Pacific Coast. On the whole, the nearer the people, the greater the number of common elements; the farther apart, the less the number. (Boas 1896, 2–3)

The article from which this quote is taken does not contain any "tabulation," but so does a German monograph to which it refers and that Boas had put together in 1895 from earlier reports documenting North Western myths in the Proceedings of

the Berlin Society for Anthropology, Ethnology and Prehistory (on Boas's early publication strategy, which relied on German academic periodicals, see L. Müller-Wille 2014). From the tables included in the final chapter of this monograph, it becomes clear that, for Boas, it was the unequal distribution of "incidents of adventures" that defined them as constituent elements of myths in the first place. The table arranges the data – page references to the preceding collection of tales that Boas had "recorded from the mouth of Indians" (Boas 1895a, v: *aus dem Munde der Indianer aufgezeichnet*) during field research in the late 1880s – in such a way that one can immediately see how the full cycle of a particular myth is present in a small group of neighboring tribes while it "dwindles down ... to mere adventures, or even to incidents of adventures" to the left and the right of the table occupied by more distantly related tribal groups (see Fig. 1).

Such a "statistical inquiry", as Boas called his investigation of Northwestern myths (Boas 1896, 3) rested on a "fundamental condition", which "differentiates our method from other investigators […], who see a proof of dissemination or even blood relationship in each similarity that is found between a certain tribe and any other tribe of the globe." The material, on which an investigation was based, had to be "collected in contiguous areas" (Ibid., p. 6). This contiguity was largely, but not necessarily a geographical one, as Boas emphasized; in addition, marriage, kinship, and social structure entered the picture. "The social customs of the Kwakiutl" – the ethnic group most intensely studied by Boas during several field trips – are, he maintained, "based entirely upon the division into clans and the ranking of each individual is the higher – at least to a certain extent – the more important the legend of the clan." Moreover, "the customs of the tribe are such that by means of a marriage the young husband acquires the clan legends of his wife, and the warrior who slays an enemy those of the person whom he has slain. By this means a large number of traditions of the neighboring tribes have been incorporated in the mythology of the Kwakiutl" (Ibid., p. 8–9). The clan system that Boas had detected among the Kwakiutl was actually even more complex than described in this quote; through

	Bilqula	Hä'iltsuk·	Newettee	Kwakiutl	Nutka	Comox	Küsten-Selisch	Fraser River
1. Nerz steigt an einer Pfeilkette in den Himmel, besucht seinen Vater und trägt für ihn die Sonne. Da er die Erde verbrennt, wird er herabgestürzt	S. 246	S. 215, 234	S. 172	S. 157	—	—	—	—
2. Erhält das Feuer von den Gespenstern	—	—	—	S. 158	—	—	S. 54	S. 43
3. Tödtet den Sohn des Häuptlings der Wölfe	—	—	—	S. 150	S. 98	S. 75	—	—
4. Greift badende Frauen an	—	—	S. 172	—	S. 108	S. 73	—	S. 26
5. Heirathet Pflanzen und Thiere	—	—	—	S. 158	S. 100	S. 71	—	S. 44
6. Lässt sich begraben	—	—	—	Vorhanden	—	S. 73	—	S. 73
7. Tödtet die Otter, um deren Frau zu erlangen	—	—	—	S. 158	—	S. 72	—	—
8. Verführt ein Mädchen, mit ihm in den Wald zu gehen und sich auf ihn zu setzen, unter dem Vorgeben, er sei ein heilendes Kraut	S. 243	S. 211	S. 178	S. 160	S. 108	S. 71	—	—
9. Geht zum Donnervogel, um dessen Frau zu rauben	—	S. 210, 211	S. 179	S. 104	—	S. 82	—	S. 34

Fig. 1 Table from Franz Boas, *Indianische Sagen von der Nord-Pacifischen Küste Amerikas* (Berlin: A. Asher, 1895a), pp. 338–39. The columns relate to groups of tribes, the rows to narrative elements of the myth in question. The full suite of incidents making up the myth is only prevalent among the Kwakiutl, while individual elements can be found in more distant tribes. The fields of the table contain page references to the preceding collection of mythical material collected and documented by Boas

marriage, the husband did not personally acquire the clan status of his wife, but he acquired it "for his son" (Boas 1897, 334–35).

By relating the distribution of mythical elements to a space whose contiguity could be ascertained in terms of geographic and socio-political relations among individuals – alliances as well as antagonisms – Boas wanted to circumvent the pitfalls of analogical reasoning in anthropology that he warned his colleagues of in the museum debate. Ironically, however, the grand picture that Boas came up with on the basis of this approach was disconcertingly fractional, and Boas would eventually give up his initial attempt to reduce the data he was presented with to some universal transmission pattern (Levi-Strauss 1988). Rationalizations of myths, whether proposed by anthropologist observers, or by the observed informants themselves, were not to be trusted:

> A great many [...] important legends prove to be of foreign origin, being grafted upon mythologies of various tribes. This being the case, I draw the conclusion that the mythologies of the various tribes as we can find them now are not organic growths, but have gradually developed and obtained their present form by accretion of foreign material. Much of this material must have been adopted ready-made […]. We are, therefore, led to the conclusion that from mythologies in their present form it is impossible to derive the conclusion that they are mythological explanations of phenomena of nature […], but that many of them, at the place where we find them now, never had such a meaning. If we acknowledge this conclusion as correct, we must […] admit that, also, explanations given by the Indians themselves are often secondary, and do not reflect the true origin of the myths. (Boas 1896, 5)

What is remarkable about Boas's "statistical inquiry" into myth is that it did not rest content with just collecting and reproducing mythical material. In order to be useful for the kind of comparative and critical analysis that Boas accomplished, this material had to be accompanied by information on how myths were produced and communicated in the places and communities from which they were originally recorded. As Stocking Jr (1974, 8) has argued, for Boas, integration of data "was not a matter of necessary or logical relations of elements." He favored "historical integration" instead, notwithstanding, or rather, precisely because of his statistical approach.

3 Boas's Data Sheets

From Boas's anthropometric surveys, a large number of original data sheets have been preserved in the archives of the American Museum for Natural History and the American Philosophical Society in Philadelphia (Jantz et al. 1992, 437). Many of these pertain to Native American tribes, and were produced in anthropometric campaigns carried out in 1891 and 1892 by Boas in preparation of an exhibition on physical anthropology he had been commissioned to organize for the World's Columbian Exposition, which was held in Chicago in 1893 (Jacknis 1985).

In this section, I am going to offer a description and detailed analysis of a small sub-set of these data sheets in order to reconstruct not only what data Boas collected,

but also how he did it, paying particular attention to the role that field observers, but also informants, played in the generation of the data. The subset in question is preserved at the American Philosophical Society, and pertains to Chickasaw individuals living in Stonewall and Tishomingo in the Indian Territory, now Oklahoma.[1] The Chickasaw had been forced to remove from the Southeastern Woodlands in 1832, and decided to settle with a closely related tribe, the Choctaw, in the Indian Territory. By the 1850s they had established settlements, including Tishomingo as the capital, and successfully resisted subsequent attempts to merge them with the Choctaw, forming a polity of their own to this day (St. Jean 2011).

The data sheets consist of forms printed on both sides that were filled out by hand (see Fig. 2). At the front top of the form, the field observer is asked to "[n]umber each record and write your name after number." From this, we know that one "Richard T. Buchanan," who indeed entered a serial number for each record taken, collected the data for Chickasaw.[2] The form consists of three sections: a first one

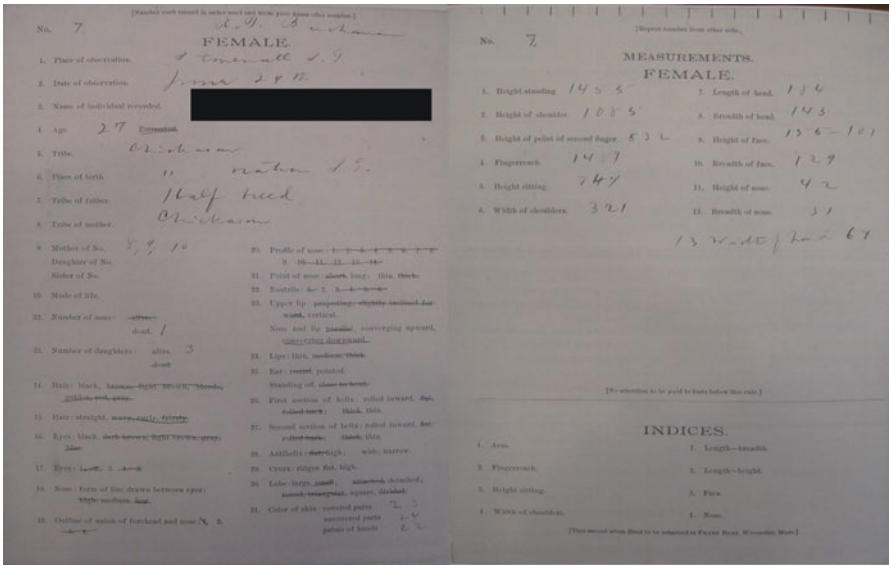


Fig. 2 Recto and verso of the data collection forms used by Franz Boas in anthropometric surveys in 1891 and 1892. Franz Boas field notebooks and anthropometric data, American Philosophical Society, Box 2, Anthropometric Data Sheets Recorded at Stonewall and Tishimingo, Indian Territory (Oklahoma). The name has been blackened by the author for anonymization. With kind permission by the American Philosophical Society

[1] Franz Boas field notebooks and anthropometric data, American Philosophical Society, Box 2, Anthropometric Data Sheets Recorded at Stonewall and Tishomingo, Indian Territory (Oklahoma).

[2] I have been unable to identify this person. It is not unlikely, though, that Buchanan was a local resident of Tishomingo; findagrave.com lists several gravestones on Tishomingo cemeteries from the nineteenth century that show the last name Buchanan, and a transcription by Tom Blake of the "largest slaveholders from the 1860 slave census schedules" in Chickasaw County lists a T. J.

providing metadata in the form of "Place" and "Date of observation" as well as information on the person observed: "Name of individual recorded," "Age," "Tribe," "Tribe of father," "Tribe of mother," relationships to other persons recorded ("Mother of … Daughter of … Sister of …"), "Mode of Life" and finally, "Number of sons" and "daughters". This is followed by a section that records a large number of qualitative physical traits, such as hair form or eye color, by offering default descriptive categories for selection. In the case of skin color, a color chart seems to have been used, as indicated by the numerals used to describe this parameter. Only then, on the verso side of the sheet, follow anthropometric variables in a separate section entitled "Measurements." The first six of these refer to overall stature, followed by six further measurements taken on head, face and nose. A third and final section is entitled "Indices," and separated from the rest of the form by a horizontal line above which it states that the field observer is not to pay "attention … to lines below this rule." As we will see further below, this section was reserved for Boas to process the data given by the measurements. The form ends with the prompt "This form when filled to be returned to Franz Boas, Worcester, Mass." There are separate forms for males and females, since some of the kin designations used, as well as some of the qualitative traits differ by gender (male forms ask for detailed information on "Beard," for example).

What is particularly striking in the series of filled out forms is that a lot of effort was spent on ascertaining the genealogical relationships between recorded individuals. Alongside relatively straightforward parental and sibling relationships, the categories of "Tribe of father" and "Tribe of mother" provide the most intriguing information in this respect. As one might expect, for each and every one of the individuals measured, "Chickasaw" is stated for the tribe he or she belongs to. Yet, answers to the questions relating to the "tribe" of their mother and father reveal very complex, mixed ancestral backgrounds. "Half-breed" is a frequently recurring designation. On the sheet reproduced in Fig. 2, for example, it is given as an answer for "Tribe of father" while the mother is stated as being "Chickasaw." If one looks at the records of the children of the female in question (sheets no. 8, 9 and 10), which record "$^{3}/_{4}$ Chickasaw $^{1}/_{4}$ white" for the tribe of mother, "half-breed" reveals itself as referring to individuals with one parent of Native American descent and one parent of European descent. Many sheets also record mixed Chickasaw and Choctaw, or "Choctaw half-breed," ancestry (sheet no. 14). The Chickasaw had been a slave-owning tribe, and while in contrast to the Choctaw they did not adopt their freedmen after emancipation in 1863 (St. Jean 2011, ch. 3), one does find quite a number of sheets were the tribe of father or mother is stated as "Chickasaw and negro" (e.g. sheets no. 2–3 and 18–19). Such assessments of mixed ancestry could reach considerable complexity. On one sheet, the tribe of mother is recorded as "$^{2}/_{3}$ Chickasaw $^{1}/_{3}$ white" (sheet 35). The only way to make sense of this proportion is to note that cousin marriage in the parental generation reduces the number of

Buchanan with 63 slaves (see http://freepages.genealogy.rootsweb.ancestry.com/~ajac/mschickasaw.htm; accessed 19/09/2018).

grandparents to six individuals, and to assume that ancestry may have been described with reference to the grandparents.

Although relatively little genealogical information is asked for by the form itself, the answers thus allow to carry out an almost complete analysis of kin relations within the cohort studied that reaches back to the grandparental, and in some cases, great-grandparental generation. Husband-wife relationships are not recorded, but can be inferred, though slightly tediously, by comparing number of children and ancestry. Conveniently, the data seems to have been recorded household by household, so that sheets for parents and their children often follow each other consecutively (e.g. sheets no. 7–10). Boas's anthropometric field campaigns were probably modeled on the 1890 US Census, which also asked for genealogical ("relationship to head of family") and racial information ("whether white black mulatto, quadroon, octeroon, Chinese, Japanese, or Indian"). He could thus assume that both his field observers, and their informants, were used to these kind of questions, the logic of genealogical analysis they presupposed, and the procedure of filling in a questionnaire.[3]

The apparent discrepancy between assigning "Chickasaw" as the tribe of persons recorded, and the mixed ancestry of their parents, is easily explained. The Chickasaw were organized by exogamous matrilineal clans, and both tribal affiliation and belongings were passed on along maternal lines (Champagne 1992, 40–41). One can therefore safely assume that the persons recorded regarded not only themselves, but also their parents as Chickasaw, as long as their parent's mothers in turn were Chickasaw. This raises the suspicion that the information on mixed ancestry might actually not have been provided by them, but by the observer. That this is not the case, however, is evident from occasional notes that the observer jotted down in the space left in the form between "Measurements" and the section "Indices," which was reserved for Boas's calculations. In these notes, Buchanan expressed doubts about the information filled in under "Tribe of father" and "Tribe of mother". Sheet no. 26, for example, where tribe of father is given as "½ Chickasaw ½ Choctaw" and tribe of mother as "Chickasaw" carries the following statement on its verso side: "The gentleman says he has heard two parents say that they had some French blood in them. He shows white blood." In another case, where tribe of both father and mother is stated as "Chickasaw," Buchannan added the note "Negro blood appears in complexion … and in shape of face" (sheet no. 53). Tribal (or racial) affiliation seems occasionally to have been a matter of negotiation between observer and informant, but in the end, the latter seems to have had the last word when it came to filling in the answers on the top front of the sheet.

[3] The form used in the 1890 United States Census is available at https://en.wikipedia.org/wiki/1890_United_States_Census#/media/File:1890B.jpg. On the complex history of racial categories in US censuses between 1850 and 1930, see Hochschild and Powell (2008). Boas himself will have had personal experience of the Prussian censuses which according to historian Christine von Oertzen developed into a data-driven exercise moving enormous amounts of paper in the 1860s and 70s; see von Oertzen (2017).

This reflects the role of mere "recorders" that Boas assigned to his field assistants. The sections of the forms dedicated to qualitative traits and anthropometric measurements leave no freedom to add personal observations. It is well known that Boas especially trained the 50 or so assistants that collected data for him in preparation of the World's Columbian Exhibition. He also modified the instruments used for measurements (Boas 1890), restricted himself to measurements where "the starting points are easily ascertained," and had the assistants perform measurements on each other, or let two observers take measurements on the same set of persons; all this in order to "reduce the personal equation, as far as possible, to a minimum" (*die persönliche Gleichung möglichst auf ein Minimum zu reduzieren*). In addition, he restricted his survey to measurements that could be performed "without disrobement" (*ohne Entkleidung*), since this would "necessarily limit the number of measured individuals" (Boas 1895b, 367). Minimizing the personal equation and maximizing population number thus actually lead to a very impoverished ontology. Observers were reduced to deleting descriptive categories prescribed on the form, and filling in a small number of mechanical measurements in generating data on the survey subjects.

Yet Boas was able, as we will see in the next section, to make a lot out of his data. Hints at how he proceeded in this can be found in the subset of data sheets on the Chickasaw. Only about a third of these show entries in Boas's hand in the section headed "Indices." Many of these entries calculate the cephalic index, i.e. the ratio of breadth of head to the length of head, and some of them the facial index in addition. What is striking about these entries is that they exclusively appear on data sheets on which the tribe of both father and mother is stated as "Chickasaw" and/or "Choctaw." In processing data, Boas apparently proceeded by grouping the data sheets, in this case separating sheets on individuals of "pure" Native American descent from those on individuals with mixed racial backgrounds. And in making this decision, Boas exclusively trusted the genealogical information that informants provided. The data sheets mentioned above, on which Buchanan had expressed his doubts about possible admixture of "French" and "negro blood," were included in the set of sheets that Boas processed in order to calculate the cephalic index. Even here, the observer's personal expertise was erased in favor of "self-identified" tribe or race, as we would say today.

4 Use and Re-use of Data

Forensic anthropologist Richard L. Jantz, who has probably done more than anyone else for recovering Boas's data from the obscurity of historical archives, expresses great admiration for the "incredible computational feat" that Boas achieved by computing "the means of height and cranial index for some 4000 individuals distributed over 60 tribes, all with pencil and paper" (Jantz 2003, 279). In this, he is referring to the only article in which Boas summarized results from the anthropometric survey carried out in preparation of the World's Columbian Exhibition. It appeared in the German Journal for Ethnology (*Zeitschrift für Ethnologie*) in 1895, and made

liberal use of tables and curve diagrams to synthesize the findings, some of which had probably already been presented in the Physical Anthropology Department of the Exposition.[4]

Boas admitted right away in this article, that the qualitative data he had collected varied too much between observers to deliver comparable results. The article therefore focused exclusively on body stature and head form, but considered these two variables not only in populations of Native American adults, but in addition in children and in "mixed bloods between Indians and other races, especially whites" (Boas 1895b, 367). The first table spanned four pages, and showed the number of individuals measured, averages, and percentaged distribution of stature in steps of 1 cm for 62 "tribes" in columns that were roughly arranged by geographic location on an East to West axis. This table was complemented by "curve plates" (*Kurventafeln)* that showed the distribution for each individual "tribe" (see Fig. 3). Boas then proceeded

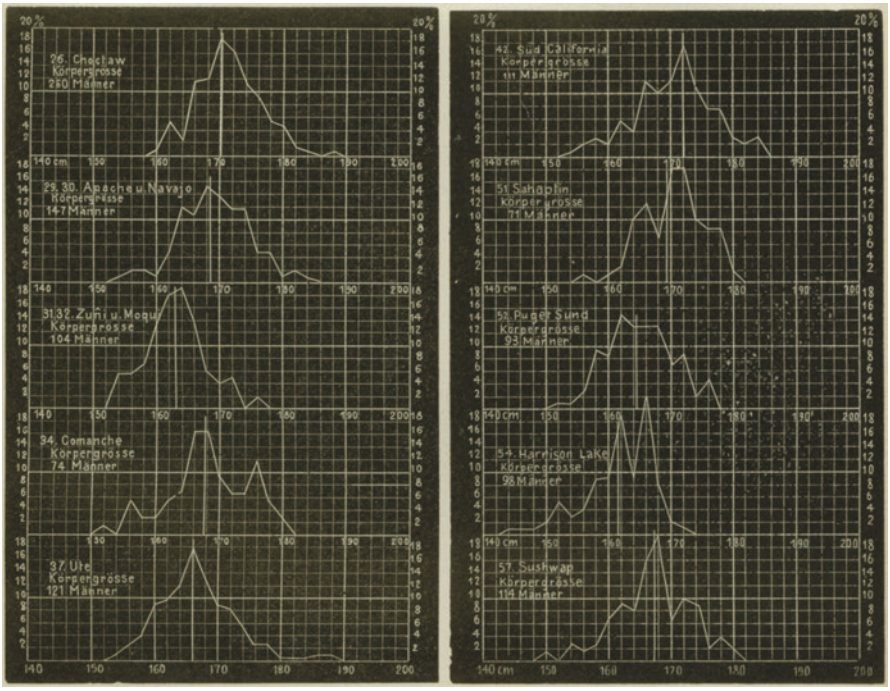


Fig. 3 Curve diagrams showing distribution of body height in several North American "tribes". The vertical axis gives percentage, the horizontal axis body height in centimeters. From Franz Boas, "Zur Anthropologie der nordamerikanischen Indianer." *Zeitschrift für Ethnologie* 27 (1895b), p. 373

[4] A guide to the Exposition states on its Ethnology Department: "For those who incline to this field of investigation, a section is devoted to physical anthropology. Here, in the skulls, *charts*, *diagrams, and models* gathered from many nations may be compared the past and present types of the human races" (Bancroft 1893, 629).

to break down his overall population, as well as populations constituting individual "tribes," by age, gender and racial descent ("full-blooded" [*Vollblut*-] vs. "half-blooded Indians" [*Halbblut-Indianer*], ibid. p. 381) in ever more complex ways. The same procedure was then repeated for head form (cephalic index) and breadth of face. He found evidence, that both parameters were influenced by environmental and hereditary factors (Ibid., 376). Their distribution at the West Coast in particular showed similar complex geographic patterns as the ones he observed in his linguistic and mythological studies, but notably without simply reproducing the latter (Ibid., p. 402). One feature that particularly fascinated Boas was that distribution curves of "half-blooded" individuals did not show simple blending of parental types, but usually two maxima indicating a "law of inheritance" according to which a "reversion (*Rückkehr*) generally occurs towards the parental form" (Ibid., 406). He pursued this topic by looking at the distribution of breadth of face. By "classifying mixed-bloods in such a way, that one group includes individuals which have more than half of Indian blood, and other individuals, which have half or less Indian blood" he even tried to demonstrate that the "Indian type" was characterized by a "stronger hereditary force" (*grössere Vererbungskraft*) with regard to this trait (see Fig. 4).

Jantz, like many others, has claimed that the 1895 article is the only one in which Boas presented and analyzed his data in detail (Jantz 2003, 279). This is not quite true. Rather, Boas seems to have re-used the data quite opportunistically in a number of publications to make particular points. Already in 1891, he published a very concise paper in the journal *Science* in which he argued for reversion, rather than blending, in human inheritance based on cephalic index data from "Oregonian Athapascans," "Northern Californians" and their "crosses" (Boas 1891). In 1894, he published an article entitled "The Half-Blood Indian: An Anthropometric Study" that made many of the points, and contained many of the illustrations, of the 1895 article, but started off with a curve diagram showing "number of children of Indian Women and of Half-Blood Women" in order to disprove the common belief that "hybrid races show a decrease in fertility" (Boas 1894, 2). And what is perhaps Boas's most important anthropometric paper, a critique of the significance of the cephalic index for indicating human types, was also based on data he had gathered in his field campaigns, now of course relating to "full-blooded" individuals, because the critique could otherwise easily have been fended off by maintaining that stability of type is generally compromised by mixture. In all of these cases, the relatively impoverished base of data was compensated by the myriad ways in which it was classified with respect to information collected on the measured individuals – place of birth, age, gender, and ancestry, in particular. While the data on physical traits covered few properties only, this data revealed that the populations under scrutiny were rich in structures that could be deployed again and again to answer different research questions relating to the role of environment and inheritance.

After 1900, Boas's interest in the physical anthropology of Native Americans seems to have dwindled. The number of preserved observations abruptly drops to 2 only in 1901, and then there are none for the remaining years. The reasons for this may have been political: With the Jim Crow laws and legislation increasingly enforcing allotment of tribal land to individual tribe members who could prove their "purity of blood" (Curtis Act 1898), having one's ancestry "questioned" was increas-

mm	Vollblut 157	3/4-Blut 85	3/8-Blut 73
124—125	0,4	—	—
126—127	0,2	—	—
128—129	—	—	—
130—131	—	—	—
132—133	—	—	0,9
134—135	1,1	0,8	1,4
136—137	2,8	2,8	3,7
138—139	4,4	7,2	7,3
140—141	3,8	6,3	12,8
142—143	8,9	6,7	20,1
144—145	13,6	9,1	14,2
146—147	15,3	13,9	11,4
148—149	14,0	16,3	8,7
150—151	11,2	16,3	10,0
152—153	6,4	9,9	5,9
154—155	6,4	7,1	1,4
156—157	6,6	1,2	1,4
158—159	3,6	0,4	0,9
160—161	0,6	1,6	—
162—163	—	0,4	—
164—165	0,4	—	—
166—167	0,2	—	—

Fig. 4 Table showing distribution of breadth of face for Ojibwa-men of different racial ancestry ("fullblooded", "3/4 blooded" and "3/8 -blooded"). From Franz Boas, "Zur Anthropologie der nordamerikanischen Indianer." *Zeitschrift für Ethnologie* 27 (1895b), p. 410

ingly becoming a highly delicate matter (see St Jean 2011, 55, for the Chickasaw). The fraught relationship between Native Americans and their "scientific observers" that this new situation must have created continues to this day and was exacerbated by the highly publicized conflicts around large-scale human genetics projects such as the Human Genome Diversity Project and the Genographic Project in the early 1990s (Reardon and TallBear 2012). It is therefore not surprising that the data from Boas's anthropometric survey have been eagerly taken up by anthropologists in past decades. While I have not found any direct mention of these conflicts in the sources I have worked with, it is revealing that one of them mentions in passing that "Boas's data offer the *only* opportunity for systematic examination of anthropometric variation among North American Indians" (Jantz et al. 1992, 456; my emphasis).

A team of postgraduate students and researchers around Jantz was the first to convert Boas's data into a "computerized database", retaining data for "individuals

who could be considered full-blooded" only, and replacing "obvious outlying values" with values "predicted from all others" following accepted statistical procedures not available to Boas (Jantz et al. 1992, 439, 442). Their analysis revealed that the data showed "strong geographic patterning" supporting "climate-morphology correlations" with exception of head-shape which showed "considerable intertribal variation" (ibid., 457). Lyle W. Konigsberg and Stephen D. Ousley – noting their gratefulness to Boas for "what, for the time, was an unusual inclusion of pedigree data" (1995, 481; cf. Jantz 1995, 351) and to Jantz for granting them access to this data in its electronic form – used a small subset of the data to test an important assumption in quantitative genetics about the proportionality between phenotypic and genetic co-variation. Using a subset of data for five "tribes", and normalizing it for sex and age, their mathematically sophisticated paper provides an impressive example for the degree to which Boas's data rendered itself amenable to the application of complex genealogical matrices (ibid., 484–485). Yet another research agenda was pursued by economic historians who drew conclusions about the historical development of nutritional status and living standards among nineteenth-century First Nations by looking at the variation of body height across time and across tribes, again thanking Jantz for granting them access to the data (Steckel 2010, 267; Carlson and Komlos 2014, 158).

The end of studies on Native Americans in 1901 did not mean the end of Boas's interest in physical anthropology. Instead, he changed subject. Reducing the number of anthropometric variables even further, he carried out an anthropometric survey on some 16,000 immigrants from Eastern Europe and Italy and their children in order to determine whether the new environment they entered resulted in a change of physical type (Boas 1912). With their obvious political significance – Boas's conclusions became part of the idea of America as a "melting pot" –, these studies as well have invited reanalysis again and again, especially since Boas took the unusual step to publish his raw data (Boas 1928). R. A. Fisher was among those who re-used this data to throw doubt on Boas's conclusions. Part of the argument pertained to the quality of Boas's data; Fisher and his collaborator Horace Gray, a medical doctor from Stanford University Hospital, suspected that it was compromised by wrongly reported paternity and inter-observer variability. This did not keep them, however, from subjecting it to Fisher's "method of analysis of variance" for the purpose of making this point by demonstrating that variability and regressions within families did not meet expectations informed by "previous biometrical work" (Fisher and Gray 1937, 92). An earlier study by Geoffrey Mackay Morant, a student of Karl Pearson, and Otto Samson had followed a similar strategy, arguing that Boas's results had been confounded by variation in age and sex while using his published raw data as evidence in favor of precisely this claim (Morant and Samson 1936).

Fisher and Gray's doubts let me return to the recent debate about whether Boas got it "right or wrong" with which I opened this chapter. The debate was sparked by another paper by Jantz, co-authored with a former graduate student of his department, Corey S. Sparks, in which the authors tested what they took to be the central conclusion of Boas's immigrant study, namely that it demonstrated "the plastic nature of the human body in response to changes in the environment." They did so

by reassessing his data "within a modern statistical and quantitative genetic framework", in particular "using pedigree information contained in Boas' data [to estimate] narrow sense heritability". The outcome was negative, with results indicating "very small and insignificant differences between European- and American-born offspring, and no effect of exposure to the American environment on the cranial index" (Sparks and Jantz 2002, 14, 636).

Unbeknownst to Sparks and Jantz, three other researchers had been carrying out a similar re-analysis on Boas's published data, the results of which they published in the March 2003 issue of *American Anthropologist*. "Using methods unavailable to Boas," just like Sparks and Jantz were doing, medical anthropologist Clarence C. Gravlee and his co-authors were led to the opposite conclusion, namely that "modern analytical methods provide stronger support for Boas's conclusion than did the tools at his disposal" (Gravlee et al. 2003a, 125). In the ensuing exchange between the two sets of authors, which was published in the June issue of *American Anthropologist*, some degree of reconciliation was reached by agreeing that Boas's claim that human head form changed with immigration was generally confirmed by his data, but that doubts remained regarding the biological significance of these changes and the nature of the causes responsible for them (Gravlee et al. 2003b, 331; Sparks and Jantz 2003, 335). What is notable about this reconciliation is that it did not hinge so much on the data used, than on the questions being asked from it. Gravelee et al. had set out to test claims that Boas had expressly made, whereas Sparks and Jantz questioned a common assumption about these claims that over the 90 years that had passed since Boas study had become part and parcel of disciplinary lore and that they considered "a burr in our bed for 90 years" (Holden 2002).

5 Conclusion

If we consider the "journey" of Boas's data as a unit of analysis, as suggested by Sabina Leonelli in the introduction to this volume, it is a journey in which the body of Boas's data as a whole did not remain untouched. Quite on the contrary, that body of data was variously partitioned, cleansed of outliers, adjusted for confounding variables like age or sex, and processed by a bewildering range of statistical procedures to produce ever new numerical and visual representations.[5] In part, as evidenced by the re-use of Boas's immigrant data, these renewed analyses of historical data sets were motivated by the impact that his critique of the race concept had on the disciplinary self-understanding of American anthropologists. The relevance of Boas's data may hence be seen to be partly due to their direct relevance for a framework of concepts and theories that had been travelling alongside them through the twentieth century. But even those who disagreed with this framework,

[5] On "data cleaning", see Boumans and Leonelli's contribution to this volume. The chapters by Porter and Bechtel, both in this volume, discuss the importance of visualizations as a medium of data travel.

like Fisher or Jantz, made use of the data, and it was also used by researchers in other disciplines, like quantitative genetics or economic history. What generated this astounding surplus of Boas's data in terms of usability in a variety of theoretical and disciplinary contexts?

My case study suggests two points in response to this question. The first concerns the importance of what I suggest to call "pattern data" for making data relevant to a variety of contexts. These are data that do not describe single, manifest properties of individual entities under scrutiny, but rather relationships among them, and hence precisely occupy the middle ground between data and metadata that I have outlined at the outset of this article. Typically, they relate to categories that allow researchers to group the subjects under study, and hence the data produced about them, in a variety of ways that are believed to be of causal relevance for similarities exhibited among these subjects.[6] Thus, Boas's anthropometric surveys are not only renowned today for their sheer scale, but also for their careful design which included collection of basic geographical and genealogical information on observed individuals (Jantz 2003, 280).[7] This information allowed Boas to classify the data that he had collected on a modest number of anthropometric variables in ways that enabled him to address an array of questions in his publications, and also explains why later researchers could turn to his data again and again to carry out new research. The poverty of data on physical traits collected by Boas, that is, was compensated by the richness of pattern data that allowed for meaningful classification (on the significance of data classification, see also Leonelli 2012; Müller-Wille 2018).

However, this richness – and this is my second point – depended on information that was provided in situational contexts in which the measured individuals themselves took on an active role, rather than simply being the passive subjects of measurement procedures. The "pattern data" Boas used in his anthropometric survey, that is, were "given" in a literal sense; in contrast to the data on stature and head form, which was extracted from individuals in a more or less mechanical manner, information on age, sex, birth place, next-of-kin, as well as tribal and racial affiliation had to rely on interviews, and was hence partly informed by common-sense notions of the persons observed. While these categories proved to be an extremely versatile tool for classifying the data in ever new ways, it also irretrievably tied it to the historical context of its production. Especially tribal and racial affiliation are categories the meaning of which, at any given point in time, has been molded by centuries of political struggle and whose application will continue to be of political relevance.

The tens of thousands of datasheets that Boas took care to preserve in his papers, and that are still accessible to researchers, thus does not only form a repository of data to be explored scientifically for what it tells us about the physical appearance

[6] In this, I take inspiration from, but do not strictly follow, the American botanist and geneticist Edgar Shannon Anderson, who distinguished between "pattern data" and "pointer data" in an intriguing article arguing for the continued relevance of natural history in genetics (Anderson 1956; see Kleinman 2016). The case of radiocarbon dating and the necessity to calibrate it by other, "relative" dating technologies described by Wylie (this volume) provides a comparable case.

[7] Ramsden provides a fine case study in this volume on the efforts that went into the design of survey technologies to satisfy the needs for date of housing planners in mid-twentieth century United States.

and genetic constitution of historic populations. Every single sheet also gives us glimpses of the life story of an individual person, and the collection of datasheets as a whole therefore forms a historical archive in its own right that can also be used to reconstruct the power relations that informed the original surveys. It is therefore unlikely that any answer to the question whether Boas got it "right or wrong" will ever bring the journey of his data to a an end. They will remain relevant as long as the historical circumstances under which they were produced, and the intervention that Boas and his collaborators made on these circumstances through their surveys, have historical bearing for the present situation.

References

Anderson, Edgar. 1956. Natural History, Statistics, and Applied Mathematics. *American Journal of Botany* 43 (1956): 882–889.

Bancroft, Hubert Howe. 1893. *The Book of the Fair: An Historical and Descriptive Presentation of the Worlds' Science, Art and Industry, as Viewed Through the Columbian Exposition at Chicago in 1893*. Chicago: Bancroft Co.

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.

Boas, Franz. 1887a. The Occurrence of Similar Inventions in Areas Widely Apart. *Science* 9 (224): 485–486.

———. 1887b. Museums of Ethnology and Their Classification. *Science* 9 (229): 614.

———. 1890. A Modification of Broca's Stereograph. *American Anthropologist* 3 (3): 292–293.

———. 1891. Mixed Races. *Science* 17 (425): 179.

———. 1894. The Half-Blood Indian: An Anthropometric Study. *Popular Science Monthly* 45: 761–770.

———. 1895a. *Indianische Sagen von der Nord-Pacifischen Küste Amerikas*. Berlin: A. Asher.

———. 1895b. Zur Anthropologie Der Nordamerikanischen Indianer. *Zeitschrift Für Ethnologie* 27: 366–411.

———. 1896. The Growth of Indian Mythologies. A Study Based upon the Growth of the Mythologies of the North Pacific Coast. *The Journal of American Folklore* 9 (32): 1–11.

———. 1897. The Social Organization and the Secret Societies of the Kwakiutl Indians. In *Report of the National Museum for 1895*, 313–738. Washington: Government Printing Office.

———. 1911. *The Mind of Primitive Man*. Boston: MacMillan.

———. 1912. Changes in the Bodily Form of Descendants of Immigrants. *American Anthropologist, New Series* 14 (3): 530–562.

———. 1928. *Materials for the Study of Inheritance in Man. Vol*, Contributions to Anthropology, 6. New York: Columbia University Press.

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.

Carlson, Leonard, and John Komlos. 2014. The Anthropometric History of Native Americans, C.1820?1890. *Research in Economic History* 30: 135–161.

Champagne, Duane. 1992. *Social Order and Political Change: Constitutional Governments Among the Cherokee, the Choctaw, the Chickasaw and the Creek*. Stanford: Stanford University Press.

Dall, William. 1887. Museums of Ethnology and Their Classification. *Science* 9 (228): 587–589.

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

Fisher, R.A., and H. Gray. 1937. Inheritance in Man: Boas's Data Studied by the Method of Analysis of Variance. *Annals of Eugenics* 8: 74–93.

Gravlee, Clarence C., H. Russell Bernard, and William R. Leonard. 2003a. Heredity, Environment, and Cranial Form: A Reanalysis of Boas's Immigrant Data. *American Anthropologist* 105 (1): 125–138.

Gravlee, Clarence C., H. Russell Bernard, and William R. Leonard. 2003b. Boas's 'Changes in Bodily Form': The Immigrant Study, Cranial Plasticity, and Boas's Physical Anthropology. *American Anthropologist* 105 (2): 326–332.

Hochschild, Jennifer L., and Brenna Marea Powell. 2008. Racial Reorganization and the United States Census 1850–1930: Mulattoes, Half-Breeds, Mixed Parentage, Hindoos, and the Mexican Race. *Studies in American Political Development* 22 (1): 59–96.

Holden, Constance. 2002. Going Head-to-Head Over Boas's Data. *Science* 298 (5595): 942–943.

Jacknis, Ira. 1985. Franz Boas and Exhibits: On the Limitations of the Museum Method of Anthropology. In *Objects and Others: Essays on Museums and Material Culture*, ed. George W. Stocking Jr., 75–111. Madison: University of Wisconsin Press.

Jackson, John P., and David J. Depew. 2017. *Darwinism, Democracy, and Race: American Anthropology and Evolutionary Biology in the Twentieth Century*. London: Routledge.

Jantz, Richard L. 2003. The Anthropometric Legacy of Franz Boas. *Economics & Human Biology* 1: 277–284.

Jantz, Richard L., D.R. Hunt, A.B. Falsetti, and P.J. Key. 1992. Variation Among North Amerindians: Analysis of Boas's Anthropometric Data. *Human Biology* 64 (3): 435–461.

Jenkins, David. 1994. Object Lessons and Ethnographic Displays: Museum Exhibitions and the Making of American Anthropology. *Comparative Studies in Society and History* 36: 242–270.

Kleinman, Kim J. 2016. Bringing Taxonomy to the Service of Genetics': Edgar Anderson and Introgressive Hybridization. *Journal of the History of Biology* 49: 603–624.

Konigsberg, Lyle W., and Stephen D. Ousley. 1995. Multivariate Quantitative Genetics of Anthropometric Traits from the Boas Data. *Human Biology* 67 (3): 481–498.

Leonelli, Sabina. 2012. Classificatory Theory in Data-Intensive Science: The Case of Open Biomedical Ontologies. *International Studies in the Philosophy of Science* 26: 47–65.

Leonelli, S. 2014. What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data & Society* 1. https://doi.org/10.1177/2053951714534395.

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

Leonelli, Sabina. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Levi-Strauss, Claude. 1988. The Social Organisation of the Kwakiutl. In *The Way of the Masks, translated by Sylvia Modelski*, 163–187. Seattle: University of Washington Press.

Morant, G.M., and Otto Samson. 1936. An Examination of Investigations by Dr Maurice Fishberg and Professor Franz Boas Dealing with Measurements of Jews in New York. *Biometrika* 28 (1/2): 1–31.

Müller-Wille, Ludger. 2014. The Franz Boas Enigma: Inuit, Arctic, and Sciences. *Montréal: Baraka Books. (Ludger is My Father's Brother.)*.

Müller-Wille, Staffan. 2018. Making and Unmaking Populations. *Historical Studies in the Natural Sciences* 48: 604–615.

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.

Powell, John Wesley. 1887. Museums of Ethnology and Their Classification. *Science* 9 (229): 612–614.

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.

Reardon, Jenny, and Kim TallBear. 2012. 'Your DNA Is Our History': Genomics, Anthropology, and the Construction of Whiteness as Property. *Current Anthropology* 53 (S5): S233–S245.

Sparks, Corey S., and Richard L. Jantz. 2002. A Reassessment of Human Cranial Plasticity: Boas Revisited. *Proceedings of the National Academy of Sciences of the United States of America* 99 (23): 14636–14639.

———. 2003. Changing Times, Changing Faces: Franz Boas's Immigrant Study in Modern Perspective. *American Anthropologist, New Series* 105 (2): 333–337.

St. Jean, Wendy. 2011. *Remaining Chickasaw in Indian Territory, 1830s–1907*. Tuscaloosa: University of Alabama Press.

Steckel, Richard H. 2010. Inequality Amidst Nutritional Abundance: Native Americans on the Great Plains. *The Journal of Economic History* 70 (2): 265–286.

Stocking, George W., Jr. 1974. Introduction: The Basic Assumptions of Boasian Anthropology. In *A Franz Boas Reader: The Shaping of American Anthropology, 1883–1911*, ed. Georg W. Stocking Jr., 1–20. Chicago: University of Chicago Press.

———. 1983. *Race, Culture, and Evolution: Essays in the History of Anthropology*. Chicago: University of Chicago Press.

von Oertzen, Christine. 2017. Machineries of Data Power: Manual Versus Mechanical Census Compilation in Nineteenth-Century Europe. In *Elena Aronova, Christine von Oertzen, and David Sepkoski*, ed. Data Histories, vol. 32, 129–150. Osiris/Chicago: University of Chicago 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.

Xie, Yu. 1988. Franz Boas and statistics. *Annals of Scholarship: Metastudies of the Humanities and Social Sciences* 5: 269–296.

Staffan Müller-Wille is University Lecturer in the Department for History and Philosophy of Science, University of Cambridge. From 2007 to 2019, he was Associate Professor in the History and Philosophy of the Life Sciences and Deputy Director of Egenis, Centre for the Study of the Life Sciences, University of Exeter (England). He also holds an Honorary Chair at the Institute for History of Medicine and Science Studies of the University of Lübeck. He received his PhD from the University of Bielefeld (1997) and subsequently worked as Scientific Curator at the German Hygiene Museum in Dresden (1998–2000) and the Max Planck Institute for the History of Science in Berlin (2000–2004). His research covers the history of the life sciences from the early modern period to the early twentieth century, with a focus on the history of natural history, anthropology and genetics. His more recent publications include an article titled "Names and Numbers: 'Data' in Classical Natural History, 1758–1859" in *Osiris* (Vol. 32, 2017) and a coauthored book on *The Gene: From Genetics to Postgenomics* (University of Chicago Press, 2018, with Hans-Jörg Rheinberger).

Radiocarbon Dating in Archaeology: Triangulation and Traceability

Alison Wylie

Abstract When radiocarbon dating techniques were applied to archaeological material in the 1950s they were hailed as a revolution. At last archaeologists could construct absolute chronologies anchored in temporal data backed by immutable laws of physics. This would make it possible to mobilize archaeological data across regions and time-periods on a global scale, rendering obsolete the local and relative chronologies on which archaeologists had long relied. As profound as the impact of ^{14}C dating has been, it has had a long and tortuous history now described as proceeding through three revolutions, each of which addresses distinct challenges of capturing, processing and packaging radiogenic data for use in resolving chronological puzzles with which archaeologists has long wrestled. In practice, mobilizing radiogenic data for archaeological use is a hard-won achievement; it involves multiple transformations that, at each step of the way, depend upon a diverse array of technical expertise and background knowledge. I focus on strategies of triangulation and traceability that establish the integrity of these data and their relevance as anchors for evidential reasoning in archaeology.

1 The Quest for an Absolute Chronology

If any data are "tragically local" (Latour 1999: 59), the fragmentary traces that make up the archaeological record would seem to fit the bill. Detached from the originating cultural events and contexts of interest to archaeologists, subject to the vagaries of preservation and the contingencies of recognition and recovery as a "record," archaeological data are often seen as presenting insurmountable obstacles to their use as anchors for evidential claims about the past.[1] A major breakthrough, foundational to the formation of archaeology as a field, was development in the nineteenth century of methods for discerning the temporal structure of the material record

[1] See Currie's reprise of and rebuttals to arguments that give rise to such pessimism (2018, chapter 4).

A. Wylie (✉)
Department of Philosophy, University of British Columbia, Vancouver, BC, Canada
e-mail: alison.wylie@ubc.ca

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_15

(Trigger 1989). The most influential of a number of chronological systems dating to this period – the "Three Age System" developed in the 1830s by Thomsen in Denmark and by Nilsson and Worsaae in Sweden – posited a cultural sequence of stone, bronze and iron ages based on the observation that, in undisturbed deposits, artifacts of these materials regularly occur together and in stratigraphic sequence (Trigger 1989: 124; Rowley-Conwy 2007: 32–47). Drawing on geological principles of superposition these assemblages were interpreted as chronological markers (Renfrew 1973: 24). To extend these sequences beyond the locales where they were established, archaeologists built fine-grained stylistic seriations that capture the orderly succession of form and design within classes of artefacts found in stratified deposits (e.g. Deetz and Dethlefsen 1967); artifacts of a similar material and design could be compared across sites and slotted into a design sequence presumed to hold for a cultural tradition or horizon. These attributions of "age" were, however, relative and of limited scope so, where possible, archaeologists made use of textual or epigraphic records to tie chronologies based on artifact typologies, seriation and stratigraphy to historically documented events and, thus, to one another. For example, coins and inscriptions testify to Roman presence in geographically distant regions, establishing (respectively) the earliest and latest dates at which the material associated with them could have been deposited. They also made use of dendrochronology and varve dating (annual sequences of tree-rings and glacial lake deposits) to anchor cultural to physical chronologies, but these too were of limited scope. The challenge was to link up chronologies of limited reach so that the trajectory of culture-transforming processes – the spread of farming, migrations and trade relations, the expansion and contraction of cultural spheres of influence – could be traced through space and time.

When radiocarbon dating was introduced in the early 1950s it was hailed as the solution to a range of chronological problems in archaeology; indeed, many expected that it would render obsolete these longstanding methods of constructing relative chronologies. The principle is straightforward. Radioactive carbon isotopes decay at a stable rate – their half-life is ~5730 years – so if you know the ratio of radioactive (^{14}C) to stable carbon (^{12}C and ^{13}C) in the environment in which a sample of organic material originated, you can use the difference between the proportion of radioactive carbon in the sample and this baseline ratio to estimate the time elapsed since "sample death" (Hamilton and Krus 2018: 198): the point at which the organic source of the sample stopped absorbing carbon and the decay process began. As Libby described the temporal data that could be captured by this means, the crucial warrant for its use as the anchor for an absolute chronology is the stability of the process of radioactive decay, a physical process that is not affected by other properties of the sample itself or its geological, much less its cultural, context.

> The rate of disintegration of radioactive bodies is extraordinarily immutable, being independent of the nature of the chemical compound in which the radioactive body resides and of the temperature, pressure, and other physical characteristics of its environment. (Libby 1952: 9, as cited by Francis 2002: 297)

By contrast to the temporal data on which archaeologists had relied, this measurable ratio of time-dependent radioactive to stable carbon clearly seemed to qualify

as "mobile immutables" in Latour's sense (Latour 1999; see also Morgan 2008, 2011). And, indeed, radiocarbon dating has had a profound impact on archaeology; in a recent retrospective Manning describes it as having "entirely restructured the practice and understanding of prehistoric archaeology around the world" (2015: 128). That said, the process of realizing its promise as a game-changing innovation has been a long, tortuous process. It is now described as proceeding through three radiocarbon revolutions (Manning 2015), each of which addresses distinct challenges posed by the multiple transformations involved in capturing, processing, packaging and interpreting radiogenic data for use in archaeological contexts.[2] In the process the Latourian ambitions that attended its initial introduction have been significantly rescaled. The radiogenic data made available by these successive revolutions is anything but "raw"; the ongoing process of refinement, calibration and interpretation affirms the robustly relational conception of data that frames this volume. I focus on two aspects of the transformations required to mobilize these data for archaeological purposes: the role of mediators, in the form of the inferential warrants and scaffolding of various kinds that make it possible to constitute material traces as temporal data; and the strategies archaeologists use to ensure the integrity of these data and, crucially, their credibility as anchors for evidential reasoning relevant to archaeological inquiry.

My aim here is to illustrate the irreducibly relational nature of data in this context where, at its inception, the radiocarbon revolution seemed poised to fulfill the most unqualified of foundationalist ambitions. I will identify a great many different kinds of objects and claims that function in archaeological contexts as data, extending an account I have developed elsewhere for a relational conception of evidence (Wylie 2011a; Chapman and Wylie 2016). On this view evidential claims are the terminus of practical arguments that, as characterized by Toulmin, originate with some "fact" or "datum" and are mediated by warrants that licence the inferential move from datum to conclusion (Toulmin 1958: 98, 218–221; Chapman and Wylie 2016: 34–36). Whether a claim counts as a mediating warrant or an evidential claim is a function of its role in an evidential argument; the material warrants that figure prominently in archaeological contexts are themselves the terminus of evidential arguments. The same is true of "data"; I concur with Leonelli that what counts as a "datum" is a function of its potential for use as evidence (Leonelli 2015, 2016) and that data are never simply "given"; they are themselves the terminus of an extended process of practice and inference that configure them as useable in a particular research context. In the case of radiocarbon data, literal journeys are involved; material traces are excavated, transported, curated, processed and incorporated into chronological models, and then put to work as archaeological evidence in a great many different contexts. But what I focus on here are primarily journeys across methodological frames. This account is itself a chronological; I chart the process by

[2] See Chapman and Wylie (2016: 147–151) for a more detailed account of the complex story of enthusiasm and ambivalence, institutional manoeuvring and competition that characterize the history of radiocarbon dating; we compare this with the reception and life history of other "external resources" imported to archaeology in recent decades.

which radically diverse types of expertise and bodies of background knowledge were brought together to refine the techniques and establish the standards that configure evolving practices for handling radiogenic data in archaeological contexts.

2 Capturing Radiogenic Data

Within a decade of the initial introduction of radiocarbon dating – the first radiocarbon revolution set in motion by Libby in the 1950s – it had become clear that its successful application to archaeological material would require a great deal of technical scaffolding. This first revolution is defined by two sets of issues: the need to establish archaeological field practices for recovering and handling samples that minimize contamination by younger or older organic material, and to refine the methods by which radiocarbon laboratories measure ^{14}C in archaeological samples.

In a Latourian analysis that provides a useful framework for considering the first of these challenges, Lucas characterizes archaeological fieldwork as an iterative process of intervention on field sites and materials – practices of disaggregation and assembly – by which an archive of material, both "found and made," is assembled in ways that are configured by the anticipated requirements of data mobility. Invoking Latour (1999), he observes that "it is precisely what is portable or mobile that…defines the archive" (2012: 244). Field sites are, in a literal and documentary sense, standardized to approximate the material form of the archive, creating legible assemblages that can be "carried over" from the field to other sites of knowledge making (2012: 230, 234, 244). Lucas's primary examples of these archive-producing processes are site survey, excavation and recording practices that are standardized within and across sites (at least, within research traditions), and designed to facilitate the translation of objects and observations "from one material form into another": "the way we intervene with [a site] is set up precisely for the manner in which we [will] read it in translation" (2012: 238–239). So, for example, the practice of excavating in stratigraphic levels, cleaning exposed features and preparing the vertical walls of excavation units is keyed to photographic documentation, and to drawing plan views and stratigraphic profiles; the site is prepared, "sculpted," so that it can be read "as if it were a drawing" or a photograph (2012: 239).

The archive in Lucas' sense is, then, an active construct, designed to encode information about context and associations that will make it possible to retrace the steps by which the contents of the archive were produced, linking material samples and artefacts, drawings and photographs, field records and notes to one another and to features of their depositional context long after they have been removed, textually translated, and dispersed to distant museums, labs, offices and classrooms. Latour describes exactly this process in connection with the stratigraphic drawings created by the team of field biologists, ecologists, and soil scientists he observed in Brazil (1999: 57–58), and it figures in Bouman and Leonelli's account of "data cleaning" – a practice documented within archaeology by Gero (2007). Traceability is crucial, especially when the field interventions are destructive, as in the case of excavation.

It is what makes possible the "iterative" analysis of an archaeological site that, on Lucas' account, constitutes the mobility of the archive; it enables archaeologists to reassess, reposition and reinterpret the data that make up an archive, and sometimes to extract from it entirely new and unanticipated data (2012: 234; Wylie 2011b; 2016). Traceability is the key to establishing "sample-to-context" relationships that make the results of radiocarbon analysis useable in archaeological contexts – a fraught set of issues that have come into sharp focus in the last few decades and a point to which I return shortly.[3] But first, consider in a bit more detail the translational processes by which radiocarbon data – the ratios of ^{14}C to ^{12}C and ^{13}C in archaeological samples – are generated.

Radiocarbon dating was initially applied to organic artefacts of known age held in well documented museum collections (e.g., Egyptian funerary furnishings). But as it became more widely available archaeologists reconfigured their field practices to anticipate the requirements of a new network of data-generating sites, specifically, radiocarbon dating laboratories. Material they had not routinely collected or that had been of marginal interest took on new significance – fragments of wood and bone, seeds and grains, the non-artefactual contents of storage and fire pits – and questions about sample collection, storage, and transport had to be addressed. As radiocarbon dating techniques evolved, the range of materials and the size of samples viable for dating changed, sometimes dramatically. With the use of Accelerator Mass Spectrometry (AMS) – based on direct detection of radiocarbon atoms – it is possible to work with samples as small as 20 milligrams, compared to the typical requirement of 10–100 grams for radiometric methods (Bronk Ramsey 2008: 258–259). At the same time the list of contaminants to be avoided has expanded from the obvious – cigarette ash and campfire charcoal – to include, for example, various types of glue, paper and cardboard that incorporate polyvinyl acetates; skin creams and lubricants in which polyethylene glycol is an ingredient; hydrocarbon-based fuels; and pesticides (biocides). Fieldworkers are advised to use glass or aluminium containers, but the specifics vary depending on type of sample, storage conditions and lab protocol; not surprisingly, given its greater precision, AMS dating is especially sensitive to contaminants.

Alongside the standardization of protocols for the recovery and handling of datable samples, radiocarbon laboratory techniques for processing them also had to be refined to control for a range of other confounds: the second set of issues that had to be resolved. These include, for example, the effects of elecromagnetic impurities, ambient radiation, radon contamination and fractionation in reactions that do not go to completion, and the need to standardize count-time and conventions for calculating and reporting margins of error. By the early 1980s protocols ensuring inter-lab reliability had been instituted, but in a review of *Radiocarbon After Four Decades* (Taylor et al. 1992), Browman observed that, while "error magnitude is no longer linked clearly to lab type," differences in the standards employed by different

[3] Shavit and Griesemer's account of "locality in biodiversity surveys" (2009) illustrates just how complex traceability to "locality" can be, even when the original methodologies of data capture are not destructive (2009).

laboratories were still an issue (1994: 378). Fifteen years later Bronk Ramsey could report that "the measurement stage of the process is no longer the most critical element in determining precision and accuracy, except for the very smallest samples" (2008: 259), but problems persisted with the pre-treatment of samples. In short, fine-tuning laboratory protocols to ensure the reliable translation of radiocarbon samples from field to laboratory has been a long and ongoing process.

As these challenges were met, a growing number of anomalies were identified in the ^{14}C dates reported for archaeological material that could not be attributed to contamination or processing error. These drew attention to the complexity of the physical processes that radiocarbon dating exploits; much more background knowledge is required to estimate time elapsed since sample death than the "immutable" decay rate of radioactive carbon. In short, it was the interpretation of radiocarbon ratios as temporal data that came into sharp focus as needing attention. It was this recognition that set in motion the second radiocarbon revolution (Manning 2015: 129): a long process of calibrating radiocarbon dates that began in the mid-1960s.

3 Calibration: Refinement and Conversion

The second radiocarbon revolution was catalysed by two concerns: that, even if the half-life of radioactive carbon is stable, the ratio of ^{14}C to ^{12}C and ^{13}C in the atmosphere is not necessarily uniform over time or space; and that plants and animals take up carbon in different ways which affect its concentration in their tissues. Together these raise questions about what baseline should be used in determining how long the ^{14}C in a particular sample had been decaying. These were first addressed in connection with the "industrial" and "bomb" effects. By mid-century the widespread use of fossil fuels had dumped an enormous amount of "dead" carbon into the atmosphere, depressing the proportion of radioactive to stable carbon isotopes, while Cold War era nuclear bomb tests had "almost doubl[ed] the concentration of radiocarbon in the atmosphere" (Bronk Ramsey 2008: 251; Gillespie 1986: 20). In the event, the global standard, as "agreed internationally by the radiocarbon community," was the average count rate for terrestrial wood dating to 1950, a choice of baseline described in the 1986 Oxford *Radiocarbon User's Handbook* as "arbitrary"; "other values could have been used with perhaps more theoretical justification" (Gillespie 1986: 18).

Establishing a global convention for calculating elapsed radiocarbon years was just a beginning. What has ensued is a process of identifying and compensating for more localized effects of sample context and composition that has depended on recruiting an enormously wide range of substantive background knowledge about the "radiocarbon life cycle": how carbon is produced, dispersed, and sequestered in diverse local environments, and how it is taken up and fixed by different types of organism (Bronk Ramsey 2008: 249–252). Where baseline carbon ratios are concerned, the complications now recognized include, for example, variation over time in the rates of radiocarbon carbon production in the upper atmosphere, which is an

effect of sunspot activity and dipole movement. This can have an impact on temperature which, in turn, affects the mixing and circulation of atmospheric ^{14}C as well as its rate of absorption into carbon reservoirs. The most significant reservoirs are marine; the rate at which radiocarbon is exchanged with the surface ocean is much slower than its dispersal in the atmosphere, and slower again in deep ocean reservoirs. A "marine offset" affects organisms sequestered in carbon sinks created by ocean currents where the proportion of radiocarbon may be considerably lower than in the atmosphere, the radioactive carbon in such an environment having decayed without being replenished. The atmospheric ratio also varies temporally and regionally. By the early 1980s it was recognized that there is a hemispheric difference in the concentration of ^{14}C, given proportionately more ocean surface in the southern than the northern hemisphere (Browman 1981: 249–67; Gillespie 1986: 26–7). In addition, as recently as 2001 two articles that appeared in *Science* reported that "a regional, time-varying ^{14}C offset can occur *within* a hemisphere" (Kromer et al. 2001; Manning et al. 2001; Reimer 2001). Wood samples from Anatolia and southern Germany, dated to the fifteenth through the seventeenth centuries AD on the basis of tree-ring sequences, had produced radiocarbon dates that diverged as much as 200 years. This discrepancy was attributed to a solar minimum that raised ^{14}C levels in the atmosphere, depressing radiocarbon relative to calendric ages, and an associated cooling effect that had seasonally different impact on trees with different growth periods (Kromer et al. 2001: 2529–30; Manning et al. 2001: 2533).

The challenges of determining baseline ratios of radiocarbon concentration for the environments in which organic samples originated is further complicated by an appreciation that processes of carbon uptake differ by type of organism. This has implications for how samples should be processed and how their measured carbon ratios should be interpreted. For example, plants that take up carbon directly from their environments have different concentrations of ^{14}C depending on whether they are terrestrial or marine, that is, whether they absorb carbon in the form of carbon dioxide or as bicarbonate. If they are terrestrial, uptake depends on the photosynthetic pathway by which they fix carbon, which differs between arid, succulent, and temperate zone plants. Radiocarbon concentrations also differ between herbivores that ingest photosynthesized carbon directly, and carnivores that get their carbon by "a more circuitous route through the food chain" (Bronk Ramsey 2008: 253). In addition, their metabolic processes may discriminate against heavy isotopes (e.g., in bone collagen) or affect the absorption of carbon by specific types of tissue (e.g., horns and nails do not continue to absorb carbon once formed).

Far from providing an autonomous and incontrovertible empirical foundation for archaeological chronologies, radiocarbon data are the conclusions of extended practical arguments that depend upon a great deal of contingent and, I will argue, local scaffolding. To be sure, the data that anchor these arguments are measurements of the carbon content of archaeological samples. However, as the process of second revolution calibration makes clear, they are only usable as a source of temporal data – that is, an estimate of time elapsed since sample death – given a complex of chain of inferences that take into account the conditions of their recovery, transport, storage, processing, and the technical details of radiometric or AMS analysis. The

inferential moves by which these measurements are converted into temporal data depend, in turn, on an immense array of mediating warrants: substantive background knowledge drawn from organic as well as physical chemistry, atmospheric science, geology, marine and terrestrial biology, to name just a few of the fields that were enlisted in the process of standardizing analytic procedures, controlling for confounds, and establishing computational and reporting conventions for radiocarbon data.[4] I use the terminology of "warrants" in the sense suggested by Toulmin (1958), to refer to all the background knowledge and assumptions that license inferences from an originating observation or measurement, mark or inscription (a "datum" on his account), to a conclusion that, in this case, takes the form of a claim about the estimated time elapsed since "sample death".

This emphasis on the substantive nature of these warrants resonates with Norton's arguments for recognizing, more generally, that inductive inference is mediated by domain-specific "material postulates" (Norton 2003: 648). In a similar spirit Woodward insists that the assumptions "required to license…reliable inference from data [to phenomena]" are "empirical," not "matters of stipulation" (2011: 172, 175). Alongside examples drawn from chemistry (determining the melting point of lead) and neuroscience (smoothing fMRI readings), he cites the assumptions on which archaeologists depend to infer temporal data (the date of a fossil) from radiocarbon decay counts: for example, "the way in which soil conditions and atmospheric exposure may affect the presence of carbon" (2011: 172). As he argues, it does not follow from the fact that such assumptions "'go beyond the data'" that they are "arbitrary, empirically unfounded, untestable, or matters of stipulation or convention" (2011: 173). The credibility of the data claim – that a given observation or measurement tracks a phenomenon of interest – depends upon the credibility of these mediating warrants. As Woodward also notes, the epistemic goals of inquiry and "attitudes toward risk" are also constitutive of these arguments (2011: 172, 174). So, for example, the claim that the radiocarbon content of an organic sample should be recognized as archaeological data depends not only on the background knowledge about confounds and offsets but also, prospectively, on its potential to serve as the point of departure for further inferences that support evidential claims about the age of a cultural feature, deposit, or site – the phenomena of interest to archaeologists.

For radiocarbon data to fulfil this function – to anchor a chronological claim that can serve as archaeological evidence – the crucial contribution of the second radiocarbon revolution has been the development of finely tuned calibration programs based on datasets that integrate the most sophisticated knowledge available about offsets and confounds of the sort I have described. To identify sources of error and correct for them archaeologists routinely rely on strategies of triangulation. They may, for example, compare carbon isotope ratios measured in material of archaeological interest against samples of "known age" that come from the same

[4] In "Circulating Reference" Latour remarks that "one science always hides behind another," registering some disappointment that the Brazilian fieldwork he observed did not, in fact, represent "the birth of a science *ex nihilo*" (1999: 32). What I foreground here is this networked interdependence among fields that comprise the trading zone in which archaeology operates (Chapman and Wylie 2016: "Archaeology as a Trading Zone," chapter 4).

(or relevantly similar) environments. The determination of "known age" may also depend on historical chronologies and on dendrochronology as well as, in some cases, stratigraphic sequences and typological chronologies – precisely the sources of temporal data that radiocarbon dating was meant to displace. The Southern German/Anatolian case mentioned above illustrates how this works in the case of dendrochronology. The annual accretion of tree-rings yields patterned sequences that can be stitched together across species and regions, so that radiocarbon dating of these samples can provide a temporal (usually decadal) profile of regional fluctuations in atmospheric radiocarbon. Varved lake sediments can support similar analyses that extend beyond the temporal reach of dendrochronological sequences. These local baseline data make it possible to refine the radiocarbon-based calculations of the time elapsed since sample death, but by no means do they resolve all the anomalies that signalled the need for calibration. At this point several ^{14}C calibration systems are available (e.g. CALIB 7.1, Stuiver et al. 2018; OxCal 4.3, Bronk Ramsey 2018). As these have been refined, "wiggle effects" have been identified such that, for some periods of archaeological interest, samples with different true ages correspond to the same radiocarbon ages, or the spread in their true ages is exaggerated, compressed, or even reversed. Here is Bronk Ramsey's appraisal of the achievements and limitations of second revolution calibration techniques:

> The problems of variable radiocarbon content in the atmosphere distort and defocus our view of the passage of time. The statistical methods now available to deal with calibrated dates act like a corrective lens to overcome these problems. However, with this clearer image other problems are also thrown into sharper focus: the statistical methods do not overcome methodological shortcomings in the radiocarbon method itself. (2008: 265)

The upshot is that, to use radiocarbon data as the basis for an "absolute" chronology – a temporal framework that, in the ideal, extends to the whole of the global archaeological archive – it has been necessary to rely on a system of warrants that effectively add contextual data back in and are valuable precisely because they are local and limited in their mobility. This predicament of locality – that secure anchoring to the local is a condition of mobility – is by no means unique to archaeology. Norton makes the point in general terms. The 'portability' of the material postulates that mediate inductive inference is invariably limited; they underwrite inference only within fields where the regularities and causal dynamics they capture can be shown to obtain (2003: 663).

4 Traceability and Triangulation

The catalyst for a third radiocarbon revolution, associated with a program of "Bayesian" chronological modelling (Bayliss and Whittle 2015),[5] is the further realization that various forms of "tragically local" data (Latour 1999: 59) are indispensable not

[5] Bayliss and Whittle describe this as a "pragmatic" Bayesian approach to archaeological problems (personal communication, 2014). Their central point, which resonates with Manning's appraisal (2015), is that any assessment of the evidential bearing of radiocarbon data on questions about

only to ensure accuracy in the translation of radiogenic into temporal data, but also when it comes to transforming temporal into chronological data that can be used to address archaeological questions. The challenge here is to determine how a measure of time elapsed since the physical event of sample death relates the timing of cultural activities that are responsible for the production, use and deposition of the organic material from which samples are drawn. This is a problem that no amount of technical refinement – in standardizing sample collection and measurement practices, or in calibrating the translation of radiocarbon ratios into time scales – can resolve. As Manning describes this current and on-going revolution, it marks a decisive shift away from the quest for temporal data that approximate an ideal of absolute immutability and unconstrained mobility. Advocates of Bayesian approaches give up the epistemic ambitions that animated earlier revolutions; rather than expecting ^{14}C dating to deliver foundational, physics-backed temporal data that can displace reliance on context-specific resources, they embrace a commitment to "fully integrate archaeological information with ^{14}C dating," including the "web" of background assumptions underlying relative chronologies (Manning 2015: 151). The emphasis in this third revolution is on integrating radiogenic data into chronological models that are archaeologically meaningful.

Whether the target of inquiry is an individual artefact or feature, a single short episode of use or occupation, a sequence of occupational layers in a densely stratified site, or a regional cultural formation that extends over millennia, the first step in the process of transforming temporal into chronological data is to assemble and appraise a set of ^{14}C dates that are potentially relevant to archaeological questions about age and chronological sequence. Traceability is crucial here. The determination of which samples to date when an archaeological archive is being created, and the choice of ^{14}C dates to include in a chronological model, depends on an appraisal of their provenance and integrity. Hamilton and Krus emphasize the need for a "holistic understanding" of the archaeological and geological context in which a sample originated that requires "at the very least…a description of the dated sample, the specific laboratory methods, and the sample's provenience in relation to the archaeological features" (2018: 193). In the case of legacy data this appraisal sometimes involves quite literally retracing the steps of those who originally recovered a sample back into the field or to the repositories and laboratories to which finds and records were dispersed, reconstructing a record of the context from which it was drawn and the processes by which it was transformed into radiogenic data (Wylie 2011b: 312). Unless these data journeys can be reconstructed – unless the chains of recovery, transport, transformation, inscription, and relocation are "reversible," as Latour puts it (1999: 61) – the samples have little value as a source of temporal data that can anchor archaeologically relevant evidential claims. Done well, this is a process of source criticism that exploits traceability as a means of making explicit

archaeological chronology must take into account how well supported a chronological model is on other grounds (its prior probability), as well as the degree to which these data support lines of evidential reasoning that are discriminating with respect to the plausibility and accuracy of the model (an appraisal of the prior and posterior likelihood of the data that anchors evidential claims).

and appraising the warrants that underpin attributions of integrity to individual samples and trustworthiness to the data claims based on them (Wylie 2011b).

In addition to traceability, triangulation provides a further check on accuracy and allows for a closer specification of the date ranges generated by ^{14}C analysis. For example, when archaeologists aggregate ^{14}C dates, rather than just calculating a mean or median date for the data set, they sometimes model the range of dates a hypothetical sample would generate, given standard margins of error, if the actual date of sample death was this calculated mean, a strategy of internal triangulation that can delimit the dispersion of pooled or averaged dates (Chapman and Wylie 2016: 152; Shott 1992: 221–223). Typically, however, triangulation strategies make use of radiogenic data drawn from different sources to cross-check the accuracy of individual ^{14}C dates and the credibility of the assumptions that inform the construction of chronological models. This may involve testing multiple samples from a single artefact or feature, sometimes submitting them to different laboratories, to control for contamination and laboratory error, or testing different types of samples drawn from a single depositional context to control for biases that can arise from relying on one type of material. It may also involve dating non-cultural, botanical and ecological samples that originated in the same environment as cultural samples in order to cross-check assumptions about baseline carbon ratios (Hamilton and Krus 2018: 195). Latour seems to have such strategies in mind when he mentions, in passing, a field practice of cross-field triangulation whereby the geomorphologist on the Brazilian field crew "adds her two cents to all the conversations, allowing her expatriate colleagues to 'triangulate' their judgments through hers" (1999: 47). Here credibility is a function of the capacity of these different types of radiogenic data to constrain one another, exposing sources of error that may not be identifiable by tracing data journeys and assessing the security of warrants for individual (calibrated) ^{14}C dates.

More expansive strategies of triangulation typical of this third ^{14}C dating revolution depend on mobilizing a range of different, non-radiogenic types of temporal data. Given practices of reuse, curation, trade and other forms of circulation that complicate the life histories of organic material in cultural contexts, datable samples may come from organisms that were cut, butchered, burned or otherwise taken out of the carbon cycle long before they were deposited in the archaeological contexts from which they are recovered. To establish a connection between the ^{14}C-datable natural event of their death and the cultural target of interest to archaeologists, a premium is put on drawing samples from organic remains that can be assumed to be "functionally related to their deposit" (Hamilton and Krus 2018: 194), to have originated in a short timeframe, or to derive from a temporally ordered sequence of deposits. Articulated animal bone or undisturbed human burials are examples of the former; geologically sealed cave deposits and the association of human remains or artefacts with extinct mega-fauna are a classic example of the latter (Chapman and Wylie 2016: 35), as are stratigraphic associations more generally: the location of a sample in relation to stratified occupational levels may set temporal bounds on its age in relation to other datable samples. The stylistic homogeneity of the artefact assemblages with which a sample is associated, and comparanda from related sites that support the seriation of particular types of artefact or feature, can also be used to establish contemporaneity or temporal sequence (Chapman and Wylie 2016: 151–155).

The point of recruiting these diverse types of data is to re-embed the much-manipulated ^{14}C mobiles in a local context of inquiry, delimiting the range of physically possible dates and margins of error generated by radiocarbon analysis and, crucially, integrating discrete traces, features, and sites into an archaeologically plausible chronology. As the number of distinct types of data built into these models is expanded, so too is the range of background knowledge – the substantive warrants – that are required to secure the inferences that link the temporal claims they support to an archaeological target. This vastly complicates the construction of chronological models, but it is also a source of epistemic credibility. The principle at work here is that the likelihood of spurious convergence on a specified range or sequence of dates is much reduced when mediating warrants are drawn from diverse sources and the material they configure as data are themselves generated by different causal processes. The credibility of the resulting chronological models is not just a function of the aggregation of individual data points or sets assessed as trustworthy; it arises from the collective capacity of these data to reinforce and to constrain one another.

5 Robustness Reasoning About Temporal Data

The strategies central to these practices of chronological modelling are recognizably a genre of "robustness" reasoning, as Wimsatt has described the diverse methods of "multiple determination" that he finds ubiquitous across the sciences (Wimsatt 1981: 123–4, 2012; Soler 2012: 3). In this case they are applied to the kind of problem Hacking explores in connection with microscopy (Hacking 1981, 1983: 186–203). They are meant to ensure that the heavily scaffolded temporal data archaeologists rely on do, in fact, track the cultural phenomena of interest, counteracting the risk that they are artefacts of, or otherwise distorted by, practices of extraction and measurement, processing and packaging for travel as elements of an archaeological archive. I have argued elsewhere that, in constructing evidential claims, archaeologists routinely exploit the causal and epistemic independence between distinct lines of evidence that originate in a common target of inquiry (Wylie 2011a: 387–389). The strategies of triangulation characteristic of the second and third radiocarbon revolution suggest that this is true, as well, of ^{14}C data. To use a metaphor of Norton's (2014: 673), the empirical objects and claims that comprise the data recruited in support of various components of a chronological model are reciprocally strengthened by being bound into "highly connected, massively tangled" and self-stabilizing systems of data-cum-evidence.

Strategies of multiple determination, coupled with traceability, can certainly mitigate the risk that convergence is spurious when diverse types of data and the evidential claims they anchor come together in support of a coherent chronological model. Nonetheless the worry remains that, absent "absolutes" in the form of immutable temporal data that can function as a decisive, wholly autonomous arbiter of chronological questions, there is an inherent nepotism in the process of mutual adjustment required to calibrate temporal data and integrate them into archaeological chronologies. The strategies for identifying, controlling and correcting for error

developed in the course of successive radiocarbon revolutions suggest four conditions that data-evidence tangles must meet if the risk of vicious, rather than virtuous, patterns of self-stabilization is to be avoided.

The first condition is a requirement that the source data and the warrants backing data claims be "secure": each, taken on its own, must be well substantiated in terms of the best technical and theoretical expertise available in the fields that make possible their capture and mobilization. This was the central preoccupation of the first radiocarbon revolution in which techniques for reliably measuring radiocarbon ratios in organic samples were the focus of attention. It figures, as well, in the long process of calibrating radiogenic data against a much expanded range of background knowledge about the nature of the samples, their contexts of origin, possible confounds that affect the measurement of carbon ratios, and their translation into the temporal scale of elapsed calendar years.

The second condition is a requirement of causal and conceptual independence between the various types of temporal data that are used to calibrate one another and to build chronological models. In the ideal, any given tangle of interlinked chains of data-cum-evidence should incorporate data that have causally distinct "life histories," and the warrants mediating the various transformations these data and their use as evidence should derive from conceptually independent research traditions. At their most effective, the triangulation strategies that figure prominently in the second and third radiocarbon revolution meet exactly this requirement.

By extension of this second condition, when one type of data is used to calibrate another, the tuning of measurement systems and the refinement of the warrants that underpin them should be justified on substantive grounds, not just because they ensure convergence. Manning describes several cases in which this was a central consideration in the process of reconciling early Cycladic and late Bronze Age Aegean chronologies with sequences of radiocarbon dates (2015: 142–150), as do Bayliss and Whittle with reference to chronological models of artefact and occupational sequences of different scales (2015: 222–230). A striking example of such reasoning recently analysed by Bokulich (forthcoming) is the decision, in 2012, to base a significant revision of the Geological Time Scale on an independent, non-radiometric measure of geologic time – a choice explicitly informed by a concern to preserve the independence of the two radiometric methods that are typically used to cross-check one another in geochronological dating.

The trajectory of the multiple radiocarbon revolutions makes it clear that traceability as well as triangulation is required. The usefulness of ^{14}C data depends on their mutability, which means that they are vulnerable to error and distortion in the course of their journeys. Detailed documentation and ongoing critical scrutiny of the transformations that comprise these journeys is crucial and, in fact, an explicit demand for traceability is a recurrent theme in the literature on chronological modelling. Hamilton and Krus emphasize the need for "transparency" with respect to model structure and the "choices and assumptions" that inform its construction (2018: 195); the hypothesized relationship between sample dates and the dates of a target event should be clearly specified, and the basis for these assumptions – background knowledge about the archaeological context and mediating warrants – should be made explicit.

A final condition might be described as a requirement of epistemic democratization: a normative implication of the relational account of data. Assessments of security and strategies of triangulation can justify privileging some types of temporal data over others as inherently more trustworthy, accurate and/or precise. However, none should be presumed to be empirically foundational "immutables," exempt from re-examination when discrepancies arise in data-evidence tangles, or when the process of retracing data journeys brings to light previously unrecognized confounds or as yet unaddressed uncertainties. This is the central motivation for the third radiocarbon revolution: that however compelling their physics-backing may be, ^{14}C data must be accountable not only to standards of credibility in their field of origin but also those that are specific to their contexts of use. This norm underwrites a commitment to treat even the most promising "silver bullet" techniques of data extraction and mobilization as tentative, the starting point for a process of epistemic iteration in which it is expected that they will be subject to continuous refinement and sometimes replacement as an evolving empirical scaffolding for inquiry (Chang 2004: 43).

These are demanding ideals, rarely fully met in practice. Nonetheless, I suggest that they are orienting norms of practice exemplified by, and responsible for, the considerable achievements of the successive radiocarbon revolutions and that have unfolded since the 1950s.

6 Conclusion

What exactly are the data in this sprawling story of extraction, processing and packaging, calibration and circulation by which radiogenic data are captured and integrated into chronological models in archaeological contexts? There are the organic artefacts and residues that survive *in situ* or are curated in the archaeological archive from which datable samples are retrieved; the carbon extracted from these samples; the isotope ratios produced by means of AMS or decay counts; the calibrated estimates of radiocarbon years elapsed since sample death; the translation of these radiocarbon dates into calendar years; and then their interpretation as dates when organic elements of the archaeological archive were created, used, and deposited. All of these constitute the "data," now repeatedly transformed, that figure as the starting point for the evidential reasoning that grounds cultural/historical chronologies. There are also all the ancillary data that back the substantive assumptions – the warrants – that mediate each step in these tangled chains of reasoning from and about the material samples, test results, and records that comprise the archaeological archive.

I submit that all of these count as data. Their status as data is a function of their role in anchoring practical arguments for a range of different types of evidential claim, not an intrinsic quality of 'givenness', closeness or similarity to the target of inquiry, much less their status as self-warranting or empirically foundational. The framing argument for recognizing that data are relational in this sense has been made by Leonelli (2015: 817, 2016: 79), and the recognition that they are as hard-won an achievement as the evidential claims they support figures centrally in the philosophical and science studies sources on which I have drawn, diverse as they are. It is also a recurrent theme in internal discussion of the vagaries of archaeologi-

cal inquiry. In addition to Lucas' account of "the archaeological record," Chippindale urges his fellow archaeologists to adopt the term "capta" rather than "data" (2000), a sentiment that resonates with Latour's admonition that one should "never speak of 'data' – what is given – but rather of 'sublata', that is, of achievements" (1999: 42).

An implication of this relational view is that the data that anchor evidential arguments are themselves the terminus of further practical arguments that depend upon their own warrants; as such, their points of origin, and each of the steps involved in capturing and transforming them into useable data are also subject to critical scrutiny, and open to demands for further backing. In the case of archaeology, building these tangles of practical argument is an achievement that depends on a genre of robustness reasoning; it is a matter of enlisting not only the data generated by physical dating techniques but also a wide range of less transportable, context-specific data. The epistemic integrity and credibility of the resulting temporal data is a function of the traceability of these transformations, a point that Latour acknowledges when he considers their "reversibility" (1999: 59, 74), not the immutability of these mobiles that he otherwise emphasizes. Bronk Ramsey captures this point when he observes that, as radiocarbon dating has become "markedly more precise (and hopefully not less accurate) we need to be even more careful…about the chain of reasoning that allows us to go from a radiocarbon measurement to an understanding of chronology" (2008: 266):

> Radiocarbon dating should not be viewed as a black box, which occasionally has to be shaken because it does not give the right answer. (Bronk Ramsey 2008: 270)

References

Bayliss, Alex, and Alasdair Whittle. 2015. Uncertain on Principle: Combining Lines of Archaeological Evidence to Create Chronologies. In *Material Evidence: Learning from Archaeological Practice*, ed. R. Chapman and A. Wylie, 213–242. London: Routledge.

Bokulich, A. forthcoming. Calibration, Coherence, and Consilience in Radiometric Measures of Geologic Time. *Philosophy of Science*.

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.

Bronk Ramsey, C. 2008. Radiocarbon Dating: Revolutions in Understanding. *Archaeometry* 50: 249–275.

———. 2018. *OxCal version 4.3*. Oxford Radiocarbon Accelerator Unit. https://c14.arch.ox.ac.uk/oxcalhelp/hlp_contents.html. Accessed 10 May 2018.

Browman, David L. 1981. Isotopic Discrimination in Correction Factors in Radiocarbon Dating. *Advances in Archaeological Method and Theory* 4: 241–295.

———. 1994. Review of *Radiocarbon After Four Decades*, eds. R. E. Taylor, A. Long, and R. S. Kra, and *Proceedings of the 14th International Radiocarbon Conference*, ed. A. Long. *American Antiquity* 59: 377–378.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Chapman, Robert, and Alison Wylie. 2016. *Evidential Reasoning in Archaeology*. London: Bloomsbury.

Chippindale, Christopher. 2000. Capta and Data. *American Antiquity* 65: 605–612.

Currie, Adrian. 2018. *Rock, Bone and Ruin*. Cambridge, MA: MIT Press.

Deetz, James F., and Edwin S. Dethlefsen. 1967. Death's Head, Cherub, Urn and Eillow. *Natural History* 76: 29–37.
Francis, Kevin. 2002. *"Death Enveloped All Nature in a Shroud": The Extinction of Pleistocene Mammals and the Persistence of Scientific Generalists*. PhD, Program in History of Science and Technology, University of Minnesota.
Gero, Joan. 2007. Honoring Ambiguity/Problematizing Certitude. *Journal of Archaeological Method and Theory* 14: 311–327.
Gillespie, Richard. 1986. *Radicarbon User's Handbook*, Monograph # 3. Oxford: Oxford University Committee for Archaeology.
Hacking, Ian. 1981. Do We See Through a Microscope? *Pacific Philosophical Quarterly* 18: 305–322.
———. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.
———. 1992. The Self-Vindication of the Laboratory Sciences. In *Science as Practice and Culture*, ed. A. Pickering, 29–64. Chicago: University of Chicago Press.
Hamilton, Derek W., and Anthony M. Krus. 2018. The Myths and Realities of Bayesian Chronological Modeling Revealed. *American Antiquity* 83: 187–203.
Kromer, Bernd, Surt W. Manning, Peter Ian Kuniholm, Maryanne W. Newton, Marco Spurk, and Ingeborg Levin. 2001. Regional 14CO2 Offsets in the Oroposphere: Magnitude, Mechanisms, and Consequences. *Science* 294: 2529–2532.
Latour, Bruno. 1999. Circulating Reference: Sampling the Soil in the Amazon Forest. In *Pandora's Hope*, 24–79. Cambridge, MA: Harvard University Press.
Leonelli, Sabina. 2015. What Counts as Scientific Data? A Relational Framework. *Philosophy of Science* 82: 810–821.
———. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: University of Chicago Press.
Libby, Willard. 1952. *Radiocarbon Dating*. Chicago: University of Chicago Press.
Lucas, Gavin. 2012. *Understanding the Archaeological Record*. Cambridge: Cambridge University Press.
Manning, Sturt W. 2015. Radiocarbon Dating and Archaeology: History, Progress and Present Status. In *Material Evidence*, ed. R. Chapman and A. Wylie, 113–127. London: Routledge.
Manning, Sturt W., Bernd Kromer, Peter Ian Kuniholm, and Maryanne W. Newton. 2001. Anatolian Tree Rings and a New Chronology for the East Mediterranean Bronze-Iron Ages. *Science* 294: 2532–2535.
Morgan, Mary S. 2008. *'On a Mission' with Mutable Mobiles*. LSE Working Papers on the Nature of Evidence No 34/08.
———. 2011. Traveling Facts. In *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*, ed. P. Howlett and M.S. Morgan, 3–41. Cambridge: Cambridge University Press.
Norton, John D. 2003. A Material Theory of Induction. *Philosophy of Science* 70: 647–670.
———. 2014. A Material Dissolution of the Problem of Induction. *Synthese* 191: 671–690.
Reimer, Paula J. 2001. A New Twist in the Radiocarbon Tale. *Science* 294: 2494–2495.
Renfrew, Colin. 1973. *Before Civilization: The Radio Carbon Revolution and Prehistoric Europe*. New York: Penguin Books.
Rowley-Conwy, Peter. 2007. *From Genesis to Prehistory: The Archaeological Three Age System and its Contested Reception in Denmark, Britain, and Ireland*. Oxford: Oxford University Press.
Shavit, Ayelet, and James Griesemer. 2009. There and Back Again, or the Problem of Locality in Biodiversity Studies. *Philosophy of Science* 76: 273–294.
Shott, Michael J. 1992. Radiocarbon Dating as a Probabilistic Technique: The Childers Site and Late Woodland Occupation in the Ohio Valley. *American Antiquity* 57: 202–230.
Soler, Léna. 2012. The Solidity of Scientific Achievements: Structure of the Problem, Difficulties, Philosophical Implications. In *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, ed. L. Soler, E. Trizio, T. Nickles, and W.C. Wimsatt, 1–60. New York: Springer.
Stuiver, M., P. J. Reimer, and R. Reimer. 2018. *CALIB Radiocarbon Calibration Version 7.1*. Accessed 10 May 2018. http://calib.org/calib/.

Taylor, R.E., A. Long, and R.S. Kra, eds. 1992. *Radiocarbon After Four Decades: An Interdisciplinary Perspective*. New York: Springer.
Toulmin, Stephen E. 1958. *The Uses of Argument*. Cambridge: Cambridge University Press.
Trigger, Bruce G. 1989. *A History of Archaeological Thought*. Cambridge: Cambridge University Press.
Wimsatt, William C. 1981. Robustness, Reliability, and Overdetermination. In *Scientific Inquiry and the Social Sciences*, ed. M.B. Brewer and B.E. Collins, 124–163. San Francisco: Josey-Bass.
———. 2012. Robustness: Material and Inferential, in the Natural and Human Sciences. In *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, ed. L. Soler, E. Trizio, T. Nickles, and W.C. Wimsatt, 89–104. New York: Springer.
Woodward, James F. 2011. Data and Phenomena: A Restatement and Defense. *Synthese* 182: 165–179.
Wylie, Alison. 2011a. Archaeological Facts in Transit: The "Eminent Mounds" of Central North America. In *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*, ed. P. Howlett and M.S. Morgan, 301–322. Cambridge: Cambridge University Press.
———. 2011b. Critical Distance: Stabilising Evidential Claims in Archaeology. In *Evidence, Inference and Enquiry*, ed. P. Dawid, W. Twining, and M. Vasiliaki, 371–394. London: Oxford University Press.
———. 2016. How Archaeological Evidence Bites Back: Strategies for Putting Old Evidence to Work in New Ways. *Science, Technology & Human Values* 42: 203–225.

Alison Wylie is Professor of Philosophy and Canada Research Chair in Philosophy of the Social and Historical Sciences at the University of British Columbia. Her philosophical analyses are case-based, chiefly concerned with archaeological practice and feminist research in the social sciences, and address such questions as follows: What counts as evidence? How should we understand ideals of objectivity given the role of values and interests in inquiry? How do we make research accountable to the diverse communities it affects? Her recent publications include *Evidential Reasoning in Archaeology* (2016) and *Material Evidence* (2015), with archaeologist Bob Chapman; "How Archaeological Evidence Bites Back: Strategies for Putting Old Data to Work in New Ways" (STHV 2017); and contributions to the *Springer Handbook of Model-Based Science* (2017), *Objectivity in Science* (2015), *How Well do Facts Travel?* (2010) and *Agnotology* (2008). Her work on feminist standpoint theory includes "What Knowers Know Well" (*Scientiae Studia*, 2017) and her 2012 APA Presidential Address, and her essays on research accountability appear in "A Global Dialogue on Collaborative Archaeology" (*Archaeologies*, 2019), *The Ethics of Cultural Appropriation* (2009) and *Embedding Ethics* (2005). She is a Past President of the American Philosophical Association (Pacific Division) and Current President of the Philosophy of Science Association.

Part VI
Ends: Data Actionability and Accountability

‘Overcoming the Bottleneck’: Knowledge Architectures for Genomic Data Interpretation in Oncology

Alberto Cambrosio, Jonah Campbell, Etienne Vignola-Gagné, Peter Keating, Bertrand R. Jordan, and Pascale Bourret

Abstract In recent years, oncology transitioned from its traditional, organ-based approach to ‘precision oncology’ centered on molecular alterations. As a result, it has become to a significant extent a ‘data-centric’ domain. Its practices increasingly rely on a sophisticated techno-scientific infrastructure that generates massive amounts of data in need of consistent, appropriate interpretations. Attempts to overcome the interpretation bottleneck have led to the establishment of a complex landscape of interrelated resources that, while displaying distinct characteristics and design choices, also entertain horizontal and vertical relations. Although there is no denying that the data-centric nature of contemporary oncology raises a number of key issues related to the production and circulation of data, we suggest that the focus on data use and re-use should be complemented by a focus on interpretation. Oncology practitioners refer to data interpretation resources as ‘knowledgebases’, an actor’s category designed to differentiate them from generic, multi-purpose databases. Their major purpose is the definition and identification of *clinically actionable* alterations. A heavy investment in human curation, of a clinical rather than exclusively scientific nature is needed to make them valuable, but each knowledgebase

A. Cambrosio (✉) · J. Campbell
Department of Social Studies of Medicine, McGill University, Montreal, QC, Canada
e-mail: alberto.cambrosio@mcgill.ca; jonah.campbell@mail.mcgill.ca

E. Vignola-Gagné
Department of Social Studies of Medicine, McGill University, Montreal, QC, Canada

Science-Metrix, Montreal, QC, Canada
e-mail: e.vignola-gagne@science-metrix.com

P. Keating
Department of History, University of Quebec at Montreal, Montreal, QC, Canada
e-mail: keating.peter@uqam.ca

B. R. Jordan
ADES, Aix-Marseille Université-EFS-CNRS, Marseille, France
e-mail: bertrand.jordan@univ-amu.fr

P. Bourret
Aix Marseille Univ, INSERM, IRD, SESSTIM, Marseille, France
e-mail: pascale.bourret@univ-amu.fr

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_16

appears to have its own peculiar way of connecting clinical and scientific statements. In spite of their common goal, knowledgebases thus adopt very different approaches partly captured by the tension between trust and traceability.

1 Introduction

In March 2018, responding to a request by the US Congress, the National Institutes of Health released a draft version of its "Strategic Plan for Data Science".[1] In its drive to modernize the "Data Repository Ecosystem", the Plan introduced a distinction between *databases* and *knowledgebases*. It defined the first as "data repositories that store, organize, validate, and make accessible the core data related to a particular system or systems", and the second as warehouses that "accumulate, organize, and link growing bodies of information related to core datasets". While admitting to "a grey area … between databases and knowledgebases" and acknowledging that some knowledgebase data "may eventually harden and become core data more appropriate for a database", the document stipulated the NIH's intention to "support each separately". In other words, this was not mere semantics: it entailed organizational and financial consequences.

While most readers are doubtlessly unaware of the database/knowledgebase distinction, it came as no surprise to us. During fieldwork for this paper, numerous respondents invoked it to characterize the computerized resources they had developed to facilitate genomic data interpretation in oncology. Given oncology's pioneering role in the development of 'precision medicine', recourse to the neologism 'knowledgebase' deserves additional investigation. What does it entail and how does it relate to the molecular reconfiguration of oncology practices? More specifically, how and to what extent does the replacement of 'data' with 'knowledge' in the portmanteau word reflect actual differences in the origin, kind, and content of the information in knowledgebases? Does the 'data journey' metaphor (Leonelli this volume; Leonelli 2016; Bates et al. 2016), often used to characterize the dynamics of data repositories, continue to appropriately describe how information elicited from journal articles or databases is incorporated and organized within knowledgebases? To begin to answer these questions we need to examine how knowledgebases are located within the sequence of activities that define genomics-driven oncology, from the initial sequencing of a patient's tumor to treatment decisions. Knowledgebases are specifically geared for *data interpretation* and as such impinge directly on discussions about the actionability and clinical utility of genomic results, i.e. the establishment of predictive relations between molecular profiling results and specific drugs (Nelson et al. 2013). Oncologists perceive them as potential solutions to a major 'bottleneck' that threatens the viability of their endeavor.

[1] https://grants.nih.gov/grants/rfi/NIH-Strategic-Plan-for-Data-Science.pdf

2 The Data Interpretation Bottleneck

In his 2011 address to the American Society of Clinical Oncology, ASCO's president discussed the challenges occasioned by the rapidly decreasing price of genomic sequencing and the ensuing 'tsunami' of genomic data:

> When data are that cheap, every patient's cancer will be informative for tumor biology. And things will get very, very complicated. (George Sledge, cited in Goldberg 2011, 4).

The issue was more than quantitative. Traditionally, tumors have been characterized by organ and/or tissue of origin and stage of development. Following the introduction of genomic platforms that identify a wide range of molecular alterations (mutations, amplifications, etc.), clinical practitioners entertained the possibility of generating an alternative categorization of tumors based on shared alterations, thus "creating a new molecular taxonomy of cancer" (Titus 2014a). Early, simplistic attempts to implement genomic medicine using a 'one cancer gene, one drug' approach, have been replaced by a more detailed understanding of the molecular bases of therapies. Cancer-related genes harbor thousands of variants that require an unprecedented level of granularity in assessing their effects. The problem has thus less to do with the actual *production* of molecular data – the required logistics, their reliability and comparability across instruments – than with their *interpretation* and consequent translation into clinical practices (Jordan 2015). As one oncologist argued, "the fundamental problem is we're generating more information than we can readily interpret as individuals" (Titus 2014b).

While precision medicine has its critics (e.g., Prasad 2016; see Subbiah and Kurzrock 2017 for a rebuttal), all major cancer centers and agencies have jumped on the genomic bandwagon. Publications commonly report on the experience of implementing 'omics' approaches (Schwaederle et al. 2015; Subbiah and Kurzrock 2016; Meric-Bernstam et al. 2013; Johnson et al. 2015). Both descriptive and performative, these publications report on the 'knowledge architecture' (Amin and Cohendet 2004) instituted by leading cancer organizations to operationalize cancer genomics. They simultaneously qualify precision oncology as an endeavor that has escaped the status of mere promissory note. All major cancers have been fragmented into a growing number of rare diseases defined not only by specific genomic variants, but also by their differential reaction to a new generation of 'targeted' and immunotherapy treatments (Vignola-Gagné et al. 2017).

The new approach associates clinical oncologists and pathologists with molecular biologists and bioinformatics specialists, modifying the equilibrium between the traditional components of oncology practices. Following the sequencing of tumor samples and the identification of tumor-specific events, these events must be annotated to establish their functional significance. Potential tumor-driving events must be interpreted, prioritized, and summarized "in the context of published literature, clinical trials, and a multitude of knowledge bases" (Good et al. 2014). Clinicians then evaluate these findings by relating them to clinical data generated from the case history of a particular patient (Van Allen et al. 2013). The increasing use of large-scale approaches, such as whole-exome or whole-genome sequencing (as contrasted with limited gene panels), has made the situation even more fraught. As Ghazani et al. (2017, 787) noted:

> [A]ssigning clinical meaning to each somatic and germ-line variant, including the therapeutic, prognostic, and diagnostic implications for individual patients, poses considerable difficulties in light of the inconsistent state of genome biological annotation.

This issue has recently become known, in the actors' own words, as the 'interpretation bottleneck'.

3 Knowledgebases and Databases

Instancing "the production of dozens to thousands of potential tumor-driving events that must be interpreted by a skilled analyst and synthesized in a report", Good et al. (2014) explained that:

> Each event must be researched in the context of current literature, drug-gene interaction databases, relevant clinical trials and known clinical actionability from knowledgebases. In our opinion, this attempt to infer clinical actionability represents the most severe bottleneck of the process.

The Good et al. (2014) paper predates the NIH distinction between databases and knowledgebases by 4 years, which suggests that the distinction has been in use for some time. While the term 'database' needs no further elaboration, having entered common parlance several decades ago, the notion of knowledgebase requires explanation. Although both 'bases' act as repositories for 'data'[2] derived from published papers, conference abstracts, datasets established by large-scale collaborations, and results of tumor profiling analyses of patients enrolled in clinical trials or undergoing routine treatment, it is not clear that we are talking about the 'same' kind of information. It is similarly unsure that both bases treat data in the same way. Are we, in other words, confronted with similar data journeys, and does this metaphor actually apply to knowledgebases?

Both kinds of repositories use equivalent software tools and packages, arguably making one a mere subtype of the other. But as scores of technology studies have shown (e.g. Bijker and Law 1992), it would be simplistic to reduce devices to their technical components. Moreover, the very fact that actors differentiate between them suggests important differences. While acknowledging that many genomic resources incorporate elements from both databases and knowledgebases, Pitel (2017) reiterates the usefulness of the distinction:

> Although data and knowledge are dependent on each other, it is important to understand that data portals contain observations, like those typically seen in the results section of an article. … Knowledgebases, on the other hand, contain critically processed data, contextualized for significance and meaning, much like what you might find in a conclusion section of an article, and are often more appropriate for immediate use in clinical laboratory practice.

At this point, we could be accused of uncritically adopting the actors' categories. Social scientists often contrast native terminology with scholarly notions that enjoy epistemic privilege. A different take on this issue has been proposed by ethnometh-

[2] Adopting an ethnographic stance, we consider as data anything that actors treat as such.

odologists through the notion of 'perspicuous phenomena', i.e. "'things' (and activity settings) that re-tune our sensibilities, so that when we return to the familiar distinctions, concepts, and debates of a social science, we can read them differently" (Lynch 2009, 114). We accordingly eschew the alternative, sometimes referred to as the topic/resource distinction (Lynch 1998, 867n88), that consists in either contenting ourselves with a description of the actors' language or in developing an analytical meta-language divorced from the actors" own meanings and practices. Instead, we seek intersections between the questions that actors ask and the questions we raise, between the practical answers they provide and the theoretical framing we offer. We focus on "topics or themes which preoccupy particular groups, and which resonate with social sciences issues" (M. Lynch, personal communication). As an actor-derived categorization, the database/knowledgebase distinction can be used both by analysts as a language of description, and by concerned groups as a language for action (Lynch 1993; Hatchuel 1996).

Figure 1 depicts the funnel running from the initial sequencing to the bottleneck of interpretation by/for the clinician. It appears that much of what precedes the bottleneck (stage 5) can be categorized as the domain of databases, whereas the bottleneck and its knowledgebases interrupt the data journey. Knowledgebases come into play when oncologists receive a sequencing report listing mutations of possible clinical import. Instead of manually scouring the entire published literature for information about those mutations, they turn to one of several interpretation knowledgebases that offer a synthetic summary and description of a given variant's clinical significance. The 'product' of a knowledgebase is the interpretation itself, an assertion about the clinical actionability of a particular variant. Although it can be traced back to a specific reference (PubMed or otherwise), a given interpretation is likely to differ from those embedded in other knowledgebases for the 'same' variant. What differs is the statement or interpretation itself, the 'level of evidence' associated with a given statement, the suggested therapy or clinical action, and the references supporting the interpretation. In this context, 'the data' no longer enter, leave, or occupy space in the 'base' as immutable entities. The core content of the knowledgebase – the interpretation – arises from the knowledgebase itself wherein the data are recombined and transmogrified into interpretative statements with multiple lineages.

Practitioners contrast databases with knowledgebases in two different (albeit complementary) ways. The first claims that knowledgebases contain *interpretation-laden* and *action-oriented* data, as contrasted with *raw* data.[3] The introduction of the database/knowledgebase distinction thus reifies the content of databases as theory-neutral data unaffected by interpretation. The distinction also elevates the status of the interpretations embedded in knowledgebases as (temporarily) reliable knowledge. The second demarcation refers to the practices and goals that establish those two infrastructures, which we can for now summarize as follows: whereas databases aim at the production of resources that will be available for use by different communities of practice, oncology's knowledgebases are typically the result of initia-

[3] Arguably an oxymoron (Gitelman 2013), the term 'raw data' is easily found in scientific publications and laboratory discussions, where it makes pragmatic sense (Cambrosio and Keating 2000, 263–265).

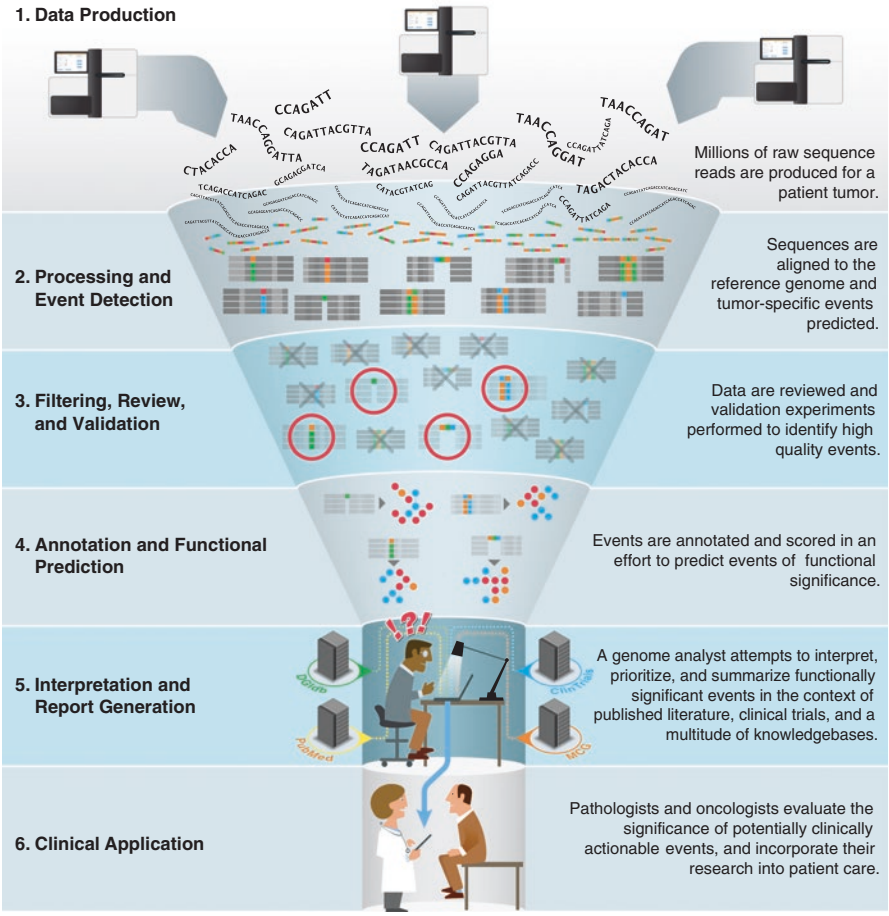


Fig. 1 "The interpretation bottleneck of personalized medicine" (Source: Good et al. 2014; Creative Commons Public Domain image)

tives derived from practical clinical concerns. As compared to much larger databases, knowledgebases address specific audiences. They are characterized by a high degree of 'situatedness' (Suchman 1987), i.e. they act as resources for clinical decision-making that are grounded in a collective understanding of possible therapeutic pathways once the local contingencies of clinical work are considered. For instance, the fact that several knowledgebases consist of an outward-facing website that only reports information with literature support, and an internal component that can exclusively be accessed by members of that institution, is justified as follows:

> In the absence of a community that understands the nuances of the potentially actionable, it's a little harder to relay that [kind of genomic] information. The treating physicians at [our institution] get a report that says: "We think this is potentially actionable because of the following reasons", and they can understand how grey that call is. That is a little more personal personalized therapy, therefore harder to do en masse, so that is indeed not reported currently on our outward-facing website. (Interview with Dr. Funda Meric-Bernstam, July 2017; henceforth FMB).

4 A Spectrum of Data Repositories

To further explore the distinction between different kinds of data repositories, consider Leonelli's (2013) analysis of an oncology database that explicitly refrained from selecting and interpreting data, namely the caBIG database, a bioinformatics initiative sponsored by the US National Cancer Institute (NCI). A key component of the 'cyberinfrastructure' destined to "empower a 'third way' in biomedical research" (Buetow 2005), caBIG was launched with great fanfare in 2003. Following recurrent criticism fueled by its overly ambitious plans, it was replaced in 2012 by a new National Cancer Informatics Program (Goldberg 2012; Thomas 2012). According to Leonelli (2013), caBIG was an "all-encompassing" database designed to provide a pluralistic community of clinical and basic researchers in oncology with easy access to a heterogeneous collection of cancer-related data. Interoperability was a key preoccupation, leaving "as much room for selecting and interpreting data as possible to their users". Otherwise put, the motley of data to which caBIG gave access had to be general enough to allow for global circulation and specific enough to fit the needs of local expert communities. A paradigmatic 'boundary object' (Star and Griesemer 1989, 393), its inability to manage this tension between two opposing demands — "fostering the global circulation of data and facilitating their local adoption" — led to caBIG's demise (Leonelli 2013). The relevant issue here is that the database design and structure were not predicated upon a shared understanding among a specific community of practice of its content and possible uses. Rather, it was supposed to "serve as many specialized uses of data as possible", with data reuse enabling collaboration or even integration across communities.

In contrast, the knowledgebases discussed in this paper seek to provide evidence-based, actionable interpretations of genomic data for use by clinical practitioners engaged in the implementation of precision oncology. From this perspective, unlike the metaphorical travelers who maintain their identity in different locations, the constitution and handling of a knowledgebase cannot be reduced to the transfer of free-floating bits of information from publications to knowledgebases through nested database systems. The issue is not simply that each database channels and filters data. Rather, data experience a process of 'extensive' manual curation, whereby, after being extracted from publications, they undergo valuation and ordering by being paired with levels of evidence, levels of actionability, and summary statements that vary from knowledgebase to knowledgebase. As a result, the information provided by knowledgebases qualifies as actionable claims or statements, rather than data, and becomes undistinguishable from the knowledgebases in which it is embedded. This fact also accounts for the difficulties encountered when curators attempt to compare or harmonize knowledgebases.

Prominent oncology knowledgebases include Vanderbilt's My Cancer Genome (MCG), launched in 2011 as the first public somatic variant interpretation resource, MD Anderson's Personalized Cancer Therapy (PCT), Memorial Sloan Kettering's (MSK) OncoKB, and Wash U's Clinical Interpretations of Variants in Cancer (CIViC). These knowledgebases rely on the biomedical literature collected in the PubMed database and in other databases such as the Catalogue Of Somatic Mutations In Cancer (COSMIC). Established in the UK at the Wellcome Trust

Sanger Institute in 2004 with just four genes, COSMIC has now become the "world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer".[4] In addition to data manually curated from PubMed, COSMIC contains other datasets such as those produced by multi-center collaborative networks (Forbes et al. 2015). In short, and as the knowledgebase developers admit, theirs and similar resources stand "on the shoulders of these other giants, these other resources that have many more variants, tens of thousands, hundreds of thousands, even millions of observations and variants" (Interview with Drs Obi and Malachi Griffith, December 2016; henceforth MOG1). Figure 2 (Ainscough et al. 2016) illustrates this dependency structure.

Given the existence of multiple knowledgebases, oncologists are confronted with a complex landscape of interrelated resources that, despite recurrent harmonization initiatives, display distinct characteristics and design choices that promote their individuality. An informant spoke, in this respect, of a "very complicated landscape of resources that are pulling multiple different resources together, integrating them in some way, helping things be visualized, or making things more user friendly, and it's a bit Wild West" (MOG1). Rather than standalone devices, these resources maintain both horizontal and vertical relations: some repositories, such as COSMIC, act as de facto quasi-standards on which others explicitly rely, extracting and embedding their content, while simultaneously maintaining an individuality that challenges the seamless interoperability of their data. CIViC, for instance, links its content to COSMIC, perceived as a complementary and yet distinct resource:

> If you have a specific variant and you find it in CIViC, then you know that someone in CIViC believes it is clinically relevant, with some documented evidence, and we link out to

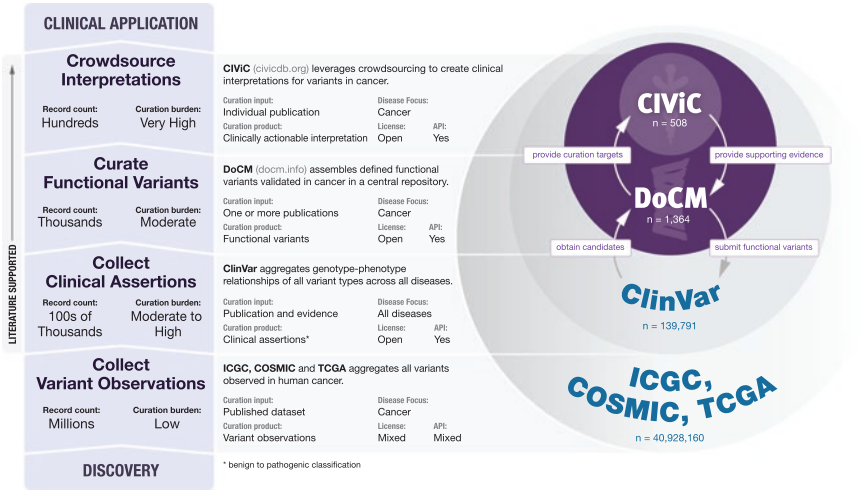


Fig. 2 CIViC in the context of related resources. Reprinted by permission from Springer Nature: *Nature Methods*, DOCM: A database of curated mutations in cancer, B.J. Ainscough et al., Copyright ©2016

[4] https://cancer.sanger.ac.uk/cosmic

> that variant’s record in COSMIC, so you can learn also what COSMIC says about that variant, about how many types of cancers have seen that variant before, and that is useful information. But there are many variants, most variants, that you won’t find in CIViC because they haven’t yet reached this level of documented clinical relevance, but they still exist in COSMIC and that’s still useful information that you could use to design an experiment or understand something about that variant, but it just doesn’t reach that level of clinical relevance for CIViC. (MOG1).

Far from being isolated and self-contained, knowledgebases function within an information ecosystem to whose intricacy and development they contribute. Their curatorial practices, for instance, include the active consulting of other databases and knowledgebases:

> When curators receive a list of gene-variants to curate, they are also given instructions to not limit their search for information to PubMed. They are trained to reference other publicly available knowledgebases such as COSMIC, Jackson Lab’s Jax CKB, and MyCancerGenome for example. Importantly, they are explicitly instructed to not copy or paraphrase the interpretations from these knowledgebases, but to use them as a resource for the primary literature on key gene variants. (Interview with Drs Debyani Chakravarty and J.J. Gao, May 2017; henceforth C/G).

They also openly relate to (or even embed) each other. As part of its data architecture, MSK has developed cBioPortal (now a multi-center endeavor), an advanced data visualization tool that draws on a number of different resources including, most obviously, OncoKB, but also CIViC, MCG, and, as one would expect given its preeminence in the field, COSMIC. When viewing the record for a given variant in cBioPortal, a user can mouse over icons for each of the above resources to bring up a brief summary of the information they contain or click through to proceed to their website. The information excerpted from those resources can thus be accessed directly via the cBioPortal interface, but the kind of information provided in the pop-up windows is different for each resource, so that inclusion of different knowledgebases provides complementary, rather than redundant information, about the ‘same’ molecular entity.

5 Practitioners’ Accounts of the Database/Knowledgebase Distinction

When asked to elaborate on the distinction between databases and knowledgebases, one of the developers of My Cancer Genome offered the following tripartite categorization, borrowed from the ‘data-information-knowledge hierarchy’[5]:

> You take data, say measurements or patient data, then you analyze or aggregate or present those data, and that would be the information, and then if you synthesize information across a bunch of different sources, that would be the knowledge. The point of MCG and CIViC and some of the other resources is really to be a ‘knowledgebase’. (Interview with Dr. Christine Micheel, July 2017; henceforth CM1).

[5] See https://en.wikipedia.org/wiki/DIKW_pyramid

Asked to clarify her statement by comparing, for instance, COSMIC and MCG, she answered that "COSMIC catalogues the alterations that have been observed in cancer, and MCG explains how that may impact therapeutic decisions" (Interview with Dr. Christine Micheel, August 2017). For his part, when asked a similar question the COSMIC director replied:

> We focused on the database angle until fairly recently. We wanted to collect as much information in the one place to empower others to investigate it to look for new genes, new targets. And we kept feeding the database. The increasing breadth and depth of that database just gives other scientists more power for their investigations. … Are we a database or a knowledgebase? We're probably focused more on the database angle of this than the knowledgebase angle. (interview with Dr. Simon Forbes, May 2017; henceforth SF2).

Rather than attempts to build robust ontological categories, these definitions of the database/knowledgebase distinction qualify as pragmatic categorizations within a rapidly evolving context. They situate each kind of 'base' in relation to the aforementioned spectrum that ranges from large-scale repositories, such as the now defunct caBIG, to single-purpose knowledgebases, via intermediate entities such as COSMIC that qualify as 'information bases' insofar as they systematically arrange information. The case of COSMIC, given its liminal position, is a useful starting point for clarifying this issue.

Compared to other endeavors COSMIC qualifies as a 'giant' because of the millions of data it contains in contrasted to the thousands typically found in a knowledgebase. Because of its 'database-ish' nature, and its comprehensive reach, COSMIC "is different things for different people … in some sense, it is just a large bucket of information that you can sift through with different perspectives in mind" (interview with Dr. Simon Forbes, February 2017; henceforth SF1). COSMIC, however, is not an undifferentiated 'bucket', but a bucket of baskets: it includes data subsets targeted to specific users. For instance, the Cancer Gene Census subset that catalogues genes causally implicated in cancer has been recently upgraded by adding annotations related to the traits that govern carcinogenesis, known as the 'hallmarks' of cancer (Hanahan and Weinberg 2000, 2011). As noted by the director of COSMIC, the CGC "the way it looks at the moment is more 'database-ish' as well, but with the new hallmarks annotations we're aiming more toward knowledge, we can describe the functional impact of each gene in cancer rather than just that it causes cancer" (SF2). This is part of a broader plan to transition from an exclusive focus on data acquisition, to the inclusion of annotations about the value of the information, leading, for instance, to the design of a "targeted, specific subset of the database toward clinicians and diagnostics". As acknowledged, however, by the same informant:

> If you're a clinician you might want to get in [COSMIC] for some clues around the impact of mutations, but it' not going to tell you that information because it wasn't really built with that in mind. We built it to gather large quantities of information. (SF1).

Is "looking for some clues around the impact of mutations" then the primary motivation for creating knowledgebases?

6 Why Knowledgebases?

Given COSMIC's pre-eminent position in the field, why did oncology practitioners feel the need to develop knowledgebases? Part of the answer lies in the need for dedicated clinical information to guide therapy. As noted by a cancer genomics researcher:

> COSMIC is just cataloguing pure genetics data online, so in the end we don't know much about clinical outcomes of these cases. It's very limited in scope. It still tells us whether a mutation has been observed more frequently than expected, which tells us something about whether it is likely to be a driver or not, but it still needs much more. We need much more data in these databases. (Interview with Dr. Marco Gerlinger, January 2016).

The missing data are bio-clinical, i.e. data that re-specify genomic entities by tying them to clinical insights; "what we're really interested in, is the clinical data that will be useful for interpretation of the molecular data, and to integrate that" (C/G). According to the same respondents, "in the development of OncoKB one thing became very clear: without clinician insight, OncoKB will be useless for clinical decision support". The information embedded in OncoKB links biological, clinical, and therapeutic information from multiple sources, which include not only the medical literature, but also FDA labeling, clinical guidelines, and abstracts from major conference proceedings, such as the American Society of Clinical Oncology (ASCO), the European Society of Medical Oncology (ESMO), and the American Association for Cancer Research (AACR).

Most importantly, in OncoKB annotations derived from these sources are not merely selected and organized by curators but vetted by a Clinical Genomics Annotation Committee (CGAC) consisting of MSK clinicians who represent leaders in their respective disease-specific fields:

> MSK has some of the best clinical and research expertise in the country. For OncoKB it was not sufficient to simply curate the available literature, our loftier goal was to capture, in a database readable format, the interpretation of these data through the lens of MSK in-house clinical expertise. (C/G).

The following example illustrates the nature of the clinicians' vetting:

> Our initial OncoKB curation efforts cast a wide net, allowing inclusion of information with any possible opportunity for clinical intervention based on the presence of a genetic variant. However, it became very clear very quickly that MSK is conservative in its definition of precision oncology. Thus, for example, we had initially included TP53 as potentially clinically actionable, based on an open phase I clinical trial testing a specific chemotherapy in TP53 mutant patients. However, the clinical committee made us immediately remove TP53 based on their real-world experience, i.e. TP53 alterations are present in 40% of patient tumors, [but] to-date there have been no therapies that have been able to effectively utilize TP53 as a predictive biomarker of activity for a targeted therapeutic. (C/G).

CIViC also focuses on data interpretation: "the meat of what we're trying to create, the data or content that we're creating, is actually the interpretation" (MOG1). The presence or absence of a clinical input is used to draw a line not only between CIviC and COSMIC but also between knowledgebases:

> Knowledgebases such as CIViC are great tools for use in the research space. They comprehensively capture the scientific literature and present this data in a research intuitive way. The development of knowledgebases such as MD Anderson's PCT and our OncoKB have been, from their inception, guided with the clinician in mind as the end-user. For OncoKB, there was an institutional mandate that physician scientists who represent disease experts were to guide the curation by specifying which information would be useful for clinical treatment decisions, and which information was considered extraneous. (C/G).

CIViC developers counter that:

> Resources like OncoKB and PCT talk a lot about their clinician review, but I haven't seen much of a structured representation of what that is, like which clinician reviewed which elements in what ways. You're just told: "You look at something in OncoKB or PCT, you should feel more confident in it because we have had it reviewed by clinicians." But that fact doesn't seem to be represented in the data model in any sophisticated way. (Interview with Drs Obi and Malachi Griffith, June 2017; henceforth MOG2).

The kind of curation, rather than the mere presence of curation, broadly defined, is thus at the very heart of the valuation processes that underlie the database/knowledgebase distinction.

7 Modes of Curation

During the Obama administration, when confronted with the challenges raised by precision medicine, the US FDA began considering a scenario according to which test developers might use information derived from a 'regulatory quality database' to support their claims. To qualify as 'regulatory quality', a database would be curated, have standards, and preferably provide levels of evidence, all of which differentiates it from a data repository (interview with an FDA official, March 2015). So, here is a first distinction: a non-curated repository and a curated database. But things are not so simple, because when asked for an example of a repository, our respondent mentioned a database that maintained in fact a relatively large team of curators. It thus looks as if it is not curation *per se* that is at stake here, but the *kind* of curation, namely research-oriented vs. clinically oriented curation. For instance, having attended a meeting of the International Society of Biocuration, one of the developers of MCG explained:

> Those are the folks that really started and maintained those research-oriented resources … I think the primary difference is the intended audience. When [Drs. Pao and Levy] conceived of MCG they were really focused on the clinician audience … both were practicing oncologists, intimately familiar with the workflows of a clinician, the way a clinician thinks, and the amount of time they have to look at a resource. The research-focused resources are really not what a clinician needs. (CM1).

The issue is not simply to avoid wasting a clinician's precious time, but, more importantly, to protect clinicians from being fed inaccurate or potentially damaging information derived from inappropriate contexts:

> For example, a patient with early stage disease is annotated to have this alteration and therefore they should get this therapy, without recognition that really in that context it is not within clinical guidelines to make that actionable. … There are a lot of manuscripts written

> without a clinical implication in mind, and saying this biomarker is associated with this drug sensitivity while the drug doses being used are clinically not relevant at all, or that biomarker association was really not a very strong one. (FMB).

This also accounts for the decision by more clinically oriented knowledgebases to include information from oncology conferences. While results presented at conferences are “generally not held to the same standard of quality or validation that a publication will be” (MOG2), they do contain relevant clinical information that is not otherwise available:

> When annotating the clinical implications of gene variants, our clinicians frequently referred us to interim clinical trial data from the proceedings of disease-specific and general clinical oncology conferences. Importantly, tumor-type specific negative data and information as to whether a drug is being discontinued from further development due to poor efficacy data is only available through conference proceedings. (C/G).

Several knowledgebases are deeply embedded in the clinical infrastructure of their parent organizations, thus providing further evidence of their situatedness. MD Anderson’s PCT, for example, acts as the external window of its Precision Oncology Decision Support (PODS) service (Meric-Bernstam et al. 2015; Kurnit et al. 2017, Dumbrava and Meric-Bernstam 2018). PODS is a prime internal resource for MD Anderson’s physicians who need assistance with the interpretation of genomic reports. It provides a rapid assessment of the quality of the testing platform, of the alterations seen in actionable genes, and of variant interpretation. In order to make it available for in-house physicians with similar patients, the information goes into a back-end database behind the institution’s firewall, whereas the information included in the external PCT knowledgebase concerns only those variants that have literature support.

A similar situation prevails at MSK, where thousands of patients are sequenced and subsequently matched with a large trial portfolio via a sophisticated IT infrastructure (Eubank et al. 2016). OncoKB annotation is included in the sequencing report that provides summaries of relevant information about alterations for which there are FDA-approved biomarkers and drugs, or compelling clinical data justifying enrolment in a specific clinical trial (C/G). The treating oncologist (who makes the final therapeutic decision) can then interact with the OncoKB team and other colleagues to further discuss the recommendations. As with MD Anderson, the public version of OncoKB does not include all internal information.

8 Trust and Transparency

Knowledgebases deploy different curatorial strategies that define how each positions itself vis-à-vis the others in a climate defined by both competition and collaboration within the oncology community. Rather than clinical expert knowledge, CIViC resorts to crowdsourcing, arguing that the sheer amount of potentially relevant references available in PubMed makes such an approach inescapable, a claim supported by the fact that the overlap between the publications curated by different knowledgebases is extremely low. CIViC’s Wikipedia-like crowdsourcing nonethe-

less involves, in addition to external curators (any user can in principle be a curator), internal curators, site editors, and domain experts in charge of ensuring quality. Crowdsourcing offers the advantage of introducing a measure of transparency:

> CIViC is the only database that actually allows a user to log in and comment and say: "Hey I disagree with this", or "You're missing this important paper", or "I would like to modify this to make it clearer". The other resources generally have behind the scenes a team of experts and they work as a sort of editorial board, almost like writing mini reviews about each variant and each gene, and they may have a collaborative process, but it's hidden and it's not occurring inside the interface and there's not the same degree of provenance about who exactly said what, and how did the knowledge evolve from its initial state to the current state, and so on. (MOG1).

To which other practitioners counter:

> Crowdsourcing as a theoretical concept is amazing. However, it comes with the assumption that clinicians, who have very limited time and bandwidth, will buy into that concept. I think one of the key factors contributing to the success of OncoKB is that MSK clinicians were mandated to guide OncoKB development since it was slated to be an institutional clinical decision support system. Additionally, we had carefully trained medical fellows and translational cancer biologists as curators who were well versed in the quality control of information that we would allow into OncoKB. (C/G).

The emergence of knowledgebases devoted to the same purpose is less an expression of redundancy than of the existence of different curatorial approaches that embed and enact each knowledgebase's strategy held together by a tension between trust (in expert judgment) and transparency (of the curatorial process). When asked what motivates the proliferation of knowledgebases, a practitioner explained:

> If you are a center or a company and you are interpreting a variant for an actual patient, a real patient, and you're acting on that information, what information do you trust? [What information] gives you confidence that you could actually act on that mutation to do something for that patient? … So, what's ended up happening is that every center just says: "We don't know who we're going to trust, so let's just recreate the whole thing over again and we control it." … There's kind of this tension between openness and trust. (Interview with Dr. Ethan Cerami, April 2017).

This tension is reflected in the different solutions adopted by CIViC and OncoKB. Both knowledgebases originated in an attempt to streamline interpretation work. Their development, however, diverged as CIViC adopted traceability and transparency as its trademark, whereas OncoKB is vetted by, and therefore representative of, MSK clinical expertise.

9 Curation, Interpretation, and Levels of Evidence

Thanks to its transparent curatorial system, CIViC offers a more granular view of those practices. The debates between curators and editors are available on the CIViC interface, and although a vast majority of them are relatively short and 'technical', some involve choices that escalate to concerns about underlying principles and the meaning of curation and data interpretation. Here is an example:

> Curator A posts evidence concerning the EML4-ALK E20 variant on the webpage.
>
> Editor B deletes part of the evidence summary arguing that it amounts to speculations. She also reduces the evidence trust rating from 5 to 3 stars.
>
> Curator A replies that he recognizes the speculative nature of his summary, but that this is part of his philosophy of evidence-statement production and his interpretation of CIViC’s mission, namely, to add context and speculate on possible connections and significance.
>
> In the ensuing discussion, Editor B asks Editors C, D, and E to weigh in on the discussion of the group’s philosophy of interpretation and evidence-statement production.
>
> She also attempts to clarify the meaning of a 5-star trust rating that should refer to highest-quality, standard-of-care studies, and be based on how well the evidence supports a given predictive statement, not the overall quality of the original paper.
>
> Concerning the deleted passages, Editor B suggests that “the additional text would be well suited to a comment at the time of submission, but I believe it to be tangential to the main point.”
>
> Editor D steps in, noting that information extracted from case reports warrants by definition a lower star rating, because of its anecdotal nature. He agrees with Editor B, and this ends the discussion.

This vignette shows how curation debates can be framed by the essential tension between the clinical purpose and utility of the knowledgebase (see the reference to standards of care), and the scientific validity and the future of evidence statements. Reminiscent of the work of guideline developers (Knaapen et al. 2010), it also highlights the textual dimension of curatorial practices, whereby data are polished into statements. A further example of this dynamic is provided by the following example:

> Following the posting of a new evidence-summary statement, the discussion focuses on whether certain kinds of lower-evidence statements, in this case about mutation co-occurrence, belong in CIViC because they could subsequently turn out to be useful.
>
> – Editor A questions the clinical utility of the evidence, whether the information actually fits into the evidence schema offered by CIViC, whether it qualifies as diagnostic, and whether it has been given the appropriate evidence-quality grade. He nonetheless acknowledges the importance and potential usefulness of the study behind the evidence statement.
> – Editor A ultimately rejects the submission, but with an encouragement to produce a new evidence statement that more clearly articulates its relevance.

It thus appears that ‘data’ excerpted from publications or databases are transformed through interpretation because they are turned into different kinds of evidence, or evidence for different things. Again, the issue is not about data or evidence *per se*, but about the *textual framing of evidence statements* and their relation to clinical utility. A key device, in this respect, is the attribution of Levels of Evidence (LoE) to statements, which act as markers of the degree of uncertainty characterizing the actionability of that statement. All the knowledgebases we investigated include LoE, and this, once again, reminds us of the centrality of this device in relation to clinical utility:

> I think the levels of evidence is instrumental, because for a clinical decision support tool to have any sort of utility a clinician needs to know: “What am I doing? Is it backed by consensus?” (G/C).

Knowledgebases have adopted different approaches to LoE. For instance, OncoKB's LoE are tied to the sum of evidentiary support that a specific mutational event is predictive of response to a targeted therapy, whereas CIViC's LoE reflect the source of the evidentiary support that comes with the statement. CIViC items are additionally accompanied by a 'Trust Rating' that indicates how compelling that evidence is judged to be. There are, moreover, differences in how knowledgebases advertise their LoE component. For instance, CIViC is described as a "community knowledgebase for expert crowdsourcing" (Griffith et al. 2017), whereas OncoKB is presented as a "a precision oncology knowledge base" that includes a distinctive system of Levels of Resistance (LoR) predictive of resistance to a specific targeted therapy (Chakravarty et al. 2017).

These differences can be compounded with the fact that establishing LoE is notoriously contentious as it involves a large degree of interpretative flexibility and because of the conflicting sources that can be used to perform that task:

> The interpretation of the genomic variants is subjective – I mean a fifty percent response rate for you is responsive? What about five percent? … For that individual patient, one of twenty that responded, this gene-drug-disease match was perfect. Just one out of twenty. Five percent. So, is this responsive if I consider a broader population? … We have one interpretation that is different from OncoKB: they have their own strengths because they have internal data, but [our source] is published, we have the connection. (Interview with an oncology data scientist, October 2016).

This brings us back to the tension between trust in the clinical expertise available at leading cancer centers and the traceability of statements to published sources. The process at MSK illustrates how the clinical consensus of an institute is captured by knowledgebase annotation:

> Several MSK physician-scientists, who represent a broad spectrum of opinion have provided insight into what a given OncoKB annotation should or should not include. One key role of OncoKB is to generate a consensus of opinion from these varied voices. Discussions and compromise have taken place through this process, no one voice has dominated, and the OncoKB annotation represents the middle ground. (C/G).

The excerpt highlights the role of local context and shared understandings in the valuation processes underlying the trustworthiness of specific statements, and thus the worth of individual knowledgebases.

10 Heterogeneity

Knowledgebases differ in terms of the kind and amount of information they carry and the assessment and interpretation of the evidence they include. In fact, they overlap very little in terms of the specific variants included and the literature they reference. When they do overlap, they may actually interpret variants differently, either because their curation relies on different publications, or because they interpret those publications differently (Patel et al. 2016).

Knowledgebases contain interpretations rather than 'data' as such (Pitel 2017). These interpretations consist of statements about associations, i.e. claims about the

evidence that a given mutation plays a particular role in cancer, and the evidence that a drug or intervention may be associated with that variant and have clinical relevance. Even in a database such as COSMIC the 'data' is not the variant itself, but the pairing of a set of genomic coordinates that represent the variant with a given biopathological process. In the case of knowledgebases, the unit of analysis consists less of 'data' than evidence records, which amount to sets of locations, cross references, and literature citations leading to an interpretation. The interpretation defines which variants are clinically relevant and the description of that clinical relevance varies from one knowledgebase to another. Factors that account for this variation include the sheer number of available publications, so that the overlap of the literature covered by a given knowledgebase can be quite small. Moreover, as noted by the developer of the PathOS decision support system (Doig et al. 2017), "a PubMed article is a pretty large body of data, and actually finding the sentence that confirms that the action is positive or negative or related to something is actually a very hard job" (Interview with Dr. Ken Doig, June 2017).

Other sources of heterogeneity include temporality and granularity. Temporality refers to the rapidly evolving knowledge in oncology, so that information presented at a conference, or even published, can be quickly disproved or replaced:

> We get a lot of requests to add [information from conference abstracts] because there are clinicians who want the most amazing cutting-edge stuff, and then you have other clinicians where we have the feedback that this published NEJM paper from three years ago [is] not good enough because it was debunked by a subsequent JAMA paper two years later, with a much larger clinical trial that was better statistically powered. (MOG2).

As for granularity, while the knowledge at the level of a gene expressed in guidelines and regulatory documents might be relatively stable, the same does not apply to gene variants:

> The FDA-labeling of approved targeted agents in a specific indication can be vague. For example, the FDA-approval of erlotinib in patients with EGFR-mutant non-small cell lung cancer was irrespective of EGFR mutation status. This is because in these cases, the drug's approval predates much of the sequencing data that determined the specific patient populations that benefit from the targeted agent. (C/G).

Similar considerations apply to guidelines that include mutations for which there are established data:

> But what does a clinician do when faced with a sequencing result that includes a known actionable gene but a lesser known variant? … That kind of information is critical in supporting clinical care, and that's where the levels of evidence represent a practical and immediate way to communicate this information. (C/G).

Knowledgebase developers are well aware of the issue of heterogeneity which is viewed as both problematic and unsurprising given the extent of the field and the complexity of interpretation. They have recently established the Variant Interpretation for Cancer Consortium (VICC), to "harmonize global efforts for clinical interpretation of cancer variants".[6] Rather than building yet another knowledgebase (a 'meta-

[6] https://genomicsandhealth.org/working-groups/our-work/variant-interpretation-cancer-consortium

knowledgebase'), the idea is to construct a portal giving access to the content of multiple knowledgebases. Thus, the field may move toward addressing the problem of heterogeneity without having to sacrifice either the latent mistrust embedded in or the pragmatic role fulfilled by locally maintained knowledgebases. This suggests that rather than a solution to the 'data interpretation bottleneck', knowledgebases and their claims and statements are still part of that same bottleneck, requiring additional bioinformatic and expert clinical human work.

11 Conclusion

Oncology has recently transitioned from its traditional, organ-based approach to a 'precision oncology' of molecular alterations. As a result, it has become 'data-centric' (Leonelli 2016). Its practices increasingly rely on a sophisticated techno-scientific infrastructure that generates large amounts of data that demand consistent, appropriate interpretations. In turn, attempts to overcome the interpretation bottleneck have led to the establishment of a complex landscape of interrelated resources that, while displaying distinct characteristics and design choices, also entertain horizontal and vertical relations. Although there is no denying that the data-centric nature of contemporary oncology raises a number of key issues related to the production and circulation of data — issues that can be explored using the 'data journeys' metaphor — we suggest in this paper that the focus on data use and re-use should be complemented by a focus on interpretation. Interpretation here refers to both the 'interpreting' activities performed by bio-clinical collectives, and to the outcomes of those activities under the guise of *actionability claims or statements*, rather than 'data'.

Oncology practitioners refer to data interpretation resources as 'knowledgebases', an actor's category designed to differentiate them from generic, multipurpose databases. While in most cases publicly accessible, albeit in a pared-down format compared to their in-house version, knowledgebases are deeply embedded in the clinical pathways of their home institutions. Their major purpose is the definition and identification of *clinically actionable* alterations, i.e. those that drive tumors and can be matched to treatments. This is no easy task, as shown by the existence of several knowledgebases that, in spite of their common purpose, adopt very different approaches partly captured by the tension between trust and traceability. To investigate what makes different knowledgebases 'valuable' to genomic practitioners confronted with a rapidly evolving domain, we have examined their structure and dynamics. The nature, amount, and quality of curation underwriting each knowledgebase appear to be major contributors to these valuation processes. A heavy investment in human curation, of a clinical rather than exclusively scientific nature is needed to make them valuable, but each knowledgebase appears to have its own way of connecting clinical and scientific statements elicited from publications, conference abstracts, clinical trials, genomic datasets, and even in-house expert statements.

The main goal of the NIH "Strategic Plan for Datascience" mentioned at the beginning of this paper is to facilitate "the modernizing [of] the NIH-funded biomedical data-resource ecosystem". The Plan refers to the development of core data

repositories to be used across different scientific domains, but also marks out a special place and a distinct role for knowledgebases within the data ecosystem. Knowledgebases that, as just mentioned, involve large amounts of human curation have been developed by “targeted communities for the benefit of scientists in that community”, and they are here to stay, as they will “still serve the functions of their own communities the way they always have, [as] distinct entities with their own priorities, their own goals and objectives” (Interview with Dr. Susan Gregurick, May 2018). While, according to the same respondent, part of the information they contain could be ‘hardened’, by for instance being made compliant with the FAIR principles for data management (Wilkinson et al. 2016), and thus transferred at some future point to a data repository, the situated and ever-changing nature of the information collected in knowledgebases make such a prospect somewhat difficult to entertain, especially in clinical domains characterized by the ongoing realignment of the normal and the pathological.

Admittedly, the database/knowledgebase distinction is ideal-typical, given that COSMIC, for instance, is shifting from its initial exclusive focus on data acquisition to highlighting the value of its data (SF2). Oncologists consult COSMIC for research purposes but also to gather information about alterations detected in their patients, although they might do so via local resources that embed COSMIC. While there is an overlap, in terms of use, between COSMIC and the more specialized knowledgebases, the latter lie at one end of a wide spectrum of resources that range from large databases to smaller interpretative resources. In the case of a database such as COSMIC that sits in the middle of this spectrum, the data journey metaphor may be used to describe how curators survey the literature, extract and refashion bits of information, assess their evidentiary strength, and decide whether and how to include them in the database. The addition of the PubMed reference number to those data in principle should allow users to travel back to the original source although, as already mentioned, this is not a straightforward task given the amount of curatorial work needed to locate specific statements. Knowledgebases, however, are less a data repository than a tool for (clinical) action, and the data journey metaphor misses this key aspect. Within knowledgebases bits of information are triangulated with other evidence, associated with levels of evidence and actionability, and embedded in carefully crafted statements that re-specify their meaning. This explains, in part, the major differences between knowledgebases, whereby the ‘same’ genomic variant is transmogrified into different entities connected to different actions.

In a domain where genomic information is becoming increasingly important for clinical decision-making, but drastically outpacing the genomic literacy of the average oncologist/clinician, knowledgebases are an attempt to fill a translational gap and provide clinicians with information about the actionability of molecular alterations, and the kind and strength of the evidence that underpins it. Knowledgebases, in this context, are designed to act, in a sense, as a virtual, *in-silico* ersatz for the multi-disciplinary gathering of oncology practitioners, molecular biologists, and bioinformaticians who come together to reach a consensus about actionable suggestions (Bourret and Cambrosio 2019). In the case of institutions such as MSK, the sheer number of sequenced patients (Zehir et al. 2017; Eubank et al. 2016) makes such a solution impossible. Instead, a tumor profiling report associated with a clinical decision support tool, OncoKB, is sent electronically to

the treating physician who can trust the provided clinical annotations because they are clinically vetted. "OncoKB", in this context, refers not merely to the knowledgebase, narrowly defined, but to the entire *dispositif*, that includes, for instance, the Clinical Genomics Annotation Committee staffed with leading clinicians.

Knowledgebases, rather than a mere data repository, embed and perform interpretations that deploy a distinctive form of bio-clinical *expertise*. Conversely, in data-centric oncology human expertise can only be enacted *via* bio-clinical collectives properly equipped with tools and devices such as those provided by knowledgebases. This apparently vicious circle becomes virtuous when those tools and devices are constituted and utilized at different places and different times by different collectives. Hence the temporal and relational nature of oncology databases and knowledgebases, which evolve in response to a number of other initiatives, for instance the introduction of new data-sharing projects sponsored by leading cancer centers. Last but not least, we should not forget the strictures that oncology, as a *clinical* domain, imposes upon knowledge production and knowledge flows, and which largely account for the difference between clinical-grade knowledgebases and the kind of databases deployed in other scientific domains.

Acknowledgements Research for this paper was made possible by the following grants: Canadian Institutes of Health Research MOP-133687 and French National Cancer Institute (INCa) SHSESP14-002. We would like to thank all the knowledgebase developers who agreed to be interviewed, in some cases repeatedly. Special thanks to Sabina Leonelli who singlehandedly coerced us into writing this paper.

References

Ainscough, Benjamin J., Malachi Griffith, Adam C. Coffman, et al. 2016. DoCM: A Database of Curated Mutations in Cancer. *Nature Methods* 13: 806–807.

Amin, Ash, and Patrick Cohendet. 2004. *Architectures of Knowledge: Firms, Capabilities, and Communities*. Oxford: Oxford University Press.

Bates, Jo, Yu-Wei Lin, and Paula Goodale. 2016. Data Journeys: Capturing the Socio-Material Constitution of Data Objects and Flows. *Big Data & Society* 2016 (July–December): 1–12. https://doi.org/10.1177/2053951716654502.

Bijker, Wiebe E., and John Law, eds. 1992. *Shaping Technology/Building Society: Studies in Sociotechnical Change*. Cambridge, MA: MIT Press.

Bourret, Pascale, and Alberto Cambrosio. 2019. Genomic Expertise in Action: Molecular Tumour Boards and Decision-Making in Precision Oncology. *Sociology of Health & Illness* 41: 1568–1584.

Buetow, Kenneth H. 2005. Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research. *Science* 308: 821–824.

Cambrosio, Alberto, and Peter Keating. 2000. Of lymphocytes and Pixels: The Techno-Visual Production of Cell Populations. *Studies in History and Philosophy of Biological and Biomedical Sciences* 31: 233–270.

Chakravarty, Debyani, Jianjiong Gao, Sarah Phillips, et al. 2017. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology* 1: 1–16. https://doi.org/10.1200/PO.17.00011.

Doig, Kenneth D., Anthony Fellowes, Andrew H. Bell, et al. 2017. PathOS: A Decision Support System for Reporting High Throughput Sequencing of Cancers in Clinical Diagnostic Laboratories. *Genome Medicine* 9 (1): 38.

Dumbrava, Ecaterina I., and Funda Meric-Bernstam. 2018. Personalized Cancer Therapy. Leveraging a Knowledge Base for Clinical Decision-Making. *Cold Spring Harbor Molecular Case Studies* 4: a001578.

Eubank, Michael H., David M. Hyman, Amritha D. Kanakamedala, et al. 2016. Automated Eligibility Screening and Monitoring for Genotype-Driven Precision Oncology Trials. *Journal of the American Medical Informatics Association* 23: 777–781.

Forbes, Simon A., David Beare, Prasad Gunasekaran, et al. 2015. COSMIC: Exploring the World’s Knowledge of Somatic Mutations in Human Cancer. *Nucleic Acids Research* 43: D805–D811.

Ghazani, Arezou A., Nelly M. Oliver, Joseph P. St Pierre, et al. 2017. Assigning Clinical Meaning to Somatic and Germ-Line Whole-Exome Sequencing Data in a Prospective Cancer Precision Medicine Study. *Genetics in Medicine* 19: 787–795.

Gitelman, Lisa, ed. 2013. *“Raw Data” is an Oxymoron*. Cambridge, MA: The MIT Press.

Goldberg, Paul. 2011. Prepare for “Tsunami” of Genomic Information, Sledge Urges in ASCO Presidential Address. *The Cancer Letter* 37 (23): 1–7.

———. 2012. NCI Bioinformatics After Kenneth Buetow: Varmus Launches Fundamental Redesign. *The Cancer Letter* 38 (1): 1–6.

Good, Benjamin M., Benjamin J. Ainscough, Josh F. McMichael, et al. 2014. Organizing Knowledge to Enable Personalization of Medicine in Cancer. *Genome Biology* 15: 438.

Griffith, Malachi, Nicholas C. Spies, Kilannin Krysiak, et al. 2017. CIViC is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer. *Nature Genetics* 9: 170–174.

Hanahan, Douglas, and Robert A. Weinberg. 2000. The Hallmarks of Cancer. *Cell* 100: 57–70.

———. 2011. Hallmarks of Cancer: The Next Generation. *Cell* 144: 646–674.

Hatchuel, Armand. 1996. Les axiomatiques de la production: éléments pour comprendre les mutations industrielles. In *La performance économique en entreprise*, ed. Jacques-Henri Jacot and Jean-Pierre Micaëlli, 35–53. Paris: Hermes.

Johnson, Amber, Jia Zeng, Ann M. Bailey, et al. 2015. The Right Drugs at the Right Time for the Right Patient: The MD Anderson Precision Oncology Decision Support Platform. *Drug Discovery Today* 20: 1433–1438.

Jordan, Bertrand. 2015. Recherche cibles, désespérément. *Médecine/Science* 31: 214–217.

Knaapen, Loes, Hervé Cazeneuve, Alberto Cambrosio, et al. 2010. Pragmatic Evidence and Textual Arrangements: A Case Study of French Clinical Cancer Guidelines. *Social Science & Medicine* 71: 685–692.

Kurnit, Katherine C., Ann M. Bailey, Jia Zeng, et al. 2017. ‘Personalized Cancer Therapy’: A Publicly Available Precision Oncology Resource. *Cancer Research* 77: e123–e126.

Leonelli, Sabina. 2013. Global Data for Local Science: Assessing the Scale of Data Infrastructures in Biological and Biomedical Research. *BioSocieties* 8: 449–465.

———. 2016. *Data-Centric Biology. A Philosophical Study*. Chicago: University of Chicago Press.

———. this volume. Learning from Data Journeys. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.

Lynch, Michael. 1993. *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge, UK: Cambridge University Press.

———. 1998. The Discursive Production of Uncertainty. The OJ Simpson ‘Dream Team’ and the Sociology of Knowledge Machine. *Social Studies of Science* 28: 829–868.

———. 2009. Working Out What Garfinkel Could Possibly Be Doing with “Durkheim’s Aphorism”. In *Sociological Objects*, ed. Geoff Cooper, Andrew King, and Ruth Rettie, 101–118. London: Routledge.

Meric-Bernstam, Funda, Carol Farhangfar, John Mendelsohn, et al. 2013. Building a Personalized Medicine Infrastructure at a Major Cancer Center. *Journal of Clinical Oncology* 31: 1849–1857.

Meric-Bernstam, Funda, Amber Johnson, Vijaykumar Holla, et al. 2015. A Decision Support Framework for Genomically Informed Investigational Cancer Therapy. *Journal of the National Cancer Institute* 107 (7): djv098.

Nelson, Nicole, Peter Keating, and Alberto Cambrosio. 2013. On Being ‘Actionable’: Clinical Sequencing and the Emerging Contours of a Regime of Genomic Medicine in Oncology. *New Genetics & Society* 32: 405–428.

Patel, Jaymin M., Joshua Knopf, Eric Reiner, et al. 2016. Mutation Based Treatment Recommendations from Next Generation Sequencing Data: A Comparison of Web Tools. *Oncotarget* 7: 22064–22076.

Pitel, Beth. 2017. *Introduction to Publically Available Knowledgebases to Aid Interpretation of Genomic Findings in Oncology*. https://www.youtube.com/watch?v=4dBh1Qkp8os. Accessed 21 Aug 2019.

Prasad, Vinay. 2016. The Precision-Oncology Illusion. *Nature* 537: S63.

Schwaederle, Mzria, Gregory A. Daniels, David E. Piccioni, et al. 2015. On the Road to Precision Cancer Medicine: Analysis of Genomic Biomarker Actionability in 439 Patients. *Molecular Cancer Therapeutics* 14: 1488–1494.

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: 387–420.

Subbiah, Vivek, and Razelle Kurzrock. 2016. Universal Genomic Testing Needed to Win the War Against Cancer: Genomics IS the Diagnosis. *JAMA Oncology* 2: 719–720.

———. 2017. Debunking the Delusion that Precision Oncology Is an Illusion. *The Oncologist* 22: 881–882.

Suchman, Lucy A. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, UK: Cambridge University Press.

Thomas, Uduak G. 2012. NCI reorganizes cancer informatics efforts; cuts some caBIG programs, moves others to NCIP. *GenomeWeb.* https://www.genomeweb.com/informatics/nci-reorganizes-cancer-informatics-efforts-cuts-some-cabig-programs-moves-others. Accessed 21 Aug 2019.

Titus, Karen. 2014a. Molecular tumor boards: Fixture or fad? *CAP Today,* October 14. http://www.captodayonline.com/molecular-tumor-boards-fixture-fad. Accessed 21 Aug 2019.

———. 2014b. From tumor board, an integrated diagnostic report. *CAP Today*, December 15. http://www.captodayonline.com/tumor-board-integrated-diagnostic-report. Accessed 21 Aug 2019.

Van Allen, Eliezer M., Nikhil Wagle, and Mia A. Levy. 2013. Clinical Analysis and Interpretation of Cancer Genome Data. *Journal of Clinical Oncology* 31: 1825–1833.

Vignola-Gagné, Etienne, Peter Keating, and Alberto Cambrosio. 2017. Informing Materials: Drugs as Tools for Exploring Cancer Mechanisms and Pathways. *History and Philosophy of the Life Sciences* 39: 10.

Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018.

Zehir, Ahmed, Ryma Benayed, Ronak H. Shah, et al. 2017. Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients. *Nature Medicine* 23: 703–713.

Alberto Cambrosio A Professor at McGill University's Department of Social Studies of Medicine, Alberto Cambrosio's recent work examines "genomics in action", i.e. as applied to concrete instances of medical work, by investigating public, academic and commercial programs that capitalize on the therapeutic insights offered by the new molecular genetics of cancer. His most recent book (*Cancer on Trial: Oncology as a New Style of Practice*, University of Chicago Press, 2012, coauthored with Peter Keating) argues that, contrary to common assumptions, clinical trials do not boil down to a mere "technology" or a few methodological principles: rather, they are an institution that corresponds to a profound transformation of biomedical activities and rise to the level of a "new style of practice". This work builds on a previous book (*Biomedical Platforms: Realigning the Normal and the Pathological in Late-Twentieth-Century Medicine*, MIT Press, 2003, also coauthored with Peter Keating) that analysed the transformation of medicine into biomedicine.

Jonah Campbell has an MA degree in Medical Sociology and is presently a Regular Research Assistant in the Department of Social Studies of Medicine. He has over 7 years’ work experience on a variety of research projects in the sociology and history of biomedicine, with a particular focus on precision medicine, genomic oncology and “Big Data”.

Etienne Vignola-Gagné is an analyst at Science-Metrix, where he conducts projects on science and innovation policies and research management. At McGill University, and the University of Vienna before (where he obtained his doctoral degree), he combined policy analysis and science and technology studies to track the history of “translational research” programs and to follow the introduction of genomics sequencing technologies in clinical oncology. He has authored or coauthored scientific contributions published in venues such as *History and Philosophy of the Life Sciences*, *Science and Public Policy* and *Scientometrics*.

Peter Keating is an Associated Professor at the University of Quebec at Montreal. Currently semiretired, he worked for many years in the history of immunology and oncology and, most recently, clinical cancer trials. He has coauthored several books with Alberto Cambrosio on these topics including *Cancer on Trial: Oncology as a New Style of Practice* (University of Chicago Press, 2012).

Bertrand R. Jordan (b. 1939) obtained his PhD in Particle Physics (CERN, 1965), then moved to molecular biology and spanned many topics during his career as CNRS Research Director (mostly at the Marseille-Luminy Immunology Institute). He notably isolated the first HLA gene in 1982 before turning to genomics, expression profiling and medical and cancer genetics. He has edited two multi-author treatises on gene expression and microarrays in diagnostics and also published 12 books on genetics aimed at the general public. He is a Consultant for several biotech companies in the field of cancer diagnostics and therapy and publishes a monthly *Chronique Génomique* (Genomic Chronicle) in the French journal *Médecine/Sciences*. He is now retired from CNRS but remains Associate Researcher at the ADÈS laboratory.

Pascale Bourret is Associate Professor at Aix-Marseille Université where she teaches sociology. She is also a Researcher at SESSTIM (Economy and Social Sciences, Health Care Systems and Societies), an INSERM-IRD-Aix-Marseille Université UMR. At the crossroads of science studies and sociology of medicine, her work focuses on the transformation of biomedical practices in connection to the development of genomic tools, with a focus on biology/clinic interface, the transformation of clinical work and the production of clinical judgement and clinical decision-making. She has published articles on bio-clinical collectives in the domain of BRCA testing, on regulation issues linked to new genomics tools and on the emergence of genomic-driven clinical trials. Her present projects investigate the implementation of precision medicine in oncology and explore the conditions surrounding the development of targeted therapies in the context of clinical and translational research.

Realizing Healthful Housing: Devices for Data Travel in Public Health and Urban Redevelopment in the Twentieth Century United States

Edmund Ramsden

Abstract The field of housing is dependent upon data from a wide range of sources, as issues of architecture, engineering, finance, sanitation, public health and social relations must all be considered in policy, planning and design. This chapter documents the efforts of housing and public health experts in mobilizing housing data across different disciplinary and social spaces in the 1930s and 40s. To overcome the immense challenge of making such extensive and diverse information available and useful, we will explore how *actionability* was built into the very methods of collecting, processing, and circulating information. New standards and appraisal techniques were devised by the Committee on the Hygiene of Housing of the American Public Health Association that would shape and determine housing data journeys in critically important ways. It was by devising new ways to simultaneously collect, organize, package and translate data in a way that was meaningful for planners and policy-makers, that led to healthful housing surveys and public health ideals playing a critical role in a period of intensive urban redevelopment and renewal in the mid-twentieth century United States.

1 The Problem of Data in Housing

Huntington Williams, Baltimore's Commissioner of Health, described the improvement of housing as the health oficer's "real opportunity" (Williams 1942, 1001). Williams was becoming a leading figure in the rapidly expanding movement to realize public health goals through urban redevelopment. While the focus on housing offered unmatched potential for preventing physical and mental disease, for planners and architects, the subject of health legitimated their expanding role in the construction and design of urban environments. The growth of this health and housing nexus was, in turn, critically dependent on data. They needed information on

E. Ramsden (✉)
School of History, Queen Mary University of London, London, UK
e-mail: e.ramsden@qmul.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_17

housing quantity and quality to identify shortages, specify problem areas, predict future needs, establish housing standards and promote new designs. They needed data on the relationship between health and housing that could travel from the laboratory and field studies of the physical, medical and social sciences to the planning offices of federal, state and municipal government.

Data was of critical importance but also generated serious problems. The data required for housing reform and urban redevelopment was extraordinarily complex, involving such areas as child development, home economics, loans and finance, social relations, engineering and construction, architectural design, sanitation, mental health and disease transmission. Data from a wide range of sources, such as physiological and engineering laboratories and social and epidemiological surveys, needed to be organized, condensed, and translated for use in the field by public health workers, builders and architects. A critical issue was, therefore, that of travel, of finding ways of transforming a highly heterogeneous mass of data into an evidence-base that would prove useful for policy and planning. In the building industry, James Baldwin of Armstrong Cork Company declared: "We must know what is going on, what has been done, and particularly what has been proven good or bad. Thus we product researchers have as our greatest problem---information. The problem is twofold…. first, how to find data, second, lack of data."[1]

This chapter documents the efforts involved in mobilizing housing data across different disciplinary and social spaces. It takes the reconstruction of housing data journeys as a window into the intertwined roles played by research, public services, and policy demands in shaping American public health interventions and building developments in the 1930s and 40s. For data to travel it needed to be useful, transmitted a way that would allow it to be applied to housing problems by a range of professionals such as James Baldwin and his fellow builders, planners, engineers and architects. This necessity of ensuring policy usefulness was an immense challenge and allows us to explore just how *actionability* was built into the very methods of data collection, processing, packaging and circulation, and would thus shape and determine housing data journeys in important ways.[2]

The chapter first examines attempts by public housing activists in the 1930s to integrate data from a variety of surveys carried out for different purposes. After building composite data on the consequences of housing quantity and quality for health, this was then circulated through public health and housing reports, newsletters, manuals and memoranda. However, this approach was soon recognized as insufficient. The most serious problem was that of travel. The data presented in city-wide surveys could not be interpreted in a way that served the practical purposes of government agencies. It was too scattered, raw, imprecise, incomparable and

[1] Building Research Advisory Board, BRAB Notes, February 5, 1954, E&IR: Building Res Adv Bd, 1950–1954, National Research Council, Archives of the National Academy of Sciences, Washington DC. Emphasis in original.

[2] This parallels Alberto Cambrosio et al., emphasis on the specific infrastructures set up to make biomedical data *actionable* by clinical researchers: they remark that this generated all sorts of interesting differences in the ways data are treated.

general.[3] The paper will then turn to focus on the work of the Committee on the Hygiene of Housing (CHH) of the American Public Health Association (APHA), established in 1937 to address this problem.

Drawing on evidence provided through a variety of laboratory and field studies, the CHH generated a series of "principles" that would need to be met for housing to be considered "healthful" – to prevent mental and physical disease, rooms needed to be of a certain size and be sanitary, ventilated and heated to a certain temperature. Data was collected, codified, and published as a more specific set of "standards" for healthful housing. Standards systematized data and conferred the credibility, validity, and authority necessary to build consensus over the healthy home. But to travel successfully, data also needed to be presented in a way that was actionable by planning agencies in relation to local contexts.[4] To this end, the Committee created an "appraisal technique", a new survey technology that put their principles into practice. It was a device that generated, processed and applied data, making it comparable across time and space. Its detailed statistical maps of urban environments served to transfer evidence on the relationship between housing and health, generated in the laboratories and field studies of medical, physical and social scientists, into the plans and designs of municipal governments.

The paper will argue that it was the CHH's success in devising ways to simultaneously collect, process, package and translate data in a way that was meaningful for planners and policy-makers, that led to healthful housing surveys playing a critical role in a period of intensive urban redevelopment in the mid-twentieth century United States. In making data that travels into the realm of policy, ensuring that it is acted upon in particular ways and in accordance with pre-defined goals or principles, a key issue is that of control. In order to encourage and enable agencies to regulate housing in accordance with the principles of public health, the CHH also needed to determine what kind of data was being acted upon and ensure that it was consistent with their interests. Thus, rather than producing repositories of data that could be applied by local agencies in various ways and to their own ends, the series of tools constructed by the CHH, most notably their standards and appraisals, allowed them to continue to exert control over key stages of data journeys from production to application. Further, these tools were mutually reinforcing, ensuring that once local agencies availed themselves of CHH information and instruments, they were encouraged to understand the city and its problems in terms of physical measures and classifications of "healthful" housing and neighborhood environments.

[3] Term "raw" will be used throughout as an actor's category. In this case it is seen as data that is unprocessed for a particular use and hence not "actionable" in a specific context (although for another actor, this data may well have been sufficiently organized for the task at hand). See "'Overcoming the bottleneck': Knowledge architectures for genomic data interpretation in the oncology domain", this volume.

[4] On this point of actionability see Cambrosio et al. They also note the distinction between "raw" and "action-oriented" data made by their actors in accordance with the processing of data for use.

2 Re-considering Housing Data

At the height of the Depression, President Herbert Hoover addressed the pressing problem of housing the Republic's rapidly growing population. The family, he declared, was "the social unit of the nation", and the home essential for its "greater happiness", a source of comfort, health, education, and morality. To provide homes for those of moderate to low-income, they needed "adequate investigation and study on a nation-wide scale".[5] In 1930 Hoover announced the President's Conference on Home Building and Home Ownership, marking the entry of the federal government into the housing field. Civic leaders, administrators, planners, architects, lawyers, social scientists and medical experts were brought together and organized into 25 sub-committees focused on issues such as slums, planning, finance, building types and homemaking. Six smaller "correlating committees" worked to pull together shared information, aims and methods regarding research, education, technology, legislation and administration.

A key member of the "correlating" Committee on Research was Edith Elmer Wood, a pioneer in the movement to realize public housing for the poorer sections of society. Wood had long argued for more objective research on which to build more successful housing policies. The organization and circulation of data identifying the housing shortage and its consequences for mental and physical health would, she believed, generate social and political support for the clearance of slums and the construction of affordable dwellings. The Committee on Research duly reviewed the work carried out by the other sub-committees and made an inventory of past studies on housing problems. Its conclusions were damning. In its examination of the "best" housing research and literature, its members were struck both by the "large mass of material" and its "inadequacy".[6] Unjustifiable assumptions had been made from the thin census data that existed and while there were some small scale studies on topics such as housing and tuberculosis, these were too localized, uneven in quality, and not comparable. It was the "fragmentary nature" of the data that was the most "outstanding revelation", and one that prevented adequate policies being devised.[7] They needed a more centralized organization of statistical data and a carefully annotated inventory of all past researches – those containing good methods and evidence stored, circulated and replicated, and those lacking merit or out of date, discarded. They also needed to find ways of ensuring that findings were utilized, as "many a housing survey revealing most undesirable conditions of living has resulted in no improvement of those conditions."[8]

[5] Preliminary Outline of the President's Conference on Home Building and Home Ownership. Edith Elmer Wood Papers, Avery Drawings & Archives Collections, Columbia University Libraries, Box 2, Folder 8.

[6] Preliminary Draft of Report of Committee on Research, Correlating Committee B, November 18, 1931, Wood Papers, Box 2, Folder 14.

[7] Ibid.

[8] Ibid.

Committee members felt that their overview contributed a more realistic "big objective": to "stimulate interest in research and then determine what research should be done."[9] Wood's role was critical in this regard. She had drawn up a list of the "basic facts which we need to know". This included the extent of poor housing in the nation, how many of these could be reconditioned and how many required demolition, the most effective methods of rebuilding and rehousing, and the distribution of income that determined what families could afford.[10] She suggested a "special national census of housing and income", or, at the very least, sample surveys supported by foundations.[11] While she had pushed, unsuccessfully, for a census of housing to be included in the decennial census of 1930, Wood's advancing knowledge and activism was rewarded with a string of advisory posts in federal and municipal government committees, as housing became one of leading policy concerns of the 1930s.[12]

First employed by the Housing Division of the Federal Emergency Administration of Public Works, Wood's role was to pull together, process and circulate the wide range of available data, "gathering together and interpreting material already in existence, either uninterpreted or differently interpreted".[13] In effect, Wood was gathering what would now be called "big data": she drew on a wide range of statistics from various social, economic, and public health surveys to quantify the extent of poor housing in the United States and correlate it with measures of various social and physical pathologies, such as juvenile delinquency, a focus of the quantitatively-minded Chicago sociologists.[14] Most significant and original was her building of data composites focused on specific cities, a means of compensating for the fragmentary nature of the data available. Prominent in these composites was data provided through a new survey technology, the Real Property Inventory (RPI) in 1932. This had been promoted by the real estate industry and was "drafted by officials who were somewhat commercially-minded."[15] The RPI, devised by the statistician Howard Whipple Green in Cleveland, was organized around the census tract, rigid

[9] Mr. Gow (James Steele, Falk Foundation), First Meeting of Correlating Committee B on Research, September 11, 1931, Wood Papers, Box 2, Folder 12.

[10] Committee on Research, Basic Facts Which We Need to Know, (Suggestions from Mrs. E. E. Wood), October 5, 1931, Wood Papers, Box 2, Folder 13.

[11] Ibid.

[12] The census of 1930 carried only 2 questions dealing with housing, namely, "home owned or rented" and "value of home, of owned; monthly rental, if rented".

[13] Wood to James Ford, Director, Research on Slums and Housing Policy, August 7, 1935, Wood Papers, Box 23, Folder 8. The Federal Emergency Adminstration of Public Works, later the Public Works Administration, was a construction agency established in 1933 to build in response to the Great Depression, stimulating the economy through employment and investment opportunities.

[14] Particularly important for Wood was the work of Clifford Shaw in the late 1920s and 30s which, in privileging an ecological understanding of delinquency, suggested to Wood a "relationship between congestion and bad conduct" - Wood, "What do delinquency areas prove?", Wood Papers, Box 69, Folder. 1 See Shaw et al. (1929).

[15] Report of Central District, State of New Jersey, State Housing Authority, Arthur J. Quinn, Central District Manager, 1935, Wood Papers, Box 15, Folder 17.

geographic units which allowed for the scientific mapping of urban areas and establishing disparities in health, wealth and social well-being.[16] From 1934 to 1936, the RPI was applied to 64 cities in 48 states under the direction of the Department of Commerce as part of the New Deal work relief program.[17] Comprising what was essentially a market survey in real estate, construction, and household equipment, the RPI collected and processed reams of data to generate a statistical portrait of a city. It carefully avoided analysis and left interpretation to individual users, the introduction to a New York City RPI declaring the uses of its data to be "probably as many as those of the Federal census."[18] The RPI included data on housing quality, defining a "substandard" dwelling to be one in need of major repairs or unfit for use, lacking in private flush toilet, bathing unit, running water, installed heating and electricity or gas for lighting. It also collected and published information on rent and occupancy. Wood promoted the RPI as the standard source of urban housing statistics, while reworking its data which had originally been generated in the interests of "hard-boiled" and unsentimental businessmen, to show just how bad things were.[19]

Wood's resulting volume, *Slums and Blighted Areas in the United States*, published in 1936, provided one of the most comprehensive overviews of the housing problem across the nation. It proved very successful, the go-to source for housing information. It was reissued in 1938 by the newly founded United States Housing Authority (USHA), an organization focused on the provision of public housing following the landmark Housing Act of 1937. Wood was then invited by the USHA to bring the volume "up-to-date and present additional graphic material."[20] In *Introduction to Housing: Facts and Principles*, Wood used the RPI as a base map on which to build a narrative that illustrated the threat posed by poor housing to American society and democracy. She adapted an earlier technique devised by Whipple Green of using transparent maps that spotted cases of disease, crime, vice, or delinquency, overlying a color map of monthly rentals by census tract.[21] Carefully selecting a series of "statistically minded" cities, Wood constructed rate and spot

[16] Originally defined as "sanitary areas", census tracts were a method developed by public health services in several cities and incorporated into the census in 1910s. By the 1930s, largely through the work of Whipple Green, they became more widely established and used by an increasing number of agencies to compare health statistics with a broad range of socioeconomic data. See Krieger (2006).

[17] RPI employed architects and engineers for enumeration and tabulation with federal funds, who would otherwise be unemployed during the Depression.

[18] Thomas S. Holden, "Foreword", in Real Property Inventory, *City of New York. Volume 4* (New York City Housing Authority, 1934), p. vii.

[19] The origins of the RPI were very significant for Wood, allowing them to counter common accusations of "sentimental bias" and for having exaggerated the failure of private enterprise to supply decent housing to unskilled labor. Wood, "Existing housing conditions in the United States", Prepared for annual meeting of Milbank Memorial Fund (MMF), April 1937, Wood Papers, Box 23, Folder 3.

[20] Catherine Bauer to Wood, May 20, 1938, Wood Papers, Box 11, Folder 14.

[21] See Wood (1936), for a discussion of this earlier method. She corresponded with numerous researchers and housing agencies to gather this statistical data.

maps, illustrating graphically through dots and crosshatching the relationship between low rental areas and disease (tuberculosis) and delinquency (criminal convictions) (Wood 1936, 1940).[22]

Wood demonstrated how to adapt, combine and circulate data from a variety of different sources. Her work was, as she noted, "quoted constantly" in the push for policies of slum clearance and public housing.[23] Wood's many facts and her identification of important of information sources, helped secure a housing census for 1940.[24] As the planner Warren J. Vinton declared, a complete record of the nation's homes, "like a mariner's chart, will enable us to steer our programs safely and accomplish the results which the Congress expects of us."[25] They would now have nationwide, unified and comparable information on the characteristics of residential structures, occupancy status, rental and home value, the unit's equipment, facilities, furniture and utilities, and home finance. Once the housing data was processed and published in a series of bulletins, it could be cross tabulated with data on family size, composition, economic status from the regular population census schedule.[26]

While housing reformers celebrated this advance from scattered local surveys, to the RPI, to a census of housing, there remained limits to use of this information. The census and RPI provided, as the urban planner Anatole Solow noted, "gross not precise data".[27] Only a few items were covered and categories such as "minor repairs needed" or "unfit for use" were general and loosely defined. As a member of the RPI unit in Washington admitted to Wood, the data on housing quality served merely to "roughly classify the buildings as to their need for repairs and demolition".[28] This data could be used to identify shortages and establish the approximate size of housing problem to politicians and the public, but it could not be used to specify how that problem was best resolved on the ground. It was useful for making a case

[22] The choice of Richmond was due to the quality of data available and the size of the city. By using a small city, they could "illustrate the principle of spot maps to show the correlation between slum areas and unfavorable social conditions." Bauer to Dr. Kimball Young, University of Wisconsin, December 26, 1938, Wood Papers, Box 11, Folder 15.

[23] Memorandum, Wood to Administrator, January 16, 1941, Wood Papers, Box 14, Folder 1.

[24] Particularly popular was her off-cited estimate that one third of the population was ill-housed, see Wood, "Existing housing conditions", ibid.

[25] Testimony of Warren Jay Vinton Before Census Committee of the House, July 13, 1939, Warren Jay Vinton Papers, Division of Rare and Manuscript Collections, Cornell University Library, Box 1, Folder: Census of housing.

[26] The Census Bureau sponsored a series of housing bulletin's that presented interrelationships between certain housing characteristics and other census data, such as characteristics of families or households occupying dwelling units. It also sponsored the production of analytical maps presenting these various housing characteristics by blocks to aid the location of problem areas and areas with inadequate housing.

[27] Anatole A. Solow, The Measurement of Housing Quality and Need: Public Health Gives a Practical Tool for Planning Action, May 1, 1947, Charles-Edward Amory Winslow Papers, Manuscripts and Archives, Yale University Library, Box 54, Folder 517.

[28] Daniel Casey, Real Property Inventory Unit, Department of Commerce, to Wood, August 18, 1934, Wood Papers, Box 5, Folder 6.

for action; much less so for the mode of action itself. What was critical, therefore, was the need for data that could be transmitted and mobilized in a way that was actionable, data that served the practical purposes of planners, architects and public health workers employed by local government agencies.

3 Public Health and Housing Standards: The Committee on the Hygiene of Housing

It was in the interest of generating more practical data which would travel into the plans and designs of municipal governments, that the CHH was founded in 1937. The Committee was organized by one of the nation's leading figures in public health and the founder of Yale's Department of Public Health, Charles-Edward A. Winslow.[29] Winslow described the home as an "instrument of health… in the wide sense of emotional and social as well as physical well-being."[30] The CHH functioned as the "technical housing body" of the APHA, conducting research in aspects of housing design, construction and occupancy which affected mental and physical health.[31] It also served as "national clearing agency", sifting through and distributing information from existing studies (Solow and Twichell 1947, 22). By bringing these "results… to administrators and technicians in the fields of public health and housing", the CHH was establishing itself as an important intermediary agency and was, in effect, bringing public health officials into the "national housing program".[32]

The evidence generated by the CHH would travel from the experiments and surveys of physical, social, biological and medical scientists to housing administrators and planners by means of a series of standards promulgated through its published reports. Standards covering occupancy, sanitation, light and air, had long played a central role in housing reform, a way of establishing and enforcing clear and workable codes to improve tenement living and defend against the unscrupulous operations of landlords and speculative builders. But, influenced by cultural norms and social interests, they had also been very inconsistent across the nation.[33] The CHH would work to develop an extensive series of standards on which "comprehensive housing regulations" could be built, as was desired by health, building and housing

[29] The Committee was organized on the request of Housing Commission of the Health Organization of the League of Nations, in which Winslow played a leading role, and as one of the national committees corresponding with that Commission. Winslow founded the Yale Department of Public Health in 1915.

[30] Winslow, The Physiology of Shelter, June 22, 1948, Winslow Papers, Box 129, Folder 731.

[31] Committee on the Hygiene of Housing, Statement of program, 1939, Wood Papers, Box 22, Folder 16.

[32] Ibid.

[33] Dread diseases were critical to the cause of early housing reform; the threat of tuberculosis and cholera demanded clean water and light and air – see Lubove (1962).

officials.[34] They would also be objective, and thus irrefutable, based as they were on rigorous scientific inquiry.

In 1938, the CHH published the first of a series of influential documents, one that would help place health at the heart of the housing problem. Their pamphlet, "Basic Principles of Healthful Housing", was described as a "preliminary attempt" to formulate the basic health needs to be served through housing. Its principles were defined as "fundamental minima required for the promotion of physical, mental, and social health", which were, in turn, based upon "fundamental biological requirements" (CHH 1938, 354). Thirty principles were then divided into four sections - physiological and psychological needs and protection against contagion and accidents. Discussion of the requirements needed to realize each principle consisted of a careful and deliberate exposition of the relevant facts, such as the need for an air change of 10 cu. ft. per person per minute to dilute atmospheric impurities. The standards proposed were tentative at this point, based on existing data that was drawn from a wide range of studies and agencies. But they noted that this data was incomplete in many areas. While the National Health Survey supplied epidemiological data on household accidents, for example, the relationship between housing and disease could not be shown through "the usual lines of investigation."[35] To successfully translate their principles for health into "concrete standards of performance for the home of the future", they needed new forms of interdisciplinary research (Winslow 1945, 20). Such ongoing studies would, in turn, continually modify such standards, making them more "precise and scientific".[36]

The CHH established a series of research committees focused on specific problems, on which appointed members were leading experts. These included physical and engineering aspects, such as building construction and household equipment, administrative and legal problems, and social and human uses. For example, a subcommittee on standards of occupancy, on which Wood served, focused on space requirements and the maximum number of individuals to be housed in each type of dwelling. They submitted ongoing reports, published in the *American Journal of Public Health* and printed as individual pamphlets by the U. S. Public Health Service, which were circulated, reviewed and abstracted by leading public health, construction, real estate and housing associations. The work of each subcommittee culminated in a detailed report that was published as part of a series on "standards for healthful housing". Winslow commented on the concluding volume on standards for occupancy published in 1950, they had presented "for the first time actual

[34] The CHH began its work with a canvassing of opinions from officials and agencies as to the most feasible types of housing control. Thirteenth Meeting of the Committee on the Hygiene of Housing, American Public Health Association, Tentative Report of the Subcommittee on Housing Legislation and Administration, February 2, 1942, Wood Papers, Box 22, Folder 16.

[35] Woods reported in Minutes of Meeting of Sanitation Advisory Committee, Washington DC, June 5, 1939, p. 6, Winslow Papers, Box 19, Folder 482.

[36] Winslow, Report of the Round table on the Hygienic Aspects of Housing, MMF, 1937, Wood Papers, Box 23, Folder 5.

concrete data on the space needed for families of various sizes."[37] The minimum space requirements were stipulated in relation to the number of individuals – 400 square feet for one person, 750 for two, 1000 for three, and so on - based on laboratory studies that measured, for example, the "atmospheric impurities" that resulted from cooking, heat sources and the human body. Yet they were also concerned to move beyond such absolute standards and grapple with the "actual conditions of occupancy", that is, how space was used by a family in its day-to-day life (CHH 1950, vi). To achieve this, they gathered together observational data on family life provided by the Swedish sociologist and subcommittee member Svend Riemer and their long-standing collaborators, the John B. Pierce Foundation's research laboratory in New Haven, a center for the study of physiological regulatory systems.[38] The result was a much more practical series of requirements that stipulated floor space in relation to family needs and activities rather than merely relying on floor area or cubic content as had been the norm in earlier housing regulations which had left "space -- the most valuable commodity housing has to offer -- … poorly designed or wholly insufficient."[39] Through their detailed examination of livability they delivered an extremely detailed list of specifications, from room sizes and their design, to floor space in relation to furniture, the placement of the bed relative to windows, the distances between cots, and the dimensions of closets and work-spaces.

The volumes were packed with research data. This gave the standards credibility, showing them to be based on the best scientific information available. It also provided a degree of flexibility. While CHH members agreed with the provision of absolute standards regarding space, such as the 400 cu ft. minimum, they also expressed concern that these could too easily become rigid, static and fixed, "not only… frozen in the minds of the designers but… made unalterable in the form of permanent buildings."[40] Modern family needs did not remain stationary and, with the growing recognition of the need for optimum space for psycho-social well-being, such crystallized standards would "likely to prove a drag upon progress

[37] Winslow to Frank Boudreau of the Milbank Memorial Fund, one of the main financial supporters of the CHH, April 27, 1950, Winslow papers, Box 54, Folder 513. The work the subcommittee was suspended during the war, as the work of other subcommittees was considered more relevant to federal agencies during the emergency, hence the delay in this final publication.

[38] Notes of Report for Dr. Maxcy to the Governing Council, 1941, Winslow Papers, Box 54, Folder 515; CHH, APHA, Essentials of Space Planning and Space Organization in Dwelling Units, Report of Subcommittee in Standards of Occupancy, March 1942, Wood Papers, Box 22, Folder 16. The CHH also worked with a wide range of agencies across the United States, carrying out field studies into conditions of heating, ventilation, lighting, and noise in summer and winter in New York City, New Haven, and Charleston, Oklahoma City, and Tennessee Valley. APHA, CHH, Statement of program, 1939, Wood Papers, Box 22, Folder 16.

[39] CHH, APHA, Essentials of Space Planning, ibid.

[40] CHH, APHA, Subcommittee on Standards of Occupancy, Principles of Space Planning and Space Organization for Low-rent Dwelling Units, Revised Draft Submitted for Criticism of the Subcommittee, prepared by Anatole Solow in cooperation with Allan A. Twichell and Harold Sandbank, March 28, 1941, Wood Papers, Box 23, Folder 1.

rather than a stimulus to progress."[41] By including such a wide-range of data, the CHH reports enabled local authorities to adapt the standards, mobilizing the data contained within them to suit their own particular requirements in terms of climate, building materials and equipment, costs and rental value, and the housing needs of particular types of families: "Local building agencies must have some leeway; otherwise there can be no variety, no adaptation to regional needs, no experiments, and therefore no real progress."[42] Consistent with features deemed to characterize successful databases, the volumes or manuals constructed by the CHH needed to be general and robust enough to encourage circulation, but also adaptable to local demands and situations.[43]

In this way, standards served as the conduit or vehicle for the movement of data from laboratory and field studies and into the plans and designs of the numerous housing associations that oversaw the mass construction of new dwellings following the 1937 Housing Act and the regulatory bodies that controlled existing housing, usually overseen by local departments of health. The CHH had worked through, simplified and condensed masses of complex data, ranging from the physical issues of construction and sanitation, to the social, addressing the dual needs for individual privacy and the opportunity for family life through design. This tidying and organizing of data is comparable to the processes of "cleaning" data so that it is amenable to analysis that is discussed by Boumans and Leonelli in this volume.[44] Drawing from Mary Douglas, Boumans and Leonelli argue that cleaning does not involve the removal of dirt, but is about ordering and classifying. In their studies of economic and plant science, this was achieved through "clustering" data into larger units of interrelated objects. The CHH similarly established groupings of data according to use by establishing standards relative to various aspects of housing, be it circulation, ventilation or occupancy. But here the categories were organized with an explicit emphasis on policy-usefulness, those relating to more fundamental problems of building structure clearly distinguished from those of building occupancy which could be attended to by local law enforcement or public health workers, for example. These were then communicated to the reader through the series of circulated reports or manuals which neatly divided and labelled the relevant information and contained clear and usable tables and charts to apply to a wide range of housing issues.

The CHH standards had considerable influence, demonstrated in new building code requirements of the National Bureau of Standards of the US Department of Commerce, the housing codes promoted by the National Association of Housing Officials (NAHO) with whom the CHH worked closely, and also regionally and

[41] Letter, Winslow to Bleecker Marquette, April 25, 1941, Wood Papers, Box 23, Folder 1. It is worth noting that Winslow argued that the public health department should have the role of adapting and improving space standards, rather than a housing agency.

[42] A Housing Program for Now and Later, February 1948, National Public Housing Conference (NPHC), Vinton Papers, Box 2, NPHC-Releases.

[43] On this point, see Cambrosio et al. who draw in turn on Leonelli (2013).

[44] M. Boumans and S. Leonelli, this volume.

locally, in committees of planning and housing regulation such as in New York City.[45] The CHH also pushed aggressively for improved standards in federal agencies charged with housing construction, declaring the lack of standards in many of the homes built for the defense industries during the war to be "shocking" and to constitute a "national scandal."[46] At the request of the USHA, they reviewed the government housing and occupancy standards for public housing, urging them to increase their room size specifications, lest they "produce a nation of neurasthenics".[47] The work of the CHH also stimulated further housing investigation and the sharing of data between experts and agencies, as Allan Pond, a public health expert and CHH member proclaimed: "Interest in housing standards and building code requirements currently is widespread and feverish. On every hand there is evidence that house design and construction standards and methods are subjects that attract the imagination of technicians and the public alike. Laboratories are humming with research activities designed to shed further light on new materials and modes of construction."[48]

4 The Appraisal Method: Transforming Standards Back into Data that Travels

The standards generated by the CHH expert committees had another important potential use. Whereas the RPI and census had provided rather crude measures of housing quality, planners and policymakers now had the means of clearly distinguishing good housing from bad, of identifying precise faults and their patterns. The CHH now sought to translate standards into a yardstick for measuring housing conditions, as public health and housing officials requested further help in ensuring their policies were "better guided" (CHH 1942, 285). The CHH now set itself a new task: "developing a method of data analysis whereby final results could be readily summarized and interpreted by local health departments and various other agencies as a guide for their policy and practice." This would supplement census and city-wide housing surveys which were useful in identifying general problem areas, but, with their breadth and generality, the "collected data do not readily lend themselves to a variety of purposes for local government agencies concerned with housing"(CHH 1942, 286).

---

[45] Densities in New York City: A Report to the Citizen's Housing Council, by The Committee on City Planning and Zoning, May 1944, Henry S. Churchill Papers, Division of Rare and Manuscript Collections, Cornell University Library, Box 2, Folder 24.

[46] Vinton to Adminstrator, United States Housing Authority (USHA), September 23, 1941, Vinton Papers, Box 28, Defense Housing Program.

[47] Wood to A. C. Shire of the USHA, November 3, 1941, Wood Papers, Box 1, Folder 21. It was this lengthy debate among and between members of the CHH and USHA that helped stimulate the CHH inquiries into family living habits and their use of space in the home. Minutes of Fourth Meeting of the Subcommittee on Standards of Occupancy, March 28, 1941, Wood Papers, Box 23, Folder 1.

[48] Pond, The application of health standards to house construction, to Annual Meeting, Connecticut Society of Civil Engineers, March 17, 1948, Martin Allan Pond Papers, Yale University Library, Box 13, Folder 229.

In order to transform standards into a workable survey technology, the complex of specifications needed to be simplified, as to conform precisely, surveys would become so large and data-laden they would lose their practical value. The solution was a "screening method": an index consisting of a limited number of factors selected as indicators or proxies for a multitude of housing characteristics.[49] For example, the presence of an inside flush toilet was not selected as an item simply because of an intense interest in the facility being present, "but because of its assumed intrinsic meaning as one element in an index of hygienic housing" (CHH 1942, 287). Index items were also selected according to the degree to which they lent themselves to precise and objective measurement, as identical information needed to be collected by different enumerators. The result was a series of items that established the quality of the building itself, its structural integrity, sanitary and heating facilities, housekeeping and facilities, its occupancy, such as area per person and number of persons per room, and its surrounding neighborhood environment, considering specific industrial nuisances, the density of land coverage, usability of open spaces, public utilities and community facilities, and specific hazards, such as heavy traffic and noxious odors.

The second innovation of the new survey method was its scoring system that generated a new dataset. A series of penalty points were scored on a scale which captured any departures from the standards of acceptability as derived from the basic principles of healthful housing. These points were weighted: very serious issues, such as the lack of a safe water supply, granted 30 points, more minor deficiencies, 1 or 2 [see Fig. 1]. These were then added together to give an overall score, and the building then placed into one of a series of quality classifications ranging from good, A, to bad, E. This method provided a more detailed and accurate analysis of housing quality and removed bias through short standardized schedules "which call for practically no subjective judgment".[50] The enumerator could move quickly and only a few days training in the technique were necessary, Winslow noting: "It's very simple."[51] The schedules were then processed by a skilled clerk who did not need to see the dwellings. Using scoring templates and summary appraisal forms, the clerk could quickly translate the data on the field schedules into numerical scores. This was then transferred to cards of the marginal punch type, allowing "rapid sorting and tabulation. The data obtained is readily analyzed and yields a measurement of housing deficiencies on a valid quantitative basis."[52] Further, by mapping out these classified buildings over an area, tabular data could be used in a

[49] For a detailed analysis of the relationship between a larger body of data and the selection of indicators to help make sense of a more complex set of phenomena and their ease of travel, see Mary Morgan's analysis of data and datum in this volume. Unlike Morgan's case, the set of indicators developed by the CHH were, necessarily for actionability, tightly bound together.

[50] E. R. Krumbiegel, "An appraisal method for housing conditions and needs: Milwaukee enforces a new housing code", Reprinted from *The Municipality*, December 1945, Pond papers, Box 13, Folder 230.

[51] Winslow, Housing Principles, May 3, 1944, Winslow Papers, Box 110, Folder 198.

[52] Krumbiegel, "An appraisal method", ibid.

Selected Deficiencies of Dwellings and Neighborhood Environment
Sample Areas Grouped by Quality Grade
Southwest District, New Haven, Conn.

Scoring Item Number	*Deficiency*	*Qualifying Range of Score: Penalty Points* [1]	*Quality Grade of Sample Area*		
			A and B Combined	*C*	*D and E Combined*
			Per cent of Dwelling Units Incurring Penalty Scores within the Qualifying Range (preceding column)		
FACILITIES	I. DWELLINGS				
2	Public Hall Daylight: Grossly Inadequate [2, 3]	5–10	0	0	10
5	Daylight Obstruction by Adjacent Structures: Serious [2]	5–15	18	28	44
8	Piped Water: Cold Only or None in Unit	7–15	4	20	42
9	Bathing Facilities: None, Shared, or No Hot Water	7–23	6	22	44
10	Toilet Facilities: Shared, Outside Unit, or Non-flush	10–40	2	1	9
12	Windowless Rooms: One or More	15–20	0	0	6
13	Installed Heating: None in at Least One-half of Rooms	10–18	7	60	64
15	Room Sizes: Area of One or More Rooms Substandard [2]	5–10	19	17	37
MAINTENANCE					
16	Yard Condition: Grossly Insanitary [2]	10–15	0	15	26
18	Structural Deterioration: Extreme [2]	20–30	3	15	37
OCCUPANCY					
21	Persons per Room: One and One-half or More	10–25	10	18	22
22	Area per Person: Substandard [2]	10–25	1	4	13
	II. NEIGHBORHOOD ENVIRONMENT		*Per cent of Blocks or Street Frontages Incurring Penalty Scores within the Qualifying Range*		
E 1	Land Coverage by Buildings: Excessive [2]	10–24	0	10	17
E 5	Land Use: 30 Per cent or More of Block Area in Industrial, Commercial or Mixed Residential Use	10–13	2	38	58
E 7	Specific Nuisances and Hazards from Non-residential Sources: High Incidence [2]	18–30	6	21	44
E 8	Moral Hazards: Considerable in the Area [2]	6–10	0	15	24
E 10	Hazards and Nuisances from Adjacent Streets: Considerable [2]	15–20	2	26	25
E 21	Public Playgrounds: Beyond Reasonable Distance [2]	8	4	44	69

1. For most of the deficiencies the range of possible scores begins with 1 or 2 penalty points. In order to show here only the really significant defects, those dwellings or street frontages with slight penalties for any item have been excluded.
2. Space limitations preclude an accurate statement here of the criteria on which this item is scored. As noted in the text, all deficiencies are reported in terms of objective characteristics, not in such loosely descriptive terms as are necessary here. Scoring is done from precise rating tables.
3. Applicable only to tenements with public halls.

Fig. 1 A table showing the scoring and classification of dwellings according to the appraisal method. From CHH (1943)

geographical form, creating a "sketch portrait of the slum block" (CHH 1942, 292). As the CHH demonstrated in a pilot survey in New Haven, by sampling every seventh dwelling in a problem area of city, they could map the quality of individual blocks or individual sections within blocks.[53] The maps provided a graphic representation of the data collected, identifying which areas were beyond saving and needed to be torn down, or areas which could be rehabilitated through some treatment to prevent further deterioration. The housing assets and liabilities of a city could now be accurately mapped to identify what problems existed and where they were concentrated: "It becomes possible to report objectively to the municipal administration the state of housing in any problem area."[54]

The technology therefore simultaneously produced data while processing and packaging it for travel into the policies and plans of local authorities.[55] The series of reports on the new "appraisal method" published by the CHH from 1942, and its three final volumes from 1945, were well received by housing and public health authorities, and even more so in the postwar era, a period of so-called urban renewal, whereby the large-scale building of new housing was to be tied to the mass clearance of slums dwellings across the nation. Programs of urban redevelopment intensified following the 1949 Housing Act which increased federal funding, but also intensified government oversight regarding the kinds of housing to be removed, renovated, and rebuilt. As a health officer in Milwaukee observed, the appraisal method was particularly well-suited to this new and more expansive approach, the objective and sharp demarcation of problematic urban areas providing for systematic and long-term programs of rehabilitation, demolition and reconstruction, replacing the "futile patch-work" of laws and regulations focused on individual dwellings on a case-by-case basis.[56] Following Milwaukee's adoption of the appraisal technique in its program of redevelopment, its city officials encouraged Philadelphia to follow suit, David Walker of the city's Redevelopment Authority declaring the CHH's yardstick to be "a most scientific method".[57] Decisions were not left to the personal judgment of the inspector, but scored objectively and "the mechanical brains of a punch card rate the quality and quantity of blight and give us an adequate appraisal of the neighborhood" (Walker 1947, 70).

The technology also encouraged closer working relationships among agencies in the city. In Milwaukee, Philadelphia, and soon Los Angeles, St. Louis, Washington DC, Baltimore, Boston, and Portland, Maine, Anatole Solow saw "the beautiful words" of "integration and cooperation", so often used in planning literature, now

[53] Winslow, "Housing Principles", ibid. The CHH carried out survey trials in 3 cities in Connecticut, New Haven, Waterbury and Stamford, testing the items and rating technique and identifying their uses for local authorities.

[54] Solow, "The measurement of housing quality", ibid.

[55] As Sabina Leonelli (2015) argues "Packaging happens at several stages of data travel and is often implemented already at the point of data production".

[56] E. R. Krumbiegel, "An appraisal method", ibid.

[57] Milwaukee was the first city to adopt the appraisal technique, outside of trial studies, completing a survey of a 16-block substandard area of the city in 1945. The field secretary of the CHH, Emil A. Tiboni, instructed the city's Health Department and Land Commission personnel in the use of the method (reported in City News in Brief, *Journal of Housing,* November, 1945, p. 204).

being realized, the police, health, building, and fire departments all using the relevant punch card data to fulfill their specific roles in the problem areas as demarcated by the appraisal method.[58] By providing a continuous record of local housing and neighborhood conditions, over time the method could generate the unity essential for successful policy. The diverse range of actors and agencies involved in housing could now visualize and interpret housing data in consistent and actionable ways. Solow described the survey as a "skeleton which gives strength to the body of planning programs. In the field of housing, a type of skeleton is now available which should permit more action than the mere rattling of bones."[59]

But in doing so the survey privileged a public health perspective in the resolution of urban problems. The data that had been used to create and legitimate standards of healthful housing, was now stripped down and simplified for travel through translating those standards into index items to map urban areas. Further, once an area had been classified, its faults dissected and listed in the local authority's survey report, the most effective means of correcting these failings, by either rehabilitating the housing that could be saved or demolishing and rebuilding in a way that would prevent future obsolescence, were to be found by turning back to the CHH standards for healthful housing. The result, therefore, were two powerful and mutually reinforcing technologies that encouraged urban agencies to understand housing in terms of preventing physical illness and accidents, disease transmission, and emotional disorder.

Finally, the appraisal method served a useful research tool. Committee members had long sought firmer evidence that better housing improved health and well-being. While the CHH publications were strong on physical illness, disease and accidents, mental health and social well-being were much harder things to measure. In their early reports they had, like Wood, relied on correlations between poor housing areas and data on delinquency and mental hospital admissions, as well as the statements of psychiatrists and social workers. In 1945, a new Joint Committee on Housing and Health was established, bringing together the CHH and NAHO; its purpose "to study the actual results of the provision of good housing" and then "translate" this information into "into the administrative practice of operating housing agencies".[60] The culmination of this committee's work was a further collaborative study with the Johns Hopkins School of Public Health and Hygiene in Baltimore.[61] The work of the Johns Hopkins Longitudinal Study of the Effects of Housing on Health and Social Adjustment began in 1954, its director, the social psychologist, Daniel Wilner, describing it as the first systematic survey that analyzed "a discrete quite measurable change in physical environment on behavior and health." It compared the mental and physical health of those in "very bad slums" of Baltimore with those relocated to new "very good housing", a

[58] Solow, "The Measurement of Housing Quality", ibid.

[59] Solow, "The Measurement of Housing Quality", ibid.

[60] Association News, Initial meeting of Joint Committee on Housing and Health, *Journal of Housing*, July 1945, p. 119.

[61] The selection of Baltimore was largely the result of the active role played by Huntington Williams in the CHH. The study was housed at the Baltimore City Health Department and funded by the APHA. See George Huntington Williams Collection, Alan Mason Chesney Medical Archives of the Johns Hopkins Medical Institutions, Box 505.

modern high-rise project for black Americans.[62] In this survey, the CHH appraisal method allowed them to first divide up areas of the city into ecological units or experimental zones of good and bad housing, and second, by using the precise appraisal data, relate specific features of the housing to specific social and psychological factors collected through a series of "psychosocial scales." Wilner's study duly showed how improved housing had led to lower rates of sickness, improved rates of school attendance, and emotional well-being (Wilner et al. 1962). By turning the appraisal method into a research tool, the CHH was able to generate and integrate data that further promoted the value of its principles for healthful housing, using the very environments that its standards and surveys had created as laboratories for testing and legitimating the principles that had helped give them birth.

5 Conclusion

By the 1960s, CHH's standards and appraisal technique were used across urban America and endorsed by national and federal associations of construction, public health, architecture and urban planning.[63] The "mariner's chart" desired in the 1930s had not been provided through a census database, but through the careful processes of curation, presentation, and packaging that made data actionable in local contexts. The striking success of the CHH can be attributed first, to its ability to make data travel; second, its ability to continue exert control over the data being produced, circulated, analyzed and acted upon. Through a series of phases, information moved by means of the vehicles constructed by the CHH and into the practices of local authorities in urban planning, health and design. First, a sequence of principles brought together existing data on a wide variety of topics to construct an argument for the necessity of healthful housing. Second, these principles were detailed by formalizing them into standards. These worked to translate a mass of complex information from a wide range of sources into clear and accessible manuals which tidied, organized and labeled data and provided flexible guidelines that allowed it to be applied to local situations. Third, the appraisal technique then translated these standards into a workable diagnostic tool that generated simple and concise information while simultaneously suggesting policy solutions. Finally, the appraisal functioned as a research tool, a means of generating further data that gave credibility to the original principles of healthful housing. We have, therefore, a certain circularity in

[62] Daniel Wilner in transcripts of Conference on the Physical Environment as a Determinant of Mental Health, Washington DC, May 28–29, 1956, John B. Calhoun Papers, National Library of Medicine, Box 63.

[63] Where they were not used, they were often in some way adapted and simplified. One of the most common complaints was that, in spite of the emphasis on simplicity, the surveys were in fact complex and expensive to complete. In Philadelphia, rather than redoing the appraisal when they needed more updated information, they devised their own version which simplified the CHH index making it less costly. Planning Division, Redevelopment Authority of the City of Philadelphia, Summary Report on the Central Urban Renewal Areas (CURA), March 1956, Churchill Papers, Box 3, Folder 27. On the CHH appraisal method and urban renewal, see Abramson (2016).

movement and a reinforcing relationship between the tools devised by the Committee. Through these technologies, the CHH was able to control data at each critical phase of movement, from identifying and circulating the problems of poor housing to generating actionable evidence for local agencies. Both the standards and appraisal technique offered authorities instruments of regulation over housing and occupancy, powers that were in turn dependent on CHH's designation, in hierarchical, tabular and cartographic forms, of the data that mattered.

The ingenuity and creativity of the CHH in making masses of complex data retrievable and actionable by a wide range of disciplines and professions ensured the centrality of the public health field in urban redevelopment from the 1940s, so much so that one official observed: "It is becoming increasingly difficult to know whether health is ancillary to housing or housing is ancillary to health."[64] When the National Commission on Urban Problems was appointed by President Johnson to address the crisis of urban unrest and violence in the late 1960s, they turned to the issue housing standards. In their report of 1969, the CHH was credited for its role in the birth of the first "modern housing code" and for showing that so many cities in the United States were failing to provide the quality of housing so critical to the health and wellbeing of their citizens (Mood et al. 1969, 10). The infrastructure generated by the CHH was sound, the Commission declared, but now needed to be updated and strengthened with more research into health and housing as there was a "paucity of valid data" (Mood et al. 1969, 33).

However, for a growing number of critics, it was precisely this entrenched and uncritical acceptance of quantitative data, housing codes and standards that was the problem. Some social scientists and activists saw the CHH as having been far too successful. While Committee members had urged flexibility and regular revision, worried that minimum standards could become obstacles to future progress, for critics it was the entire infrastructure developed by the CHH that was problematic.[65] Local governments and agencies were trying to resolve social problems through instruments that simply could not account for the complexity, variety and dynamism of urban life however regularly they might be informed by new data. In Boston the appraisal method had been put to work in the West End, designating most of the housing in a 48-acre site obsolete and beyond rehabilitation. The West End project report described the buildings as "dilapidated" and the area as "overcrowded" with a "severe lack of any open space."[66] On the basis of these findings, the Boston Redevelopment Authority razed the site, displacing some 2700 families, to make

[64] John C. Leukhardt, "Health centers and health services in housing programs", 18th Annual Conference, Milbank Memorial Fund, April 2–3, 1940, Wood Papers, Box 23, Folder 6.

[65] As the editors have suggested, Bowker and Star's notion of "infrastructural inversion" is a useful way of conceiving of the growing challenges to urban planning in this period emanating from the social sciences, as critics began to question the long held assumptions that advances in housing quality were to come from ever stronger and more precise housing codes and standards to which home-owners, landlords and tenants would be forced to comply. See Bowker and Star (2000, 34–46).

[66] West End Project Report: A Preliminary Redevelopment Study of the West End of Boston, March 1953, Urban Redevelopment Division, Boston Housing Authority, Herbert Gans papers, Rare Book & Manuscript Library, Columbia University, Box 2, Folder 3. See also O'Connor (1993).

way for five residential high-rise (and high-rent) apartment complexes that fulfilled the CHH specifications. The case of the West End became celebrated by opponents of urban renewal thanks largely to a research project carried out under the direction of Erich Lindemann, Chief of the Psychiatry Service at the nearby Massachusetts General Hospital. The West End project, titled "Relocation and Mental Health: Adaptation under Stress," sought to devise new methods of understanding how individuals and communities adapted to severe stresses, such as the loss of home, and build a more effective social and psychiatric support network (Ramsden and Smith 2018). The project's publications, based on in-depth interviews and participant observation, documented the devastating effects of urban renewal on a community and criticized the CHH methods of classification (Fried and Gleicher 1961; Gans 1962). The appraisal method relied on a series of items that could be objectively measured and translated into action and hence, they were straightforward, physical and quantitative. They allowed large sections of the city to be mapped and classified for demolition. While the CHH had enabled the clearance of slums and the building of new and better housing throughout the United States, with this attention to the physical, it could not hope to capture the complex social lives and varied need of different communities in the city. The portability of data had come at the cost of its detachment from the lived experiences of city dwellers and the social meanings of shared urban spaces. While the West End was crowded, its buildings dilapidated, it was a healthy and mutually supportive working-class community.

Having witnessed the power of the CHH's technologies in the West End and using this experience as an exemplar to contest and critique methods of data processing and application, social and behavioral scientists began to demand, collect and circulate new kinds of data. They devised questionnaires, surveys and observational studies that could better capture, organize and translate how people experienced space and how it could be better designed in accordance with the interests of the users. While the APHA would continue to update its guidelines on healthy housing through to the 1980s (Mood 1986), long after the CHH had disbanded, the unity between planning, design, construction, medicine and the social sciences, contributed to in no small part by the Committee's technologies, began to break down. With the CHH having played a critical intermediary role, there was now little to link together the large housing databases generated by the census and smaller user-oriented surveys of environmental quality applied by independent agencies. The influence of the CHH had waned gradually, and in the wider social and political climate of the 1960s and 70s, this was hastened by growing criticism of large-scale urban renewal and public housing development from a wide range of sources, not only fiscal conservatives and neo-liberal policymakers, but liberal critics of the "urban bulldozer" that appeared to demolish communities as well as slums, and simply served to shift poverty to other parts of the city (Anderson 1964; Jacobs 1961).

In this study, we have seen how the CHH functioned very successfully as an intermediary organization that helped generate and circulate data that was credible, authoritative, easily transferred and acted upon. In the various cases brought together in this volume, patterns of data journeys have been examined and critical issues, conditions, and practices of configuration, visualization, transformation and linkage explored. In the case of housing data, with its explicit practical role in planning and

designing the built environment, grounding physical interventions in the world, we have had an opportunity to examine just how accountability and actionability can be built strategically into data journeys. We have seen how public health and planning experts, housing activists, and policymakers were attracted by the promise of big and open data which would grant authority, credibility and power to housing reform. Yet they also wanted to ensure that data was usable in specific contexts. The CHH managed to combine elements of universality in terms of the objective facts of the healthy home and the minimum standards required to construct and it, with local demands for usefulness, adaptability and actionability. By virtue of its role in the development of a centralizing infrastructure of new technologies that could simultaneously generate, process, standardize, organize and circulate data, and further, make these technologies available to local authorities, the CHH was able to determine the kind of data that was put to work in planning, building and design, and thereby secure the public health perspective in housing policy throughout the nation.

References

Abramson, Daniel M. 2016. *Obsolescence: An Architectural History*. Chicago University Press.

Anderson, Martin. 1964. *The Federal Bulldozer: A Critical Analysis of Urban Renewal, 1949–1962*. Cambridge: MIT Press.

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., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and its Consequences*. Cambridge: MIT Press.

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.

CHH (Committee on the Hygiene of Housing). 1938. Basic Principles of Healthful Housing. *American Journal of Public Health* 28: 351–372.

CHH. 1942. An Appraisal Technique for Urban Problem Areas as a Basis for Housing Policy of Local Governments: Illustrative Results from Three Test Surveys, Report of Subcommittee on Appraisal of Residential Areas. *Public Health Reports* 57: 285–296.

———. 1943. A New Method for Measuring the Quality of Urban Housing—A Technic of the Committee on the Hygiene of Housing. *American Journal of Public Health* 33: 729–740.

———. 1950. *Planning the Home for Occupancy*. Chicago: Public Administration Service.

Fried, Marc, and Peggy Gleicher. 1961. Some Sources of Residential Satisfaction in an Urban Slum. *Journal of the American Institute of Planners* 27: 305–315.

Gans, Herbert. 1962. *The Urban Villagers: Group and Class in the Life of Italian-Americans*. New York: Free Press.

Jacobs, Jane. 1961. *The Death and Life of Great American Cities*. New York: Random House.

Krieger, Nancy. 2006. A Century of Census Tracts: Health & the Body Politic (1906–2006). *Journal of Urban Health* 83: 355–361.

Leonelli, Sabina. 2013. Global Data for Local Science: Assessing the Scale of Data Infrastructures in Biological and Biomedical Research. *BioSocieties* 8: 449–465.

———. 2015. What Counts as Scientific Data? A Relational Framework. *Philosophy of Science* 82: 810–821.

Lubove, Roy. 1962. *The Progressives and the Slums: Tenement House Reform in New York City, 1890–1917*. Pittsburgh: University of Pittsburgh Press.

Mood, Eric W. 1986. *APHA-CDC Recommended Minimum Housing Standards*. Washington, DC: American Public Health Association.
Mood, Eric W., Barnet Lieberman, and Oscar Sutermeister. 1969. *Housing Code Standards: Three Critical Studies*, Research Report no. 19. Washington, DC: National Commission on Urban Problems.
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.
O'Connor, Thomas H. 1993. *Building a New Boston: Politics and Urban Renewal, 1950–1970*. Boston: Northeastern University Press.
Ramsden, Edmund, and Matthew Smith. 2018. Remembering the West End: Social Science, Mental Health and the American Urban Environment, 1939–1968. *Urban History* 45: 128–149.
Shaw, Clifford R., Frederick M. Zorbaugh, Henry D. McKay, and Leonard S. Cottrell. 1929. *Delinquency Areas: A Study of the Geographic Distribution of School Truants, Juvenile Delinquents, and Adult Offenders in Chicago*. Chicago: University of Chicago Press.
Solow, Anatole A., and Allan A. Twichell. 1947. Housing Objectives in Terms of Health. *Journal of the American Institute of Planners* 13: 22–25.
Walker, David M. 1947. Urban Redevelopment – Making a Beginning on the Job... Philadelphia. *The Journal of Housing* 4: 69–70.
Williams, Huntington. 1942. Housing as a Health Officer's Opportunity. *American Journal of Public Health* 32: 1001–1004.
Wilner, Daniel M., Rosabelle P. Walkley, Thomas C. Pinkerton, and Matthew Tayback. 1962. *The Housing Environment and Family Life: A Longitudinal Study of the Effects of Housing on Morbidity and Mental Health*. Baltimore: Johns Hopkins Press.
Winslow, Charles-Edward A. 1945. Landmarks of 1944: Housing and Public Health. *American Journal of Public Health* 35: 18–21.
Wood, Edith Elmer. 1936. *Slums and Blighted Areas in the United States*. Washington, DC: GPO.
———. 1935. Housing: Public and/or Private. *Survey Graphic* 24: 5–7.
———. 1940. *Introduction to Housing: Facts and Principles*. Washington, DC: United States Housing Authority.

Edmund Ramsden is a Wellcome Trust University Award Lecturer in the history of science and medicine in the School of History, Queen Mary University of London. His current research is focused on the history of experimental animals in psychology and psychiatry and on the influence of the social and behavioural sciences on urban planning, architecture and design in the twentieth-century United States with a particular interest in public housing.

From Washington DC to Washington State: The Global Burden of Diseases Data Basis and the Political Economy of Global Health

Jean-Paul Gaudilliere and Camille Gasnier

Abstract This chapter takes the origins, development and uses of the Global Burden of Disease database as lens to interrogate the political economy of global health, focusing on the intended logic of this massive accumulation and manipulation of epidemiological data, and the ways in which it informs the management of public health programs and activities. Following the GBD's journey from its first embodiment as a World Bank tool in the early 1990s to its present day development at the Institute for Health Metrics and Evaluation helps understand how epidemiological data travel to become actionable data, revealing the complex interactions between data gathering on political purpose and their effective uses (or non-use) in specific contexts. The GBD database was first conceived following an accounting logic closely linked with planning: by aggregating epidemiological as well as financial data, the aim was to achieve triage, i.e. balance health budgets and prioritize investments. Nevertheless, as we argue, the specific context of global health and its mode of government have given way to different and contrasting uses of the database. GBD data are now most referred to as indicator: in global "donors" discourses they figure as numerical pictures of suffering distribution across the globe and signs of emergency.

1 Introduction

Everybody paying a short visit to the Institute for Health Metrics and Evaluation's website can experience the wealth of data on diseases and on their impact world wide it offers. Indeed, this research center in global health, financed by the Gates Foundation and located in Seattle at the University of Washington, has elaborated an impressive database on the "Global Burden of Disease" made available through the site's interface. This software and its inexhaustible stock of charts and maps display

J.-P. Gaudilliere (✉) · C. Gasnier
Cermes3, Inserm-EHESS, Paris, France
e-mail: Jean-Paul.GAUDILLIERE@cnrs.fr; camille.gasnier@ehess.fr

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_18

Disease Adjusted Life Years (DALYs) lost world wide because of illness, enabling the comparison of the burden of illness across time, countries and/or pathologies.

The logic underlying this unique tool is that the "burden of diseases" is measurable at a global level based on the aggregation of local and national data collected through a network of hundreds of institutions and ten times as many collaborators; and that the impact of mortality, disability and risks can be reduced to one single standard unit: the years of life lost as a consequence of illness (what they call the "DALYs"). The GBD data basis thus provides a general equivalent for assessing the "global" impact of health disorders – both in the geographical and the epistemic meanings of the world global. This burden is massive amounting to hundreds of millions of years of life lost and IHME's implicit statement in making it visible is that such burden hinders the growth of the economy and the progress of social life on a grand scale.

But there is more to the GBD, specifically what it highlights is the question of non-communicable disorders and comorbidity. A long assumed vision of health in the global South stresses the importance of infectious disorders and a scenario of the epidemiological transition mimicking the twentieth century Northern history of the replacement of infectious diseases by chronic disorders. In contrast, the GBD data (Fig. 1) reveal that diseases in low- and middle-income countries are increasingly double in nature with infectious as well as non-communicable disorders like depression, cardiovascular or pollution related pathologies affecting the population of these countries and their people individually.

Who are the intended viewers of such data? To some extent they are research physicians and public health specialists but as IHME leadership and its sponsors explain in every presentation of the GBD: this is a tool for action (Gates 2013). Its envisioned users are in the first place donors, public or private, who must decide where to put their money and how to make the biggest difference in the future of the world's health with their investments. Underlying the display of objective health data is therefore an ethos of intervention and fast response to emergencies: "we" (the donors) are able to know what counts in global health, we are able to know how to prioritize actions, we are able to know how to evaluate outcomes and measure efficiency or performance.

Critical analysts and actors of global health alike have commented on the emergence of the GBD and its relations to global health (Adams 2016; Arnessen and Nord 1999; Birn 2009). One frequent thread of analysis is to approach it as the highly visible symbol of the transition from "international public health" to "global health". This transition has been analyzed as shifting power alliances, the World Health Organization and its member governments finding their dominant role challenged by a series of organizations that emerged in the 1990s (nongovernmental organizations, transnational corporations, influential foundations such as the Gates Foundation) (Brown et al. 2006; Muraskin 2005) to target specific emergencies (malaria, tuberculosis, HIV, and – rarely – non communicable diseases). This new political order was also reflected in a shift in practices, with the diffusion of new tools (standard programs for access to drugs and other technologies) sometimes

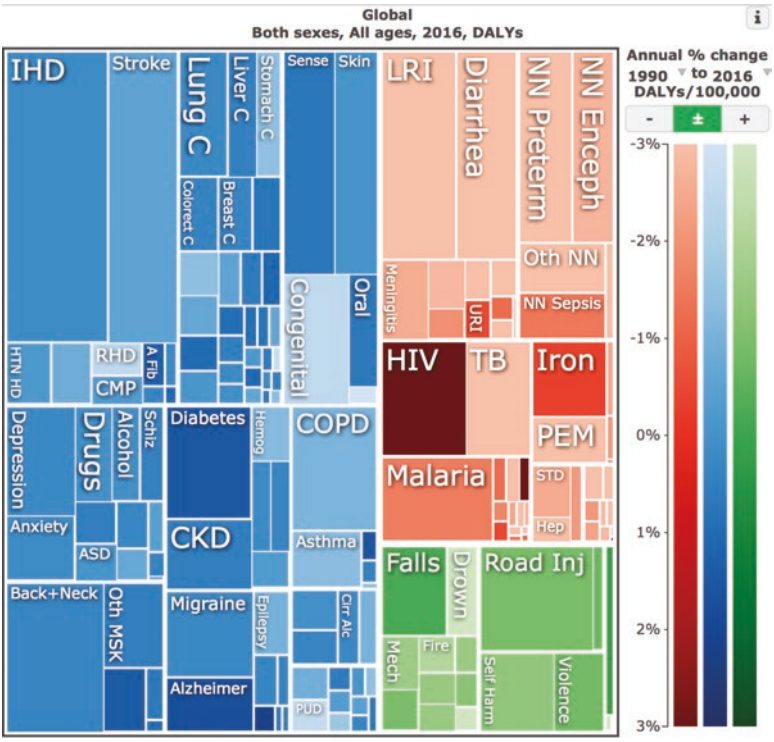


Fig. 1 Representing the burden of accidents, infectious and non-communicable diseases worldwide
Source: IHME website, GBD Data visualizations, GBD compare, http://www.healthdata.org/data-visualization/gbd-compare, accessed November 10th, 2018

derived from corporations' management (standard procedures, accounting systems, performance indicators) (Reubi 2018).

Private initiative, markets, management and individual choices are the keywords of this new world of health beyond nation states; a world, which has often been identified as one more manifestation of the big *neoliberal* transformation of government originating in the 1980s critical evaluation of Keynesian economic policies (Chorev 2007). Historians, anthropologists and sociologists have therefore regularly opposed global health and the postwar decades of international public health when development, nation-states, UN institutions and "health as a right" (as the WHO constitution proclaimed in 1946) dominated the landscape of health and population government, within and beyond the borders of nation-states (Brown et al. 2006; Chorev 2012; Randall 2016).

In this chapter, we use the GBD database as a lens to refine this contrast and interrogate the political economy of global health, focusing on the intended logic of this massive accumulation and manipulation of epidemiological data, and

particularly on the question of triage[1] as key issue and core practice in the management of public health programs and activities. The GBD's journey from the World Bank in the early 1990s to present day IHME helps understand how epidemiological data travel to become actionable data, revealing the complex interactions between data gathering on political purpose and their effective uses in specific contexts. Indeed, tracing the first GBD, the one, which surfaced in the World Development Report issued by the World Bank in 1993, the first programmatic document of this sort focusing on health, highlights the political goals grounding the invention of this new metrics whose main purpose was to allow the economic comparison and therefore the triage between different health interventions. In resonance with M. Morgan's analyses in the present book, we could say that the GBD database was first conceived following an accounting logic closely linked with planning: by aggregating epidemiological as well as financial data, the aim was to achieve triage, i.e. balance health budgets and prioritize investments. Nevertheless, as we argue, the specific context of global health and its mode of government have given way to different and contrasting uses of the database. GBD data as produced by the IHME are most referred to as indicators: in global "donors" discourses they figure as numerical pictures of suffering distribution across the globe and signs of emergency rather than tools for systematic comparison and prioritization. The journey from Washington to Seattle has therefore changed the nature of both the exercise and the data.

2 DALYs as Global Metrics: The World Bank and Economic Triage

The landmark in the World Bank sanitary turn was the publication of the 1993 World Development Report "Investing in health" (thereafter 1993 WDR), which made official and rationalized the Bank investments in health as decisive elements in its strategy to alleviate poverty. It thus departed from decades-long commitments to a vision of development centered on the building of infrastructures, on the rise of agriculture productivity and – when it came to deal with populations as such – on birth control (Devesh 1997; Ruger 2005; Staples 2006).

The change was not a sudden, crisis-like event, solely grounded in the new sanitary disorder of the time associated, for instance, with the dramatic impact of the AIDS epidemics. Rather, it had deep roots in 1980s internal debates on the meaning and targets of development, which remain to be properly mapped. For instance the Bank's Population, Health and Nutrition division priorities were deeply impacted by the 1970s and 1980s contestation of population control programs in the global

[1] Triage is a concept used so far by anthropologists to refer to "clinical triage", i. e. the local decisions regarding who will or not benefit from therapeutic and other interventions (see for instance Lachenal et al. 2014)

South and the acknowledgment that some demographic transition was happening worldwide without much correlation with these programs. One critical aspect of these debates was the mounting importance of "human capital" as category for analysis and action as reflected in a wave of reports issued in the 1980s and 1990s (for instance Becker 1995). Human capital theories thus backed a gradual displacement of issues toward health, education and women empowerment reflected in the growing number of projects the PHN division launched and their shifting focus away from nutrition and population control.

A second dimension in the Bank's sanitary commitment was the complex relationship the shift maintains to structural adjustment policies. Investing in health did not officially contradict the latter's conditions for granting loans to nation-states caught in the debt crisis, namely the urgency of budget balancing and privatization. True, the Washington consensus singled out health and education as priorities. However, in practice, public investments in the social sector were very often severely cut as an effect of structural adjustment policies. Moreover, all along the 1980s cost recovery in the health system was a persistent motto in World Bank's reports and memos of understanding with countries (De Ferranti 1985). This provided the background for the famous 1987 Bamako declaration through which African countries expressed their willingness to engage in the generalization of patients' fees for hospitals services and drugs with the background motive that these fees would ease the financial burden of health institutions, provide rolling funds to improve supply and make "pseudo-clients" more responsible and attentive to the quality of what was provided. This agenda deeply backlashed and critiques escalated beyond the usual circles including public health circles and international organizations like WHO and UNICEF.

In the early 1990s World Bank officials knew it even if they disagreed about the interpretation of such developments, i.e. whether they should be considered as intrinsic flaws of the policy or signs of a misguided implantation by governments marginally interested in human capital development. A World Bank paper issued in 1995 thus tried to put adjustment's impacts on health into perspective, stating that countries that had undergone adjustment policies were allowed to spend more on health when adjustment ended and when their economy recovered, their spending on health growing faster than in countries that had not followed adjustment policies.[2] The 1993 WDR was a *de facto* response as it offered an alternative by strongly endorsing the idea that markets cannot by themselves provide for health care, which is in most instances a public good. *Investing in health* thus meant in the first place strengthening public, meaning nation-states based, health systems.

Strong elements of continuity with structural adjustments nonetheless prevailed. In the Bank's eyes, public management of health was only thinkable if cost-effective, if performance was placed center-stage, if targets were carefully accounted and outcomes measured. The introduction of the Disability Adjusted Life Years (DALYs)

[2] World Bank Archives, Folder 392721, Memo Yazbeck A., Tan J-P, Tanzi V, "Public Spending on health in the 1980s: the impact of adjustment lending programs", Background Paper of the 1993 World Development Report, August 1995.

was therefore not only a way to take into account health problems neglected using the usual mortality/morbidity statistics but also and more importantly the introduction of a measure, which could help balance problems and solutions, could for instance help decide whether, in a world of limited resources, tuberculosis chemotherapy was worth doing and putatively more effective than HIV prevention. Even if the DALYs were eventually criticized internally for their medical rather than economical nature – they could not help decide if states should invest in genetically improved crops or in health centers – the dream of a general equivalent, money-like, was not far away.

Calculating the DALYs implied aggregating mortality and morbidity data under the umbrella of lost years of life and therefore mobilized two different calculations. The first one amounted, for each disease category, to weighting the distribution of death numbers associated with age groups against the life expectancy specific to each country on the basis of coefficients factoring in the decreasing economic usefulness of people according to their age. The main novelty regarded the addition of a certain number of years of life lost due to "disability" based on a fractional equivalence between a year of normal life and a year of impaired life with the disease in question. The coefficients applied for each disease to compute the impact of disabilities, i.e. the number of years of life lost due to bad health were in fact defined by using an average of the answers of a small number of experts to adapted questions in surveys designed to reveal preferences: "You are a decision maker who has enough money to buy only one of two mutually exclusive health interventions. If you purchase intervention A, you will extend the life of 1000 healthy (non-disabled) individuals for exactly one year, at which point they will all die…. The alternative use of your scarce resources is intervention B, with which you can extend the life of *n* individuals with a particular disabling condition for one year. If you do not buy intervention B, they will all die today; if you do purchase intervention B, they will die at the end of exactly one year." (Arnessen and Nord 1999, p. 1424). Experts had then to choose the value for *n* that would make them indifferent between the two programmes (Murray et al. 2002).

Following the publication of the 1993 report, the calculus of DALYs has been much discussed including the ways in which the GBD numbers incorporate a productivity-based understanding of the value of life or a quantified understanding of how valuable, how normal, is a year of life with tuberculosis, diabetes or cancer. Bringing the impact of disease down to a single indicator based on age, gender, the disability situation and the moment when the disease began was justified by the need to build a comparison tool allowing decision makers to choose their interventions by comparing the incomparable, by evaluating, for instance, the difference between the cost of one year of life for a child suffering of vitamin deficiency and cost of one year of survival for a 50-year-old with cancer. Put it differently, the DALYs were an attempt to seize all kinds of suffering in a commensurable way, in order to compare the effects of very different health interventions and choose the most efficient ones in budget constraints contexts, i.e. to try to optimise investments in health by choosing the interventions that would be the most effective in relieving the "burden" of suffering. In resonance with M. Morgan's analyses in the present

book, the invention of the GBD could be understood as accounting data, data linked to the government's need to monitor the economy in a constraint budget, to arbitrate between different social investments (others than health, also education, see M. Morgan's chapter) and to evaluate and optimize the returns.

From this point of view, the DALY works in a similar way to the QALY (Quality Adjusted Life Years) in health economics. In fact, the discussion of DALYs in the 1990s was similar to the numerous debates about QALYs, their advantages and limitations; for example, targeting the coefficients used to give the deaths of children or old persons less weight (Gavin 2002), or the arbitrary nature of the assessments regarding the value of one year of life with various disabilities or the value of impaired functioning due to a disease, as if human misery, "evaluations of severity and its cost [could] be validly standardized across different societies, social classes, age cohorts, genders, ethnicities and occupational groups" (A. and J. Kleinman 1996). More important for this paper is however the connection the 1993 report made between the DALYs and the measurement of cost-effectiveness. This was a central ingredient in the valorization of the GBD as basis for triage. This linkage has been overlooked since – for reasons discussed below – it disappeared from the exercise when the GBD machinery moved from the World Bank/Harvard/WHO complex to the Gates/IHME nexus.

The calculus of DALYs was actually combined with a general evaluation of performance in national health systems. For the poorest countries, the recommendation was to stop financing high-technology hospitals and expensive care infrastructure, only of benefit to the middle and upper classes, and to privilege interventions that would meet the needs of the most destitute, populations "at risk", less because of their peculiar exposure to pathogens than because of their social and economic vulnerability. Hence, the World Bank experts recommended reorganizing protection by defining a publicly offered and freely accessible basic system of care (the only one for which direct, centralized and evaluable action was possible).

One should not be mistaken, the point was not to leave out private actors, on the contrary, but to make a critical distinction between basic and more individual needs, between countries rich enough to cover costs, whatever the mechanism (taxation, insurance or patients' contribution) without drastic triage and low- and middle-income countries with very limited resources, where most households were not in a position to provide for their health needs, where triage was operating *de facto*, without much rationalization, favouring the urban middle class, and where the public provision of an "essential package" of interventions (through both public and private, first of all NGOs, services) was indispensable: "Perhaps the most fundamental problem facing governments is simply how to make choices about health care. Too often, government policy has concentrated on providing as much health care as possible to as many people as possible, with too little attention to other issues. If governments are to finance a package of public health measures and clinical services, there must be a way to choose which services belong in the package and which will be left out." (World Bank 1993 p. 59).

This plea for targeted investments was delineated in a much more detailed and prescriptive way with the selection of 47 interventions for which the Bank panels of

economists and health specialists computed costs and numbers of years of life saved in order to provide cost-effectiveness ratio. This complex operation actually started before the writing of the 1993 WDR, namely in 1988, with the establishment of a "Disease Control Priorities" (DCP) working group within the PHN division of the Bank whose initial aim was to develop new tools for measuring the effectiveness (rather than the monetary benefits) of health investments.[3]

The DCP project relied on another kind of triage to define the rationale for government's involvement in health, not only in prevention but also in curative services. Indeed, published as background material to the WDR report, the DCP working group final document relied on the attempts by panels gathering epidemiologists and economists working on one pathology, i.e. tuberculosis, or one medical issue, i.e. mental health, to assess legitimate interventions in their field, gather all available economic evidence on their costs and outcomes under optimal conditions and – when possible – provide numbers for the cost per DALY avoided. These numbers were then used to rank interventions according to their effectiveness. A major result was that – in contrast – to the classical divide health economists were making between prevention and treatment with the former considered as "public good" due to the importance of externalities and the impossibility to accrue individual benefits to a putative buyer, economic legitimacy crossed the line with highly cost-effective clinical intervention such as tuberculosis chemotherapy and poorly cost-effective preventive intervention such as water sanitation (Fig. 2).

The final outcome of this ranking effort was the proposition of an "essential package of health services" in developing countries. Beyond effectiveness expressed in terms of cost for one DALY avoided, overall spending was critical in the selection: World Bank experts estimated unrealistic to bet on a massive increase of public expenditures in low- and middle-income countries even if they spend much less than developed ones in proportion of their GNP. The essential package was thus limited to a doubling of what low-income countries were already spending to reach the level of $12 per person per year. The package prolonged and provided new legitimacy to existing priorities like immunization, STD treatment, prenatal care, family planning or Aids prevention. Decisive novelty resided in a few items in the category of non communicable diseases prevention and clinical treatment: tobacco control or the more important "limited care" cluster focusing on the treatment of skin allergies and injuries on the one hand, access to medications for pain relief, diabetes, hypertension, and tuberculosis on the other hand (Bobadilla et al. 1994).

Cost-effectiveness and performance – the values imbedded in the first GBD as well as the associated expectations for an economically rationalized triage to some extent confirm the neo-liberal scenario with one qualification, which is to recognize that privatisation of health services was only marginally the issue while triage was the fundamental one. As C. Murray, the man who had so strongly pushed for the creation of the GBD summarized: « Decision-makers who allocate resources to

[3] World Bank Archives, Folder 19831130, de Ferranti, « Sector financing an overview of the issues », Draft Population Health & Nutrition Paper, November 30th, 1983.

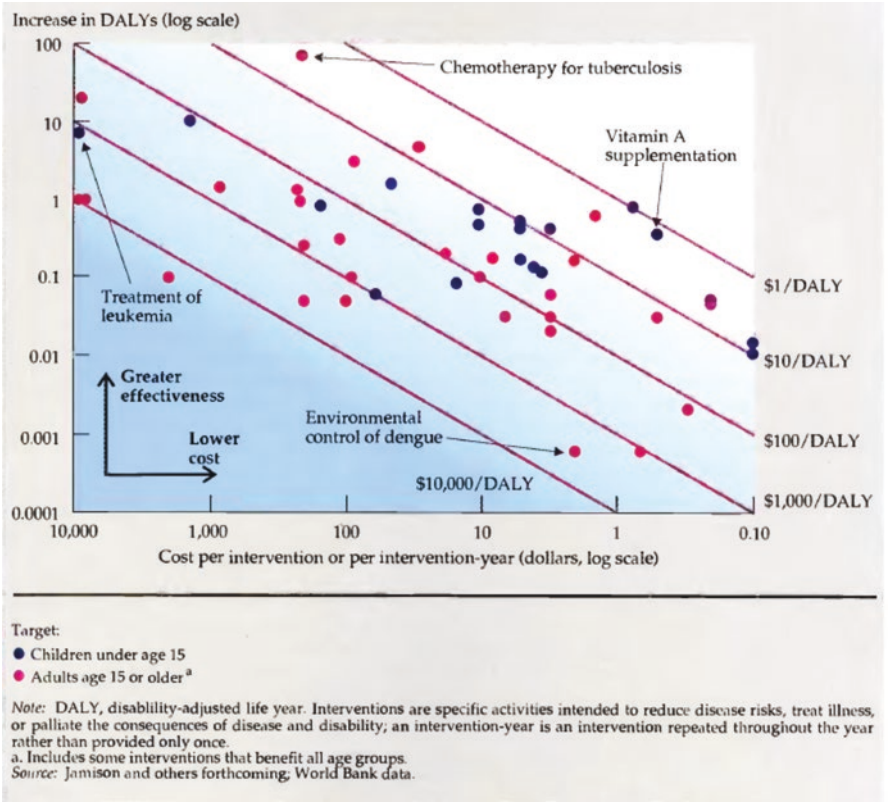


Fig. 2 Benefits and costs of forty-seven health interventions
Source: Courtesy of the World Bank (Jamison et al. 1993)

competing health programs must choose between the relative importance of different health outcomes such as mortality reduction or disability prevention. Because money is one-dimensional, the allocation of resources between programs defines a set of relative weights for different health outcomes. The only exception to this is in a completely free market for health care where such decisions between competing health programs are not made by a central authority but by individuals, one health problem at a time. » (Murray 1994).

The logic of economic triage and package design thus exemplify the accounting nature of the DALYs calculus. The comparison of interventions for their cost-effectiveness and their putative inclusion in a public package of services operated within the framework of an imaginary budget balancing exercise, namely a search for the "best" equilibrium between "inputs" (financial resources, most often combining state and donors) and "outputs" (the costs of selected interventions) that may include increasing inputs but most often meant adjusting outputs to preset levels of

inputs (see below the example of Bangladesh). Thus, rather than privatizing, states were now mandated to focus on investment performance. They were invited to enter what may be described as an audit culture (Power 1997) based on the use of a whole new range of evaluation and ranking instruments (Gaudillière 2014, 2016).

It was therefore not mere rhetoric when the authors of the report made an unexpected link with the old WHO primary health care strategy, explaining in their overview of the report that: "Provision of cost-effective health services to the poor is an affective and socially acceptable approach to *poverty reduction*. Most countries view access to basic health care as a human right. This perspective is embodied in the goal "Health for all by the year 2000" of the conference held by the WHO and UNICEF at Alma-Ata in 1978, which launched today's primary health care movement. Private markets will not give the poor adequate access to essential clinical services or the insurance often needed to access such services. Public finance of essential health services is thus justified to alleviate poverty. Such public funding can take several forms: subsidies to private providers and NGOs that serve the poor; vouchers that the poor can take to a provider of their choice; and free or below-cost delivery of public services to the poor." (1993 WDR, p. 5).

3 Health System Data and Political Triage: Primary Health Care at WHO

In its own way, i.e. the equation of primary health care with an essential package of intervention, the World Bank *homage* highlights one of the main elements of continuity between global health and international public health as typified in the 1978 Alma Ata strategy, namely the politics of scarcity and the permanence of triage. The Primary Health Care (PHC) strategy looked at health as human basic need whose fulfillment could not be thought of and worked out in isolation from other "sectors" of development. This had two major sequences.

The first one was that health was an object of planning. Late 1970s and early 1980s WHO texts on PHC thus endlessly repeat that PHC is a national strategy that presupposes public investments that need to be coordinated in a plan for the entire country and for "all the people" with a special view on those at the periphery, the most needy rural populations. The second element was that health had to be integrated in a general planning balancing investments in the health and other sectors with an eye on the multiple links between the various ingredients of progress. Deemed of special importance within this perspective were "boundary" questions of food and nutrition, water supply and sanitation, family planning and education policies.

PHC targeted "appropriate, affordable and acceptable technologies" and the establishment of "local" and "integrated" centers addressing "basic needs". The critical question was therefore how would such needs be selected and become ingre-

dients in the national health planning. The first response was that basic needs should be defined on the basis of epidemiological and public health knowledge mobilized by experts. The second and most forgotten response was to admit that the selection of "basic needs" was in the first place a political choice and that communities should – at least in the discourse – be granted a say.

Unsurprisingly political triage of the first sort dominated. Rooted in a long process of internal consultation involving all regional offices of the health organization – and in spite of a lack of transparency regarding the criteria and tools to achieve selection – the conference of Alma Ata ended with this list of targets: "education concerning prevailing health problems and the methods of identifying, preventing and controlling them; promotion of food supply and proper nutrition; an adequate supply of safe water, and basic sanitation; maternal and child health care, including family planning; immunization against the major infectious diseases; prevention and control of locally endemic diseases; appropriate treatment of common diseases and injuries; promotion of mental health; and a provision of essential drugs."

Political triage however did not imply an absence of data and numbers but these were epidemiological in the first place but not exclusively. Long before WHO and the World Bank started to collaborate (with all the tensions some participants have highlighted) in the making of the GBD, WHO and the Bank's PHN division had inaugurated exchanges of information, personnel and launched common ventures. Alma Ata was in this respect a turning point: from the World Bank perspective because it made its director and personnel consider that something new was happening at WHO that seized health as a system, linked to infrastructures and to other areas of development; on the later side because the Bank appeared not only as a resource for the funding of programs like sanitation projects but also as a source of expertise in the increasingly acrimonious debate on the feasibility of the PHC strategy, its broadness and the alleged need for a more selective approach. WHO thus sought WB help in defining a strategy for financing PHC, obtained WB collaboration for several new programs, including tropical disease research, maternal and reproductive health, extended immunization with the consequence that both institutions engaged in regular almost yearly strategic consultations. One specific dimension of this emerging common ground was the problem of health systems evaluation. Soon after Alma Ata, the WHO Director General started to look for Bank's know-how in the evaluation of development projects and system analysis having in view WHO ability to help nation states in the design of systemic reforms rather than vertical, operation or disease oriented projects. This resulted in the creation of a dedicated team, which started to collect data and produced in 1980 the first reference document on 'health systems' indicators and evaluation, which ideally included non-epidemiological data on budget, personnel, buildings, access to care and coverage.[4]

[4] WHO archives, draft memo « Indicators for monitoring progress toward 'Health for all' », 10 July 1980.

4 Missing and Alternative Numbers: The Low Visibility of DALY-Based Triage

Once the intimate relations between the GBD and economic triage as well as their difference with political triage and its associated data are acknowledged, the question becomes that of whether this unique global metrics is actuality used by donors or public health authorities. Such use has been postulated but rarely shown. It is true that DALYs figure in many discourses on global health emergencies or programs but this is in most instances as legitimizing argument, in isolation, with no connection to cost-effectiveness and without any significant comparison across interventions or diseases.

A good example is that of tuberculosis chemotherapy. The disease and the treatment figured prominently in the 1993 WDR as it appeared as one of the most cost-efficient intervention with a ratio a little above $1 per DALY avoided. The calculation originated in a nation-wide experimentation of a new regimen based on the short-course administration of standard antibiotics association conducted by the International Union Against Tuberculosis and Lung Disease (IUATLD) in Tanzania (and later in Malawi and Bostwana) during the 1980s (Gaudilliere et al. forthcoming). Employing initially a vintage regimen that combined streptomycin, isoniazid and thiacetazone, the project focused on operational improvements, care, and epidemiology to give existing regimens traction throughout the country. Examining what IUATLD introduced, we can imagine what was lacking before: systematic reporting of new cases and treatment was mandated; diagnosis through sputum microscopy – rather than X-ray – was made the compulsory standard; drugs were provided free of charge.

IUATLD experts kept an eager eye on efficiency of the program. Such efficiency was of course organizational, i.e. choosing the right protocol and building proper institutions. With the consequence that treatment failure would still be blamed on the non-compliant patient, while social conditions that drive epidemics and complicate therapy drop from the radar. The issue of strains resistant to antibiotics and therefore of drug sensitivity testing was given low priority when developing the NTLP (National Tuberculosis and Leprosy Program) along IUATLD lines. While diagnostic and treatment capacities with regards to bacteriologically positive cases were greatly expanded, diagnostic facilities in relation to drug sensitivity testing remained insufficient with one functional sputum culture laboratory throughout the whole period 1979–1988. Efficiency was not only a problem of epidemiology. It was also a matter of costs and choice of priorities; for instance aiming for a cure rate above 90% was considered useless at it would disproportionately increase costs and therefore should be avoided.[5] Cost efficiency was thus aimed at before the Tanzanian TB program got picked up in the World Bank calculus of DCP. Such interest resulted

[5] ERC GLOBHEALTH archives (Cermes3, Paris), Karel Styblo papers, Progress report «Results of the NTLP after the First Ten Years», June 1988.

in joint publications of IUATLD main expert Karel Styblo and of Christopher Murray, the architect of GBD, who in those days worked for Harvard University's Center for Population Studies (Murray et al. 1991a, b). Styblo and Murray combined public health epidemiology with economic analysis. Their evaluation of cost-effectiveness of short course therapy rested not just on curing more patients than standard therapy but also on the projected number of deaths averted or treatment costs avoided in future population.

Towards the end of the 1980s the Union project thus changed context and became the basis of a global strategy. The World Bank, from 1991, initiated a large-scale trial of the same regimen in China. The WHO, changing course after two decades of relative neglect for tuberculosis care, declared TB a global emergency in 1993. It used the IUATLD trials as examples and condensed the approach into the Directly Observed Treatment Short-course (DOTS) strategy that it put in practice from 1995 onward. As defined at the time, DOTS consisted in five elements deemed critical for success: the existence of a national program with significant political priority, the reliance on passive detection rather than active search for patients, bacteriological diagnosis, proper supply of drugs and delivery free of charge, and what had strongly come to the fore in the shift from trial to strategy: directly observed treatment meaning a form of ambulatory treatment such that patients would receive and absorb the drugs under the supervision of health personnel or of dedicated community workers. Advocating for DOTS in the mid-1990s, WHO made an abundant use of the 1993 World Bank cost-effectiveness calculus but dissociated what concerned tuberculosis chemotherapy from the entire discussion about an essential package of care and from the comparison with other interventions and their ranking. The $1 per DALY avoided was singled out and aligned with more clinical and epidemiological numbers like the regimen 90% efficiency or the projected numbers of deaths averted.

This did not imply that the logic of economic triage did not play a role in the 1990s global government of tuberculosis. In 1997 the World Bank reached an agreement with the Government of India, providing the latter $100 millions in order to reorganize its National Tuberculosis Program and implement the DOTS strategy. Launched in the 1960s the latter was the first program of its sort in developing countries and during the years WHO engaged in the PHC strategy, it played an exemplary role there. In the mid-1990s, when India began negotiations with the World Bank to get funding for a new programme, WHO and World Bank experts' evaluating this legacy were barely impressed: the Indian old program had certainly suffered from "serious lack of funds", but, according to the World Bank experts, it had been badly designed and organized with "insufficient trained staff (…), reliance on X-ray instead of sputum analysis for diagnosis (…), with a proliferation of drug regimens (…), a private sector which treat[ed] over 50% of new TB with an extraordinary variety of ineffective and potentially harmful drug regimens (…), a lack of quality control and regular supply of drugs (…), a reluctance of service providers to give adequate information to patients because of stigma (…), a poor recording and monitoring system (…) a lack of quality control of laboratory results." (World Bank 1997). In other

words, the programme was the opposite of the DOTS strategy, which was to be implemented on a massive scale, i.e., during the first 5 years, in 102 districts with a total population of 270 million persons.

Performance was not defined from the point of view of costs, which was done *beforehand* by calculating the total amount of the contract based on the average cost of chemotherapeutic treatment. The target criteria selected by the World Bank and the Indian Ministry of Health were medical and epidemiological: number of persons detected and treated (2 millions) and success rate of the cure (85%, where success is equal to the disappearance of TB-bacteria from examined sputum). All of this made it possible to anticipate a significant reduction in the incidence of tuberculosis.

Health data were however also proxys for another kind of performance, this time an administrative one, as is seen in the list of the specific risks of the programme identified by the World Bank experts, which were the risks of: difficulties in persuading providers and patients to accept the practice of directly observed treatment and the rigorous features of the DOTS strategy; poorly administered short-course chemotherapy and poor quality anti-TB drugs, which would increase the probabilities of developing drug resistance; the inability of the Central and State TB Cells to provide the leadership and services required to ensure proper implementation of the programme; an uneven supply of drugs combined with the availability of large quantities of drugs which could be misused, especially in light of "the spotty record of drug deliveries in India". (World Bank 2006).

Many development economists, including within the World Bank, however share the vision that GBD and the DALYs are not proper economic instruments since they favor an "internal" public health orientation that allow for comparison between diseases or interventions but do not provide for any rational for core economic questions like the level of investments, the allocations of resources between health and other social and economic sectors, the kind of care provision that can be left to the market; all questions central to the management of *national* health systems. As a consequence there is a low visibility of DALYs-based cost-effectiveness calculus even in the Bank's own assessment of health related investments and packages.

A good illustration of this is the late 1990s negotiation of health system reform in Bangladesh, during which the design of the basic package of essential services provided by the state took place without any mobilization of the kind of ranking involved in the 1993 WDR even if the process started with the $12 package proposed in the 1993 WDR report. The donors' mission and the local authorities had to take into account the fact that the Ministry of Health budget amounted to spending in the range of $3,5 per capita only. A technical group gathering practitioners, public health specialists and health authorities was gathered to select the interventions falling into four priorities areas: child health, reproductive health and population, communicable diseases, simple curative care whose costs would be computed with the help of an economist paid by WHO. Assuming that no significant increase of resources was politically feasible, triage to align the package with the $3,5 ceiling was the preparation team's next task. The technique used was not only cost-

effectiveness ranking but a scoring of each intervention involving five criteria for triage: costs, provision feasibility, potential health impact, burden of disease and economic status (whether the intervention could be considered as public good and the importance of its externalities). The dominance of general economic criteria in the process was reinforced in the last stage of triage since eliminating interventions from the package through scoring proved very difficult: it failed meeting the $3,5 target, leaving a deficit in the range of 15%. In order not to stall the negotiations with the donors and secure the help of World Bank, USAID and Northern Europe aid agencies, the preparation team finally agreed to draw a *balanced* "contingency plan", which the Bengali considered as a first step. As the World Bank expert participating in the negotiations later explained: "while (cost-effectiveness) is an economic evaluation tool, public health specialists, much more so than economists, swear by it as a primary prioritization tool." (Yazbek 2002).

One must add that the use of DALYs as instrument of economic triage has also been impaired by the difficulties associated with data collection for comparison on such grand scale and the recurrence of doubts or mistrust originating in its "missing" numbers and the complexity of the modeling involved in finding "proxys". This may be illustrated with one of the major outcomes of GBD, which has been to give an unprecedented visibility to mental health and psychic disorders in the global South. In a recent study of global mental health in West-Africa Anne Lovell has thus shown that for most of the region the numbers available in the GBD did not rely on local studies, did not mobilize the data originating in the operations of local health institutions, but originated in a complex set of correlations between the burden of mental health disorders and various epidemiological, social and economical variables worked out in countries benefiting from more reliable statistics (Lovell forthcoming).

The absence of policy-oriented use of the GBD of the kind that was expected at its origins doesn't mean that the GBD database is no longer conceived as data that could be used to enhance political decision-making relying on cost effectiveness analysis. Indeed, in spite of its infrequent use, the Disease Control Priorities project is still alive, at least as a modeling enterprise. In the mid-2000s, in parallel with its investment in IHME, the Gates Foundation started funding a follow up of the 1993 WDR, helping Dean Jamison and his colleagues produce a second DCP, and more recently a third DCP. The latter, based on complex computer simulation models and selecting 93 interventions and updating their ranking, has become part of mounting contemporary debates about "universal health coverage". Thus, economic triage based on DALYs has not disappeared, the irony being that it now finds renewal in debates regarding "universal health coverage", which reveals strong stands in favor of public investments compared to private markets but also deep tensions regarding universality understood as access for all but also universality understood as care for all health "needs", thus acknowledging although in an oblique manner the importance of political triage.

5 Conclusion

After its birth in between Washington DC and Harvard University, the GBD went through a period of difficult existence as its production was no longer supported by WHO or the World Bank (Smith 2015). It was finally rescued in 2007 when the Gates Foundation decided to fund the enterprise and have it relocated in Seattle at IHME, an independent institute at the University of Washington. The enterprise then took the form presented above, that of a global epidemiological data-basis whose connections with health economics is scant, namely embedded in the history of the DALYs on the one hand, acknowledged with the inclusion in the IHME data basis of financial information regarding global health investments on the other hand, thus leaving out of the scene the links between the two, i.e. cost-effectiveness, intervention ranking and economic triage.

There is no doubt that data and the GBD machinery travelled from Washington (DC) to Washington (State), but can this journey account for the changing epistemic status of the enterprise, for the fact that economics and triage have been put at arms' length? The response is certainly positive if the journey is considered not as a geographical displacement but as a shift from one social world to another. GBD moved from being strongly associated with a key financial institution whose main activity is the triage of development related loans to being inserted in one of the hotspots of global health academic life.

The move has been discussed as a consequence of personal tensions between its initiator, C. Murray, and other players in global epidemiology, or as an effect of WHO bureaucracy and entrenchment in outdated data production (Smith 2015). Given the origins and meaning of the first GBD, a much more critical question is that of the World Bank not following up and integrating the GBD in its operations; of the World Bank not producing any second WDR on health after 1993. As suggested in this paper, in so far as a non-event can be interpreted, this non-investment has deep roots in the tensions underlying the genesis of the tool.

In her chapter "The Datum in Context: Measuring Frameworks, Data Series and the Journeys of Individual Datums" of this volume, Morgan stresses the multiplicity of data sets economists have designed emphasizing the importance of their internal logic, i.e. the relationship between bits and whole. She accordingly distinguishes the accounting logic typical of highly integrated data sets like the matrix of national economies and the indicators logic of loosely articulated sets of numbers like those associated with the Millennium Development Goals. The difference resonates with the distinction we make between the uses of the GBD as instrument of economic triage, central to the design of packages, the comparison and optimization of investments on the basis of their cost-effectiveness on the one hand, and the uses of GBD data in an isolated manner, as measurement of the worth of isolated interventions or projects in order to legitimize choices made on the basis of other metrics and/or criteria be they epidemiological, organizational or social. The $12 package of the 1993 WDR report is emblematic of the accounting mode; WHO use of DALYs to

argue for the rationality of the DOTS strategy for controlling tuberculosis fits the indicator mode. Typical of the problematic life of the accounting mode is the fact that it is also in such a mode that the Gates Foundation uses the GBD. Analyzing the Bloomberg Initiative, a philanthropist association spearheaded by the Bloomberg and Gates Foundations and aiming at reducing tobacco use, David Reubi quotes an epidemiologist involved in the Initiative, explaining how Bill Gates relied on Murray's work on the GBD to design the Initiative. (Reubi 2018).

Morgan's perspective is also that these contrasted kinds of data strongly constrain their possible uses and ability to travel. Accordingly, the GBD data, originally elaborated to compare cost effective health interventions and choose the most efficient one, are nowadays more often used as isolated indicators of the geographical distribution of suffering worldwide and linked to different causes rather than as accounting data with strong relations to the whole, here economic growth. The trajectory of the GBD however reveals less direct relationship between kinds of data and political decision-making, more complex patterns for which the question of context in general, the political economy of global health in particular can't be avoided. This is quite obvious when considering the rise of triage based on economic data and performance assessment and the many ways in which this form of calculus and resource allocation contrasts the political triage of international public health and its logic of health needs. What this paper shows is that a simple reading of contemporary economic triage either as consequence of data sets design or as straightforward manifestation of the neo-liberal paradigm can't account for GBD uses and non-uses. Multiple political economies of health as well as heterogeneous institutional configurations were and are at stake resulting in differentiated modes of accounting as the difference between the vision of health financing underlying the 1993 WDR and the more neo-liberal one the "Population Health and Nutrition" division of the World Bank developed in the mid-1980s.

An important factor to be considered is thus the fact that the global health field in which the World Bank operates since the late 1990s is no longer that of nation-states "planning" development and making budget allocations. The global health world is a world of competing "causes" and vertical programs, which do not, or only marginally, target systems. It therefore does not require broad comparisons of interventions across the health sector, not even speaking of comparisons across the entire spectrum of development targets. Even the World Bank, that invented the GBD and that is supposed to be a development bank that invests in health systems strengthening, also massively targets vertical programs, mimicking its competing partners of "transnational humanitarianism" (Fassin, 2011). In fact, the global health movement at large doesn't seem to need a global metrics such as the GBD, its players being much more interested in indicators, that is to say data informing projects' symbolic and technical performance (David 2018).

Acknowledgements The research grounding this paper was conducted within the framework of the research project GLOBHEALTH (From International to Global: Knowledge, Diseases and the Postwar Governement of Health) funded by the European Research Council Grant 340510.

References

Adams, Vincanne. 2016. Metrics and the Global Sovereign. In *Metrics: What Count in Global Health*, ed. V. Adams. Durham: Duke University Press.

Arnessen, Theo, and E. Nord. 1999. The Value of DALY Life: Problems with Ethics and Validity of Disability Adjusted Life Years. *British Medical Journal* 319 (7222): 1423–1425.

Becker, Gary. 1995. *Human Capital and Poverty Alleviation,* World Bank Working Papers. Washington, DC: The World Bank:.

Birn, Ann-Emmanuelle. 2009. The Stages of International (Global) Health: History of Successes or Successes of History. *Global Public Health* 4 (1): 50–68.

Bobadilla Jose-Luis, Cowley Peter, Musgrove Phillip and Helen Saxenian. 1994. *The Essential Package of Health Services in Developing Countries*, World Bank Report 1993 Background Papers Series Number 1. Washington, DC: The World Bank.

Brown Theodor, Cueto Marcos, and Fee Elizabeth. 2006. The World Health Organization and the Transition from 'International' to 'Global' Public Health. *American Journal of Public Health* 96 (1): 62–72.

Chorev, Nitsan. 2007. *Remaking US Trade Policy.* From Protectionism to Globalization. Ithaca: Cornell University Press.

———. 2012. *The World Health Organization Between North and South*. Ithaca: Cornell University Press.

De Ferranti, David. 1985. *Paying for Health Services In Developing Countries. An Overview*, World Bank Staff Working Papers Number 721. Washington, DC: The World Bank.

Devesh, Kapur, John P. Lewis, and Richard C. Webb. 1997. *The World Bank. Its First Half a Century*. Washington, DC: The Brookings Institution.

Fassin, Didier. 2011. *Humanitarian Reason: A Moral History of the Present*. Berkeley: University of California Press.

Gates, William. 2013. *How We Measure Impact to Improve Lives*. Bill and Melinda Gates Foundation, January 29th. http://gatesfoundation.org/Who-We-Are/Resources-and-Media/Annual-Letters-List/Annual-Letter-2013.

Gaudillière, Jean-Paul. 2014. « De la santé publique internationale à la santé globale: l'OMS, la Banque Mondiale et le gouvernement des thérapies chimiques ». In *Le gouvernement des sciences à l'échelle globale,* D. Pestre (dir.), 65–96. Paris: La Découverte.

———. 2016. « Un nouvel ordre sanitaire international ? Performance, néolibéralisme et outils du gouvernement médico-économique ». *Ecologie & politique*, 2016/1 (52): 107–124.

Gaudillière, Jean-Paul, Christoph Gradmann, and Andrew McDowell. forthcoming. « The not so distant past, tuberculosis and the DOTS challenge ». In *From International to Global Health*, ed. C. Beaudevin et al. Manchester University Press.

Gavin, Yamey. 2002. Have the Latest Reforms Reversed WHO's Decline? *British Medical Journal* 325 (7372): 1107–1112.

Jamison, D.T., W.H. Mosley, A.R. Measham, and J.L. Bobadilla, eds. 1993. *Disase Control Priorities in Developing Countries*. New York: Oxford University Press.

Kleinman, Arthur. 1996. The appeal of Experience; The Dismay of Images: Cultural Appropriations of Suffering in Our Times. *Daedalus* 125 (1): 1–23.

Lachenal Guillaume, Lefève Céline, Nguyen Vinh Kim. 2014. *La médecine du tri. Histoire, éthique, anthropologie*. PUF, coll. Science histoire et société.

Lovell, Anne. forthcoming. *Global Mental Health Metrics in Sub- Saharan Africa: So 'Poor Numbers' Matter for Public Health?*.

Muraskin, William. 2005. *Crusade to Immunize the World's Children*. Los Angeles: USC Marshall – Global Biobusiness Initiative.

Murray, Christopher. 1994. Quantifying the Burden of Diseases: The Technical Basis for Disability-Adjusted-Life-Years. *The Bulletin of WHO* 72 (3): 429–445.

Murray, Christopher J., Styblo Karel, and Rouillon Annick. 1991a. Tuberculosis in Developing Countries: Burden, Intervention and Cost. *Bulletin International Union Tuberculosis* 65 (1): 6–24.

Murray, Chiristopher J., et al. 1991b. Cost Effectiveness of Chemotherapy for Pulmonary Tuberculosis in Three Sub-Saharan African Countries. *The Lancet* 338 (8778): 1305–1308.
Murray, Christopher J., et al. 2002. *Summary Measures of Population Health. Concepts, Ethics, Measurement and Applications*. Geneva: WHO.
Power, Michael. 1997. *The Audit Society. Rituals of verification*. Oxford: Oxford University Press.
Randall, Packard. 2016. *A History of Global Health. Interventions into the Lifes of Other Peoples*. Baltimore: Johns Hopkins University Press.
Reubi, David. 2018. Epidemiological Accountability: Philanthropists, Global Health and the Audit of Saving Lives. *Economy and Society*, *in press*.
Ruger, Jennifer P. 2005. The Changing Role of the World Bank in Global Health. *American Journal of Public Health* 95 (1): 60–70.
Staples, Amy S. 2006. *The Birth of Development*. Kent: Kent State University Press.
Smith, Jeremy. (2015). Epic measures. One doctor, Epic Measures. One Doctor, Seven Billion Patients, New York: Harper-Collins.
World Bank. 1993. *World Development Report 1993 Investing in Health*. New York: Oxford University Press.
———. 1997. *Staff Appraisal Report*. India. Proposed Tuberculosis Control Project, Rapport No. 15894-IN, 6 January.
———. 2006. *Implementation Completion Report on a Credit in the Amount of US$96.7 Million to India for a Tuberculosis Control Project*. Report No: 34692, 29 June.
Yazbeck, Abdo S. 2002. *An Idiot's Guide to Prioritization in the Health Sector,* HNP Discussion Paper. Washington, DC: The World Bank.

Jean-Paul Gaudillière is a Senior Researcher at the Institut National de la Santé et de la Recherche Médicale and Coordinator of the ERC project "From International to Global: Knowledge, Diseases and the Postwar Government of Health". His recent work focuses on the history of pharmaceutical innovation and the uses of drugs on the one hand and the dynamics of health globalization after World War II on the other hand. Amongst others, he has coedited, with Volker Hess, *Ways of Regulating Drugs in the 19th and 20th Centuries* (Basinkstokes, Routledge-Palgrave); with U. Thoms, *The Development of Scientific Marketing in the Twentieth Century: Research for Sales in the Pharmaceutical Industry* (New York, Pickering & Chatto, 2015); and, with L. Pordie, *Industrial Ayurveda: Drug Discovery, Reformulation and the Market* (Asian Medicine, vol. 9, 2014–2015).

Camille Gasnier is a Sociologist. She has been working on norms and the new forms of health and environmental management in the private sector. She is currently Postdoctoral Fellow at Cermes3 within the framework of the ERC project "From International to Global: Knowledge, Diseases and the Postwar Government of Health".

Data Journeys in Art? Warranting and Witnessing the 'Fake' and the 'Real' in Art Authentication

Catelijne Coopmans and Brian Rappert

Abstract This chapter approaches questions about data and data journeys by examining demonstrations of fakery and expertise in popular accounts by forgers and their pursuers. We examine how relations between tellers and audiences are configured – who can be trusted, and what can be relied on when it comes to knowing the real from the forged. The various ambivalences regarding the nature of art, of perception, and of expertise, as well as the ways in which moves and techniques (re)produce the expert-teller in fraught conditions, bring a shiftiness to the constitution of data and evidence in this domain. Taking our cue from STS scholarship on the fixation and circulation of visual evidence in scientific practice, we discuss moves and techniques that point *to* (particular) features of a work of art to resolve authenticity questions, as well as those that point *away from* and negate features. More though, at stake in the case of forgery is not just how individual objects get rendered discernible, but also whether there is anything to discern at all. This chapter examines how experts find a place to stand as they account for the potentially unfaithful objects under their care.

1 Introduction

The BBC's *Antiques Roadshow* is a programme that has, for decades, featured experts travelling from town to town to appraise artworks owned by the locals. It banks on the element of surprise: frequently, what owners believe or assume about their possessions is turned on its head in the process of expert appraisal. In the early

C. Coopmans (✉)
Institutionen för Tema, Linköping University, Linköping, Sweden
e-mail: catelijne.coopmans@liu.se

B. Rappert
Department of Sociology, Philosophy and Anthropology, Egenis, University of Exeter, Exeter, UK
e-mail: b.rappert@exeter.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_19

days, Impressionist and Modern Art expert Philip Hook remembers being able to deliver delightful surprises regularly, by revealing bygone objects as prized treasures. Later, as "members of the public […] got more and more optimistic about their property" (Hook 2014, 263), the reversal went the other way, the surprise less pleasant: "'Value? Not very much, I'm afraid. But it's such an interesting thing. Take it home and enjoy it.'" (Hook 2014, 263).

The surprises and reversals on the *Antiques Roadshow* depend on experts recognizing objects for what they are, positioning them relative to other objects ('rare', 'common', 'exquisite', 'decorative', and so on), and assigning them prospective market value. The objects presented do not always make such identification and valuation easy. Akin to the way scientists review and process instrument traces or specimens (Lynch 1985; Amann and Knorr Cetina 1990; Halfmann this volume), a careful adjudication comes into play, a reckoning with the possibility that things may not be as they seem. Materials are read for their trustworthiness, and spoken of – as Martin and Lynch (2009, 262–263) have argued in relation to cell biology – "as agents of their own visibility and identity: as showing and hiding themselves; presenting deceptive appearances; obediently complying with procedures or remaining recalcitrant."

One reason why art experts account for the possibility of deceptive appearances is the risk posed by *forgeries* – works intentionally designed to pass as those by a valued artist. Part of the job of appraisal is precisely to distinguish a genuine Constable from "a modern painting in the style of Constable which has been oven-baked in order to produce an apparent early-nineteenth-century craquelure in the paint surface and then claimed as a genuine Constable" (Hook 2014, 212). The matter of art forgery is not always black and white: the lines between a fake and an unintentional misattribution, a fake and a heavily restored item, or a high-quality fake and a low-quality original, are in specific instances blurred (Jones 1990; Hook 2014). Yet because the determination of origins and authorship matters so greatly to the price an item can fetch on the market, the epistemic game of authentication pulls towards a binary: *is it* or *isn't it*…

In this chapter, we complement this volume's analyses of data journeys in the sciences with an excursion into the efforts and complications of making features of artworks warrantable and witnessable in light of questions about authenticity. We thereby follow Steven Shapin's argument, in 'The sciences of subjectivity', that so-called subjective forms of knowledge-production merit study for how they are anchored and go beyond the idiosyncratic. Shapin observes, for instance, that much of the talk of wine connoisseurs, "is referential, that is, it points to characteristics in the wine that connoisseurs come to know about, and taste communities can and do coalesce around more or less stable way of designating these characteristics," (Shapin 2010, 178).

In art authentication, the stakes involved in distinguishing the 'fake' from the 'real' inform efforts to determine what's given and what can be relied upon in the face of possible ambiguity and deception. When concerns about authenticity emerge, claims about what is known and knowable, seen and seeable, for whom and when, cast artworks in the role of would-be 'data' to be mobilized as evidence for

determining true origins. In this chapter we highlight some key dynamics and variations of such casting, and discuss the accountability relations between experts, audiences and works of art that are thereby enacted.

The following tour of some of the colourful characters, controversies and efforts at revelation in the artworld is based on published 'insider accounts', documentaries and news reports – materials aimed at inviting a broad audience into an appreciation of the ways artworks may be designed to deceive and how such deception may be detected. Adopting an agnostic perspective on the possibility of turning artworks into data that travel into public accessibility, we find in the metaphor of data journeys an impetus for exploring what gets made available to make sense of contested works of art, how such warrants for sense-making are circulated beyond the expert realm, and what sorts of complications arise. Sabina Leonelli's (2016, chapter 3) relational definition of data proposes we think of data as what can be circulated and exhibited to others in corroboration of claims; she also points to data becoming salient *qua* data in and through the material form or "packaging" of information. In the first part of the chapter, we discuss practices and complications of "packaging" visual difference so as to make fakery available for 'all to see'. In the second part, we discuss practices and complications of treating artworks as material traces of their own origins. As our focus is on the public face of authentication, we will pay close attention to what becomes witnessable and portable, and for whom, amidst attempts to recognize objects for 'what they really are'. We will also show how such attempts, in turn, cannot be divorced from those of producing and preserving art's market value.

2 Here, See

To begin then, how are artworks, or aspects of these, mobilized in relation to claims to knowledge? How does art get worked up such that what is passed off as one thing can be exposed – in ways warrantable and witnessable – as *actually* something else?

Let us consider one prominent attempt to both spot fakes and cement the status of the teller: Thomas Hoving's (1996) *False Impressions: The Hunt for Big-Time Art Fakes*. In it he asserts that as much as 40% of the works he examined as the director of the New York Metropolitan Museum of Art were phony or tantamount to being so. Yet dealers, collectors, curators and artists were said to be reluctant publicly to speak about this phenomenon and reveal the art world's seedy underbelly – something Hoving took upon himself to remedy.

False Impressions forwards itself as an authoritative guide to the techniques and machinations of faking and 'fakebusting' in a number of ways. It 'proves' itself by: First, detailing years of experience and seniority of Hoving as a well-placed insider. Second, the extent of examples that serve as data points for the argument. Forgery, after forgery, after forgery (after forgery…) is presented to the reader: a sixth century BC Greek bronze statue, a fourteenth century enamel plaque, a Renaissance cup, a twentieth century sketch, etc., etc., etc. in page after page after page. Third,

the specificity of claims. Hoving names names: former colleagues' experiences and mishaps are detailed – friend, foe, and neutral person alike are identified. Such features shore up the book's said ability to spill the beans about "the real world of fakes." (Hoving 1996, 7).

Especially given the rampant and yet largely unremarked deception said to be afoot, much is at stake in how Hoving practically marshals visual materials in support of contentions about the status of specific objects. To understand how he does so, consider one example. In the main text of *False Impressions*, Hoving describes his immediate appreciation of two versions, one authentic and one a forger's copy, of a sketch of a boy called Henri Leroy by the nineteenth-century artist Jean-Baptiste-Camille Corot. Hoving claims he had no trouble spotting the fake:

> I, for one, instantly selected the drawing (figure 48) that combined an unmistakable heavy-handed and academician's touch […] with congeries of tiny mistakes, most of them in the rendering of the child's costume. [The] phony seemed all too obvious, far too plodding and deliberate for the nervous and carefree genius of Corot (Hoving 1996, 193).

These sketches, which Hoving found in a book by art forger Eric Hebborn, are reproduced in a black and white photographic insert in *False Impressions*. They are accompanied by a caption specifying which is which, followed by another pointer regarding the grounds for this determination: "The academic correctness gives the phony away." (Hoving 1996, insert, figures 15&16)

Both the main text and the caption provide a firm upshot of what is on display. Despite Hoving's making a personal statement ("I, for one…"), readers are solicited plainly to 'see' for themselves certain observable features that make the imitation different from the original ("unmistakable heavy-handed and academician's touch", "mistakes"). They are instructed where to see some of these features (the child's costume) and given to understand that these are tell-tale anomalous giveaways ("the phony seemed all too obvious").[1]

Just as laboratory science proceeds on the assumption of the "in-principle distinguishability of 'natural' from 'constructed' objects" (Lynch 1985, 82), here the fake Corot is charged with exhibiting features of artificiality ("academic correctness") in a way the genuine 'specimen' does not. In Hoving's account, the distinction is delivered in the way his remarks work up the two sketches in relation to one another. This achieves a 'fixation of evidence' (Amann and Knorr Cetina 1990), whereby visual expectations regarding what fits the "nervous and carefree genius of Corot" are mobilized as the background against which tell-tale signs of fakery can be located.

Here we find the familiar dynamic of multiplying the witnessing experience (Shapin and Schaffer 1985, 25). Despite being clearly in a position of being presented *to*, in being shown the materials that gave rise to Hoving's determination, readers can warrant the outcome of the comparison to themselves. By getting his audience to engage with the perceivable difference that bears out his assessment,

[1] We had wanted to show the sketches in this chapter, so readers could undertake their own visual comparison exercise, but did not get permission for reproducing the fake! To see them, try http://mountshang.blogspot.com/2009/11/which-is-fake.html

Hoving makes it at least notionally verifiable. The reader is invited to locate within the sketch designated as "the phony" the giveaway signs of its dubious status – and it is the comparison with the sketch designated as the Corot that makes for these signs' repeatable, widespread witnessability.

By evoking the relevant background for seeing the difference, the figurative gesturing Hoving does is more akin to an open palm that casts regard rather than an index figure that surgically singles out. This opening of palms is both a showing of *what's* at hand and a drawing of attention to *who* is providing this opening – that is, a gesturing *back* at the gesturer. The seeing enabled, then, warrants the fake at the same time as affirming that Hoving is in possession of the requisite knowledge and skills to know the significance of what is being displayed. It achieves a particular distribution of expertise, configuring the reader as one who does not know but can appreciate (cf Woolgar 1991). As an opening of palms, Hoving's consideration of the sketches is also notable for its non-exhaustiveness and generality. As a result, the gesturing sets up a situation in which readers' own efforts and abilities to see what is treated as plainly there for Hoving, are made accountable, as well.

3 See It, See It Not

Hoving's use of the sketches of Henri Leroy follows, as we have mentioned, an earlier such pairing by English forger Eric Hebborn in his autobiography *Drawn to Trouble* (1991). As part of his book, Hebborn too cites decades of professional experience, describes forgery, after forgery, after forgery (he crafted), and names names. And he too details a murky art world wherein the small-time scams of individual dealers and collectors complement the institutionalized dishonesty ingrained in the art trade.

As he claimed, anyway, Hebborn made it a point not to attribute his works. In his own dealings he simply offered the 'fortuitously found' works themselves (mainly old master drawings), letting those in the business of proffering attributions derive their own conclusions. Because of this practice, Hebborn argued that he did not delude; dealers, scholars, and collectors deluded *themselves* by seeing what they wanted to see. The frequency with which this happened he attributed to a number of factors, ignorance and greed among them, but also the mistaken belief that imitations or forged works can be straightforwardly identified for being of inferior aesthetic quality.

Let us consider then how such arguments inform Hebborn's presentation of the sketches of Henri Leroy. More directly than Hoving, and more playfully, too, Hebborn appeals to readers to get involved in appraisal by using their own eyes. The sketches are given side by side – the text reads:

> It might perhaps amuse you to test your own abilities as a connoisseur, and decide for yourself which of two photographs (Figs 48 and 49) represent a detail from the original. Even if you happen to be Joe Bloggs in person, you will still have a fifty-fifty chance of being right. Look carefully, take your time, and seek the hesitant line of the copyist as opposed to the strong sure line of Corot. The answer is given at the bottom of the page. (Hebborn 1991, 226)

Having teased readers regarding their ability to spot the difference, but also given them information to check if they picked the right original – in tiny letters at the bottom of p.226: "Answer: Fig 48" – Hebborn then goes further:

> Now, having read the solution, look at the two drawings again and you will suddenly notice how poor my version is, how faulty the construction, how harsh the modeling, and all sorts of ghastly errors which escaped your notice before.

The guessing game from before here develops into distinct valuations. The tongue-in-cheek disparaging of the 'fake' that "suddenly" appears "poor" also prods readers to recognize the dependency of their seeing on their understanding of what they are looking at. When a work is branded a 'fake' the eye seeks features that confirm its inferior status.

In yet another twist though, Hebborn carries on from the previous text to state:

> But what if I should now tell you that the answer at the bottom of the page is wrong?

With this, the features just established to anchor the distinction threaten to become mirage-like. Overall, the side-by-side juxtaposition combined with the textual instruction solicits a comparison between the sketches that 'packages' perceivable difference into tell-tale signs, but ultimately in a way that renders these unreliable as evidence.

Didactic comparisons of the kind employed by Hoving and Hebborn are a familiar device to educate non-experts about fakes; they have, for example, been a staple feature of museum exhibitions about fake art dating back to the 1950s (Lenain 2011, 264; Casement 2015). Philosopher Nelson Goodman (1983) distils the utility of this technique by arguing that, for the novice, the side-by-side juxtaposition of original and forged works:

> (1) stands as evidence that there may be a difference between them that I can learn to receive, (2) assigns the present looking a role as training toward such a perceptual discrimination, and (3) makes consequent demands that modify and differentiate my present experience in looking at the two pictures.

Through the deliberate placement of works side-by-side, Goodman claims novices become aware of the possibilities for learning-to-distinguish; by implication, proficient viewers can confirm their skill.

As we have seen in the previous section, Hoving sets the stage for comparison much in the manner that Goodman describes. The sketch identified as the real Corot is established as the measure by which the flaws of the other become perceivable. The reader is thereby aided in locating Hoving's assessment about 'academic correctness' in the way the fake differs from the original. It is aid that is more akin to someone giving directions by waving their hand along a bearing rather than pointing to a dot on a map, but aid nonetheless. In Goodman's terminology, the reader is cued to (1) see for themselves that the two sketches are different; (2) make sense of that difference in terms of the expert's assessment of what this difference amounts to; and (3) appraise one as an original, whose features then are seeable as character-

istic of Corot's style, and the other as a fake, whose features then are seeable as flawed imitations.

By contrast, if there is any training at all in the twists and turns of Hebborn's side-by-side game, it is to make readers aware of their susceptibility to priming. With his parody – the pointing to what "suddenly" becomes apparent when the answer has been revealed, and then the playful reversal – Hebborn drives home this point. When a work is made seeable as derivative of another, the perceivable difference between the two is cast in terms that makes the former seem inferior to the latter. Goodman's three features of the side-by-side technique are thereby questioned as enablers of learning. Readers are invited to recognize their own limits, and to become critical of the way a baseline for their seeing is provided. Hebborn doesn't contest that there are differences between the two sketches, but questions the way these are mobilized as tell-tale signs from which the matter of authenticity can straightforwardly be resolved. Difference may be witnessable, and may thereby be forwarded as 'data' for authentication, but its status as prospective evidence is shaky.

Hebborn's trickery seeks not only to confront his (lay) readers with their limitations, but also to undermine trust in expert determinations. A key refrain in his book is his assertion that experts had on several occasions mistaken his work for old master drawings, even as they would categorically deny that forged art can be of high artistic quality. His side-by-side game suggests that their confident claims and their efforts to anchor verdicts in specific features, may be built on quicksand. Yet, while on the one hand thus undermining expertise, Hebborn also relies on it to bolster his own status as a top-notch faker/artist:

> Just as there could be little satisfaction in scoring a goal in the absence of a goalkeeper, so it is that to sell a master drawing to someone lacking the necessary expertise to make a proper appraisal of it is at best a hollow victory. In other words, only the experts are worth fooling, and the greater the expert, the greater the satisfaction of deceiving him. (Hebborn 1991, 218)

As a result, for wanting to uphold the notion of "proper appraisal" as well as cast scathing doubt on the claims and warrants put forth by authentication experts, *Drawn to Trouble* is pitched in an arguably tension ridden tone. Amidst the exposure (again and again) of misattributions, self-deceptions, and bias rampant in the art world, Hebborn reaffirms the tradecraft expertise of the fooled. A similar tension is evident in the fact that the quality of Hebborn's work is contested among art historians and connoisseurs, with some considering them to approximate 'perfect' fakes (Lenain 2011, 269), and others arguing that his fakes are not nearly as convincing as he claimed (Jones 1990, entry 257; Hoving above). The questions of which differences matter, how to demarcate what belongs and doesn't to an artist's visual signature, and who can reliably do so, thus add complexity to the assumption that there is something 'wrong' with the fake that, once located, becomes available for all to see.

4 In the Blink of an Eye

In the previous sections, consideration of the sketches of Henri Leroy provided an illustration of techniques through which fakery is made accountable and anchored in what is at hand, along with how this shores up the credibility of experts. The juxtaposition of the two sketches provided the basis for turning perceivable difference into 'data', for acknowledging how attributions direct the seeing of fakery, and for indicating how things can get, well, befuddling. Although Hoving and Hebborn differed in how they orientated to the sketches and what they demonstrated, for both the side-by-side placement opened the possibility, using the words of Cohen and Cohen (2012), for a kind of 'hot' authentication based on direct perception by viewers rather than the 'cool' authentication gained by reading expert pronunciations. In this sense, the visual 'data' warranting authenticity verdicts is made available to anyone.

In this section, we move from practices and complications of "packaging" visual difference to discussing how different approaches to authentication treat artworks as material traces of their own origins. We do so mainly on the basis of the well-known tale of Teri Horton, an American truck driver who tried to get a painting she had bought in a thrift shop for $5 authenticated as a Jackson Pollock. This story, particularly as told in the documentary film 'Who the #$&% Is Jackson Pollock?' by Harry Moses (2006), also allows us to begin to explore how attempts to warrant authenticity intersect with commercial stakes.

Moses' and other accounts of the Horton case (Cole 2004, 2006; Hoving 2008; Grann 2010) feature two seemingly diametrically opposed forms of expertise that were brought to bear on the determination. The first is connoisseurship, focusing on the general stylistic impression of the painting and its resemblance (or lack thereof) to the general stylistic impression given off by Jackson Pollock's work. In the documentary 'Who the #$&% Is Jackson Pollock?' by Harry Moses, Thomas Hoving appears as the poster boy for this type of expertise, proffering a negative verdict for the painting:

> My instant impression, which I always write down, you know, the blink, the one-hundredth of a second impression, was: Neat. Dash. Compacted. Which is not good. He wasn't neat. He wasn't compacted. It's pretty. It's superficial and frivolous. And I don't believe it's a Jackson Pollock. (Moses 2006)

His nemesis in the film is art forensics expert Peter Paul Biro, who represents the second type of expertise. In a scene shot within his lab, Biro explains how he was able, with microscopes and high-powered photographic equipment, to locate a fingerprint on the back of the painting, which he then successfully matched to another print lifted off a blue paint can in Jackson Pollock's studio in East Hampton, New York. Another expert named Andre Turcotte demonstrates the match by pointing to the bifurcation pattern on both prints, presenting viewers with an animation that purports to demonstrate the overlay point by point. As Biro argued elsewhere in relation to his technique:

> Connoisseurship relies on an expert's close comparisons of a given work with closely related examples in order to discern where it "belongs" in terms of place, date and maker. By applying forensic methods, the process of attribution takes a novel and remarkable turn: it can use the evidence of fingerprints to trace a work of art back literally to the artist's hand. In effect, when the paper trail is missing or broken, forensics can at times fill in the gap. (Biro 2010, 157)

Biro's emphasis on material traceability represents an effort to make art into data that is markedly different from stylistic appraisal. In addition to his use of fingerprints, in the film Biro also explains how he matched gold particles found on Horton's painting with gold particles found on Pollock's studio floor.[2] In his efforts to construct material provenance trails, Biro's approach differs from the more customary way imaging technologies and materials science are drawn upon as a line of defense against forgery – namely by testing for 'deeper-layer' manipulation that does not show on the surface of the work, and for "glaring anachronisms of materials and technique" (Craddock 2009, 1).[3] Too, most notably in the work of Simon Cole (2004, 2006), fingerprint expertise has been argued to be more similar to, than different from, connoisseurship in its reliance on comparative judgment.

But Biro nevertheless is the poster boy for 'science' in 'Who the #$&% Is Jackson Pollock?', and the difference between his and Hoving's expertise appears large and unassailable. Connoisseurship appears highly inscrutable – certainly the caricatured way in which Hoving appears in the film makes him the epitome of an old-fashioned authority vested in the experience and trusted judgment of particular persons (privileged white males in particular), and offends modern sensibilities regarding the accountability of experts (Porter 1995). With Biro, on the other hand, the emphasis shifts from the expert to the evidence, in line with "the desire to democratize" art authentication by "scientificizing" it (Grann 2010). The making of *new data* from the work of art – fingerprints, paint sample readings – and the demonstration of technical methods designed to make extraction and comparison of such data systematic, replicable and verifiable (having set the terms for fallibility[4]), correspond to cultural notions of objectivity in investigating material links between the art object and its maker. To Hoving's put-down that "scientists are very interesting but come after the true connoisseurs" Biro retorts that connoisseurs need to update their understanding of what is and isn't a Pollock based on the evidence he uncovered (Moses 2006). 'Who the #$&% Is Jackson Pollock?' ends with the matter unresolved but also leaves viewers shaking their heads at 'ivory-tower' connoisseurship that refuses to reckon with material findings.

Hoving's approach gets a more positive billing in Malcolm Gladwell's *Blink: The Power of Thinking Without Thinking* (2005). Far from anti-systematic, the "instant

[2] And, an exercise of a different order, he matched the drip patterns of Horton's painting to those of an undisputed Pollock.

[3] For example, the fate of a purported Frans Hals sold by Sotheby's was sealed when scientific analysis in 2016 found "synthetic pigments that the artist, in the seventeenth century, could not have used" (Subramanian 2018).

[4] We thank Niccolo Tempini for providing this articulation.

impression" is here presented as a valid, if not easily explicable, way to know. The 'blink' points to a special kind of learnt receptiveness, developed through deep study of an artist's oeuvre. The knowledge that something is off can manifest in impressions such Hoving's above, or bodily reactions such as feeling cold ("as though there was a glass between me and the work", said one connoisseur), repulsed or uncomfortable (Gladwell 2005, 5) – something in line with a "sixth sense" (Hoving 1996, 19). The appeal to this acquired sense makes credible certain felt intuitions that might otherwise be dismissed as idiosyncratic reasonings, personal hunches, etc. The inability to isolate and nominate particular features as the grounds for a 'fake' verdict then does not invalidate such a verdict.[5] To paraphrase Gladwell, it is possible to know without knowing *why*, to know without being able to articulate *how*.

In the way these different approaches to authentication position art, knowledge, and appraisal to the public, as Cole (2006) has commented, "the truth ultimately comes down to which expert you believe." That 'you' are not yourself an expert is thereby underlined. At the same time, from the perspective of the artwork under investigation, the question of data remains highly relevant. Hans-Jörg Rheinberger's (2011) distinction between 'materials', 'traces' and 'data' helps to outline this. The artwork is construed alternately as a *material* to be interacted with, from which *traces* can be generated – this is what Biro was doing in scanning Horton's painting and Pollock's lab for fingerprints and paint samples – and *a complex trace in its own right*, in its totality a manifestation of its own history of becoming – this is what Hoving was engaging with. For Rheinberger, writing about experimental practices in the life sciences, the conversion of 'trace' to 'data' is about storing the information-content of precarious organic traces for future retrieval and pattern recognition (similar to Leonelli's definition of data as that which has been organized for witnessing, circulation, retrieval). In our case, treating art as *trace* is central and endemic to the work of authentication, but converting traces into *data* is *not*. 'Blink' thinking proceeds on a different basis than the ability to pinpoint data that can be appreciated or reactivated as key ingredients for authenticity adjudication; the subjectivity of the appraiser and the materials on which the verdict is made are here much harder to separate (cf. Shapin 2010).

At the end of 'Who the #$&% Is Jackson Pollock?, Teri Horton has not succeeded in having her work included in Pollock's oeuvre, but she has received an offer for it, for $9 m. As one of her friends remarks in the film: "Horton brought this painting to life": an "ugly" thrift shop painting nobody much cared about has been upgraded to a *possible* Pollock, disputed and in limbo. The expert appraisals have helped propel, if not exactly a data journey, a *journey of valuation* composed of a great many moves, including:

- A local art teacher pointing out to Teri that her painting 'could be' Pollock's work (prompting from her the question with the curse word that gave the documentary its name);

[5] Even as Hoving (1996), for one, recommends following it up with a detailed examination that might produce such features.

- Teri and her son Bill's persistent efforts to find experts and brokers in the art world who would not dismiss the painting out of hand;
- The technical investigation of Peter Paul Biro, which provided a turning point in giving the painting the 'weight' it needed to qualify as a possible Jackson Pollock;
- The favourable testimonies of Nick Carone, a friend and contemporary of Pollock (who said the painting is technically consistent with how Pollock painted) and art forger John Myatt (who said he could not have forged this work).

In and through these moves, a once-unremarkable object is shifted into a realm of ambiguity. Such a shift is also evident in relation to the work's undocumented provenance, highlighted by experts in the film as a big problem. Initial dismissal out of hand – "there are no Jackson Pollocks in thrift stores" – gives way to ambiguity as we learn that Pollock apparently *did* throw away work he wasn't happy with. Allan Stone, an art dealer, *did* get a genuine Pollock out of the dumpster in East Hampton. Lee Krasner, Pollock's widow in charge of the inventory, may not have kept proper track of all paintings that left the studio, etc.

A particularly interesting additional character introduced in the film is Tod Volpe, an art dealer contacted by Horton after she read his book *Framed: America's Art Dealer to the Stars Tells All*. Volpe gets involved to try and "put money behind the painting", specifically by interesting a collective of Hollywood actors and Wall Street finance professionals to buy it from Horton as a way to improve its provenance. Backed by such august owners, it can then be sold again, with a much-enhanced exchange value.[6] In an analysis of the case, Tay Yong Chiang (2016) called this effort to make the painting attract money a form of "commercial proof". The term is significant for how it puts Volpe's work on par with the authentication work of connoisseurs and scientists as three possible modes of proving that shape the object's journey of valuation. New datapoints are added: dollars and owners' names. The creation of commercial proof reminds of the way auction houses like Sotheby's and Christie's perform themselves as "temple[s] of civilised style and judgment" (Lacey 1998, 3).[7] Through glossy catalogues, private viewings, staff in smart evening outfits, the presence and commentary of experts, the style and demeanour of the auctioneer, and the way items are prepared for their moment in the spotlight, objects attract hefty sums.

So a once-ignored thrift shop painting is brought to life through a combination of moves that include extracting material traces from it (Biro), constructing the possibility of genuineness on the basis of testimony (Carone, Myatt), and offering it as an investment (Volpe) so it can begin to circulate in the artworld proper, *as* artwork proper. The way these moves together warrant the possibility of genuineness shows the entanglement between efforts to know Horton's painting for 'what it really is' and efforts to build commercial success for it. Significant, too, is the way the film

[6] Volpe did not, in the end, succeed in getting the painting sold this way, partly because Horton would not part with it.

[7] We thank CF Helgesson for suggesting this connection.

exhibits the mechanisms of which it tells. The idea for the documentary came from none other than Tod Volpe, who thought introducing the story to a broad audience would help build the painting's value, and who brought together the parties to make it happen. Clearly, the entanglement that shows up here between commercial stakes and authentication efforts – both put before the public's eye – complicates the common-sense notion that authentication precedes, and is a prerequisite for, sales. The casting of artworks in evidential roles intersects with their circulation in the market in a more complex fashion, as we will continue to unpack in the next section.

5 Is There Anything to See? Is There Anyone to See it?

Public demonstrations of the approaches by which works of art are made to speak to their own genuineness have their shadow side in the various ways that routes to knowing are presented as barred. The problem of revealing fakes is not only one of sizing up troublesome objects that trick our perceptions. It is also one of contending with a destabilizing doubt about the universe of objects deemed 'art' as well as the experts that speak for them. The knowing and knowability of art forgery entails a complex mix in which the truth is variously treated as available and elusive, publicly demonstrable and beyond simple verification, given up and held back.

Let's return to the writings of Hoving and Hebborn for examples of such oscillations.

Hoving moves from general arguments that fakes are easy to foil because there is always *something* that gives the fake away, to other general arguments that it may be difficult to know what to look for:

> [One] work of art can be proven a fake because the drapery is too nervous in style; another because the drapery is not nervous enough. A statue of the fourteenth century can be fake because it is too refined, too beautiful. Another statue of the same period can be condemned because it is not sweet and pretty enough. (Hoving 1996, 22)

As well as appealing to 'blink' judgment as a basis for distinctions, Hoving also, at times anyway, appeals to the even less specifiable spiritual quality of art. As he writes, "I tend to look upon works of art as partly spiritual and mysterious and partly human and fragile. Their lofty nature helps me break free from the mundane" (Hoving 1996, 16). In contrast to this glimpse of the sublime, phony pieces are "nothing but mockeries, dead things" (Ibid., 333). In attributing such a "cult status" (Benjamin 1936) to art, the distinction between the fake and the real remains clear, but rather than located in visible material features it gets bound up in some way with that which transcends (and thereby also with the learned receptiveness and sensitivity of the connoisseur). Hebborn performs a similar slipperiness in denying the visibility of forgery in terms of lack, fault, or inferiority, while upholding a sense that there is 'good' and 'bad' art. The former is the domain in which he locates his own efforts, which he looks to genuine experts to confirm on the basis of "proper appraisal".

More broadly, the nature of *False Impressions* and *Drawn to Trouble* as exposés of the art market does not allow us to rest entirely assured that there is a way out

from the havoc wreaked by ambiguities and close resemblances. The hidden nature of art forgery, the way in which it aims for close resemblance and for passing unnoticed, means that, as Sergio Sismondo has written for the practice of ghost management in medical publications, "we cannot tell how common it is from published exposés," (Sismondo 2007, 1429). Equally, discussions of the possibility of very convincing fakes that are hard to catch can be read as shoring up, but equally as destabilizing, the status of the teller by being self-serving without providing a stable reference point. Hoving attempts to uphold the notion that such reference points exist, while Hebborn embraces the slipperiness, acting as both exposer and trickster. These exposés then oscillate between upholding and destabilizing, for their audiences, the prospect of seeing and knowing, begging the question what there is to be witnessed in this space.

Routes to knowing are also presented as obstructed in reports about the legal pressures that keep art experts from making public what they privately know. In its basic contours, the dilemma is not new (Easby and Colin 1968) – negative authenticity assessments destroy market value, making owners and dealers lose money – but the large fortunes at stake in the art market now have made them more extreme. According to art lawyer Ronald Spencer, scholars are "nervous about taking a $500 fee and getting sued for $10m" (quoted in the Economist 2012). There are reports of experts refusing to make their doubts about new discoveries public, and of a scheduled debate about the authenticity of a set of Francis Bacon drawings being "cancelled a week before it was to have taken place […] due to 'the possibility of legal action'" (Economist 2012). The Andy Warhol Foundation dissolved its authentication board in 2012, citing the exorbitant costs of defending against legal challenges as the reason (Kinsella 2012). Authors of the *catalogue raisonné*, the authoritative list of works by a particular artist, report having received bribes and death threats (Cohen 2012; Economist 2012).

At the same time, getting proof of art forgery to hold up in court is difficult, as was shown in a recent case against two art dealers allegedly working for a crime syndicate that was flooding fake Russian modernist art into Germany:

> After five years of investigating the 1800-work collection in collaboration with more than 10 international experts […] authorities were ultimately unable to determine the authenticity of the bulk of the collection, after only four paintings were declared to be fakes. (Neuendorf 2018)

In this case, warring art experts making opposite claims did not help, and the court seems to have put most stake in the scientific analysis of paint samples, to which, for reasons the account does not provide, only a fraction of the works in the collection were subjected.

The rules of what to do with art that is assessed as fake are also not clear cut: the works may be confiscated or – in rare cases – destroyed, they may be stamped or marked in some way, but also may simply be returned to a dealer or previous owner in exchange for restitution. So it is not unheard of to have artworks previously discredited as fakes *resurface* in the market after some time (Cohen 2012). Efforts to recognize a work for what it really is are hampered when it is difficult to mobilize evidence of fakery in a court of law, and comparatively easy to cut loose the ballast of unfavourable verdicts and re-enter an artwork into circulation.

Such pressures and troubles encountered by art experts in turn affect the market: in the absence of a *catalogue raisonné* or the possibility of expert certification, "savvy art-buyers" are reportedly "spending less than they otherwise would" (according to a source quoted in the Economist 2012). They also change the accountability relations between experts, audiences and works of art. Some institutions have decided to do away with the *catalogue raisonné* in favour of something less definitive. For certain artists' works, online repositories are being developed that allow collectors and others to make their own determinations. In the words of a representative of the foundation that maintains the estate of artist Alexander Calder: "You determine if your work is fake or not with the data we present" (Cohen 2012).

This last statement brings full circle this chapter's survey of efforts to make fake vs. real art witnessable and warrantable. It leaves the viewer to assemble the case, as experts and the organizations that employ them put materials on display but stay clear from making evidential arguments. Such an outcome reminds of dynamics of devolved judgment and tension-ridden witnessing when expert testimony is calibrated to lay juror assessment, prompted by the question of whether evidence can and should 'speak for itself' (Goodwin 1994; Jasanoff 1998).

Overall, the knowing and knowability of art forgery is subject to shifting orientations. Firm grounds for determination are both gestured towards and withdrawn, and those giving accounts of the world of fakes can, among other things, displace offering a definite depiction; display fact after fact to build a credible argument; defer to some individuals as authoritative experts; devolve meaning making to viewers; and indicate obstacles to being able to see what is shown. As part of these strategies, expertise is varyingly circulated around, shifted in a zero-sum fashion, mutated, or pushed on to elsewhere. As a result, readers or viewers are varyingly barred, invited, and demanded to partake in the process of sense making.

6 Conclusion: Varieties of Data and Journeys of Art

In *Data-Centric Biology*, Sabina Leonelli refers to Paul Edwards' (2010) discussion of "data wars" in climate science to distinguish two ways in which data are handled and valued in scientific research. In one model, associated with weather forecasting – where "original sensor data may or may not be stored; usually they are never used again" – "the idea of "raw data" is not highly valued and scientists tend to work with models of data built through statistical tools" (Leonelli 2016, 22). In the other model, associated with climate science, the collection and curation of diverse data supports work on "a variety of research questions", and the point is for the data to be 'there' and available to be accessed at different points in time.

What we have described may be a third variety of how data and data journeys feature in the production of knowledge, one that applies to instances where the stakes revolve around recognizing things for what they are, and assigning them their proper 'place'. Art authentication in this respect finds common ground with archaeology, which as Alison Wylie (this volume) asserts, "depends fundamentally on discerning the temporal structure of the material record of the cultural past." In both

cases, inferences are made about the past from traces that endure into the present. Art authentication also finds common ground with forensics and medical diagnostics, which use the "clues" provided by what we can access *here* and *now* to make inferences about a *there* and *then* (Ginzburg 1989). The speculative point here is that data gain significance (or don't) insofar as they can be mobilized in relation to questions of origin, cause, perpetrator, instigating circumstances, etc.

Wylie shows such mobilization to be a "hard-won achievement", dependent on background assumptions, procedures of triangulation and the specific arguments archaeologists seek to make. In art authentication, too, efforts to determine and demonstrate what's given are informed by assumptions about art and authorship, about the craftiness of forgers (and the limits thereof), and, perhaps most fundamentally, about the artwork as a trace that, given the right approach, will speak to where it came from. The shiftiness we have documented in this chapter, especially in attempts to create public witnessability, resides in the varying ways these assumptions are embraced or contested. It also resides in how art is produced as covetable commodity and as investment, with works by popular artists fetching increasingly large sums as they change ownership. On the one hand, the determination and demarcation of what is and is not original has become more important as prices have risen; on the other hand the mere *possibility* of genuineness can spur financial speculation, and the pressure to avoid lawsuits has generated interesting readjustments of accountability relations between experts and art buyers ("You determine if your work is fake or not with the data we present.") As the epistemic challenge of authentication meets this commercial push-and-pull, what is extracted or pulled from the artwork, brought into view through comparison, gleaned through "blink" thinking, or added as information associated with the work, is attended to in various ways as secure or provisional.

More mobile than these data of different kinds, the examples in this chapter seem to suggest, is the artwork itself. What gets mobilized as evidence for inauthenticity at one point may not take the work out of circulation forever, and may be (temporarily) forgotten, marginalized, or erased. And yet, with the circulation of a work also circulates the possibility of revisiting it as evidence of its own origins. When, how and by whom that possibility is activated is circumstantial, but the importance of assigning artworks their proper "place" (in the double sense of *origin* as well as *resale value*) positions works in what we might call a permanent state of being proto-data.

7 Coda

It is worth noting that Hoving (at times anyway), Biro, and Hebborn all agree on the general availability of works of art to be open for inspection, despite differences in *how* they are made to speak and what they speak *about.* Whatever their varying moves, the underlying similarity those surveyed in this chapter share is the potential for discernment – if only we care to look properly.

And yet, despite the manner objects are positioned as sites available for "close looking, the making of fine distinctions" (Nagel 2004, section 13), some accounts of successful art forgery point in an opposite direction. British forger John Myatt, who was involved in the "the biggest art fraud of the 20th century", was known to have made poor-quality fakes for which he used "an easily detectable household emulsion paint developed in the mid-60s, decades after most of the paintings were supposed to have been executed. In some cases, he used K-Y Jelly as a medium to add body and fluidity to his brushstrokes" (Landesman 1999). American Mark Landis, a forger who successfully donated his works to prestigious museums, said of his method:

> I know everybody's heard about forgers that do all these complicated things with chemicals and what-have-you […] I don't have that kind of patience. I buy my supplies at Walmart or Woolworth – discount stores – and then I do it in an hour or two at most. If I can't get something done by the time a movie's over on TV, I'll give up on it. (quoted in Caffrey 2015)

Far from doing their utmost to confound scrutinizing gazes or scientific probes, Landis and Myatt present forgery as superficial. The commonplace notion of the gaze of the discerning viewer, expert, etc. that needs to fooled is thereby rendered into a mere trope. Just as deceptive objects are not what they seem, neither might be the practices of deception.

Acknowledgements Thanks to the editors of this volume, and to members of the Values group at Tema-T in Linköping, for helpful comments on an early version of this chapter.

References

Amann, Klaus, and Karin Knorr-Cetina. 1990. The Fixation of (Visual) Evidence. In *Representation in Scientific Practice*, ed. M. Lynch and S. Woolgar, 85–121. MIT Press: Cambridge, MA.

Benjamin, Walter. 1936. *The Work of Art in the Age of Mechanical Reproduction*. https://www.marxists.org/reference/subject/philosophy/works/ge/benjamin.htm. Accessed 18 Sept 2019.

Biro, Peter Paul. 2010. Fingerprint Examination. In *Leonardo da Vinci, "La Bella Principessa": The Profile Portrait of a Milanese Woman*, ed. M. Kemp and P. Cotte, 156–173. London: Hodder & Stoughton.

Caffrey, Jason. 2015. America's Most Generous Con Artist. *BBC World Service*, 31 March. http://www.bbc.com/news/magazine-31818367. Accessed 18 Sept 2019.

Casement, William. 2015. Fakes on Display: Special Exhibitions of Counterfeit Art. *Curator: The Museum Journal* 56 (3): 335–350.

Cohen, Patricia. 2012. In Art, Freedom of Expression Doesn't Extend to 'Is It Real?'. *The New York Times*, 19 June http://www.nytimes.com/2012/06/20/arts/design/art-scholars-fear-lawsuits-in-declaring-works-real-or-fake.html?_r=0. Accessed 18 Sept 2019.

Cohen, Erik, and Scott A. Cohen. 2012. Authentication: Hot and Cold. *Annals of Tourism Research* 39 (3): 1294–1314.

Cole, Simon. 2004. Jackson Pollock, Judge Pollak, and the Dilemma of Fingerprint Expertise. In *Expertise in Regular and Law*, ed. G. Edmond, 98–120. Aldershot: Ashgate.

Cole, Simon A. 2006. ART; A Little Art, A Little Science, A Little 'CSI'. *The New York Times*, 31 December. https://www.nytimes.com/2006/12/31/arts/design/31cole.html. Accessed 18 Sept 2019.

Craddock, Paul. 2009. *Scientific Investigation of Copies, Fakes and Forgeries*. Amsterdam: Elsevier.

Easby, Dudley T., and Ralph F. Colin. 1968. The Legal Aspects of Forgery and the Protection of the Expert. *The Metropolitan Museum of Art Bulletin* 26 (6): 257–261.

Economist. 2012. Collectors, Artists and Lawyers: Fear of Litigation is Hobbling the Art Market. 24 November. https://www.economist.com/news/business/21567074-fear-litigation-hobbling-art-market-collectors-artists-and-lawyers. Accessed 18 Sept 2019.

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

Ginzburg, Carlo. 1989. Clues: Roots of an Evidential Paradigm. In *Clues, Myths, and the Historical Method*, 96–125. Trans. John and Anne C. Tedeschi. Baltimore: The Johns Hopkins University Press.

Gladwell, Malcolm. 2005. *Blink: The Power of Thinking Without Thinking*. New York/Boston: Little, Brown and Company.

Goodman, Nelson. 1983. Art and Authenticity. In *The Forger's Art*, ed. D. Dutton, 93–114. London: University of California Press.

Goodwin, Charles. 1994. Professional Vision. *American Anthropologist* 96: 606–633.

Grann, David. 2010. The Mark of a Masterpiece. *The New Yorker*, 12 July. https://www.newyorker.com/magazine/2010/07/12/the-mark-of-a-masterpiece. Accessed 18 Sept 2019.

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.

Hebborn, Eric. 1991. *Drawn to Trouble*. New York: Random House.

Hook, Philip. 2014. *Breakfast at Sotheby's: An A-Z of the Art World*. London: Penguin Books.

Hoving, Thomas. 1996. *False Impressions: The Hunt for Big-Time Art Fakes*. New York: Simon & Schuster.

———. 2008. The Fate of the $5 Pollock. *Artnet* 4, November 4. http://www.artnet.com/magazineus/features/hoving/hoving11-6-08.asp. Accessed 18 Sept 2019.

Jasanoff, Sheila. 1998. The Eye of Everyman: Witnessing DNA in the Simpson Trial. *Social Studies of Science* 28: 713–740.

Jones, Mark. 1990. What Is a Fake? In *Fake? The Art of Deception*, ed. M. Jones with P. Craddock and N. Barker, 28–57. Berkeley: University of California Press.

Kinsella, Eileen. 2012. A Matter of Opinion: Concerns About Liability Have Led Several Artist's Foundations to Stop Authenticating Their Work. *ARTnews,* 28 February. www.artnews.com/2012/02/28/a-matter-of-opinion/.

Lacey, Robert. 1998. *Sotheby's – Bidding for Class*. London: Warner Books.

Landesman, Peter. 1999. A 20th-Century Master Scam. *The New York Times Magazine,* 18 July. http://www.nytimes.com/library/magazine/archive/19990718mag-art-forger.html. Accessed 18 Sept 2019.

Lenain, Thierry. 2011. *Art Forgery: The History of a Modern Obsession*. London: Reaktion Books.

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

Lynch, Michael. 1985. *Art and Artifact in Laboratory Science: A Study of Shop Work and Shop Talk in a Research Laboratory*. London: Routledge & Kegan Paul.

Martin, Aryn, and Michael Lynch. 2009. Counting Things and People: The Practices and Politics of Counting. *Social Problems* 56 (2): 243–266.

Moses, Harry. 2006. *Who the #$&% Is Jackson Pollock?* Documentary film.

Nagel, Alexander. 2004. The Copy and Its Evil Twin: Thirteen Notes on Forgery. *Cabinet*, 14 (summer). http://cabinetmagazine.org/issues/14/nagel.php. Accessed 18 Sept 2019.

Neuendorf, Henri. 2018. Case Against the Alleged Mastermind of a Russian Avant-Garde Forgery Ring Ends with Convictions on a Lesser Charge. *Artnet*, 15 March. https://news.artnet.com/art-world/alleged-forgery-ringleader-conviction-1245590. Accessed 18 Sept 2019.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Rheinberger, Hans-Jörg. 2011. Infra-Experimentality: From Traces to Data, from Data to Patterning Facts. *History of Science* 49 (3): 337–348.
Shapin, Steven. 2010. The Sciences of Subjectivity. *Social Studies of Science* 42 (2): 170–184.
Shapin, Steven, and Simon Schaffer. 1985. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton: Princeton University Press.
Sismondo, Sergio. 2007. Ghost Management: How Much of the Medical Literature is Shaped Behind the Scenes by the Pharmaceutical Industry? *PLoS Medicine* 4 (9): e286. https://doi.org/10.1371/journal.pmed.0040286.
Subramanian, Samanth. 2018. How To Spot a Perfect Fake: The World's Top Art Forgery Detective. *The Guardian*, 15 June. https://www.theguardian.com/news/2018/jun/15/how-to-spot-a-perfect-fake-the-worlds-top-art-forgery-detective. Accessed 18 Sept 2019.
Tay, Yong Chiang. 2016. *A Proof/Truth Analysis of 'Who the #$&% is Jackson Pollock?'* Lecture at Tembusu College, National University of Singapore, 17 October.
Woolgar, Steve. 1991. Configuring the User: The Case of Usability Trials. In *A Sociology of Monsters: Essays on Power, Technology and Domination*, ed. J. Law, 28–57. London: Routledge.
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.

Catelijne Coopmans is a Research Fellow in the Department of Thematic Studies, Technology and Social Change (TEMAT), Linköping University, Sweden. Her research interests include the dynamics of seeing/knowing and of expertise in areas such as medical diagnostics, business data visualization and art authentication. Among her publications are "Visual Analytics as Artful Revelation" (2014), "Eyeballing Expertise" (with Graham Button, 2014) and "On Conveying and Not Conveying Expertise" (with Brian Rappert, 2015). She is a Collaborating Editor at *Social Studies of Science* and a Member of the editorial board of the *East Asian Science, Technology and Society: An International Journal* (*EASTS*).

Brian Rappert is a Professor of Science, Technology and Public Affairs at the University of Exeter. His long-term interest has been the examination of the strategic management of information, particularly in the relation to armed conflict. His books include *Controlling the Weapons of War: Politics, Persuasion, and the Prohibition of Inhumanity*; *Biotechnology, Security and the Search for Limits*; and *Education and Ethics in the Life Sciences*. More recently, he has been interested in the social, ethical and political issues associated with researching and writing about secrets, as in his books *Experimental Secrets* (2009), *How to Look Good in a War: Justifying and Challenging State Violence* (2012) and *The Dis-eases of Secrecy: Tracing History, Memory and Justice* (with Chandre Gould, 2017).

Part VII
Afterword

Afterword: Data in Transit

Helen E. Longino

Abstract The naïve fantasy that data have an immediate relation to the phenomena of the world, that they are "objective" in some strong, ontological, sense of that term, that they are the facts of the world directly speaking to us, should be finally laid to rest by the papers collected in this volume. In this afterword, I propose that these papers, investigating data journeys in fields from particle physics to urban planning, show that even the primary, original, state of data is not free from researchers' value- and theory-laden selection and organization.

The naïve fantasy that data have an immediate relation to the phenomena of the world, that they are "objective" in some strong, ontological, sense of that term, that they are the facts of the world directly speaking to us, should be finally laid to rest by the papers collected in this volume. One might think that "data journeys" catalog the way that raw givens are transformed as they move from their original context to other contexts, whether higher levels of abstraction in the same field or other fields of inquiry. These papers, investigating data journeys in fields from particle physics to urban planning, show that even the primary, original, state of data is not free from researchers' value- and theory-laden selection and organization. Once reactivated in a new context, once they have journeyed, the mutability of the data is even more starkly revealed. But it is just this mutability that demands of researchers' creativity and diligence in the preparation and transport of data. Data are the currency of science and, even though not nature announcing itself to us, must be protected as if they were, because they are the closest we have. Where philosophers have in the past treated data as brute, unanalyzed, givens for purposes of inference to hypotheses or theories, the essays in this volume approach data as themselves the outcome of research practices. The practice perspective reveals a multiplicity of data production and manipulation processes. I will underscore the ways activities reported in four of these papers demonstrate this point. I will then engage in some general

H. E. Longino (✉)
Department of Philosophy, Stanford University, Stanford, CA, USA
e-mail: hlongino@stanford.edu

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_20

reflection on the lessons from the workshop that spawned this volume for thinking about data. The journeys are various, the fields even more so.

The contributions from Ramsden and Müller-Wille illustrate the challenges of obtaining information from data. Of course, one can count and measure any number of things. The trick is to measure the right things and to measure and report them in a way that will serve cognitive and practical purposes – one's own and those of others. Ramsden's paper tracks the progressive improvement in the quality of data on housing in mid-twentieth century United States, setting it alongside changes in the perception of the lives of the poor by middle-class professionals. Mueller-Wille's follows anthropologist Franz Boas's even earlier development of standardization in anthropometrics. This is set alongside changing demographic preoccupations in the US. Both papers concern themselves with the technologies researchers developed to make physical data relevant to social and cultural questions, as well as with the institutions, norms, and practices within which those technologies were deployed.

Ramsden focuses on the efforts of Edith Wood to obtain data that could be used to create national housing standards. When she began her work in the early 1930s she could complain that what data existed were locally variable, so that generalizations across states and cities were not possible. Wood "knew" that housing stock was inadequate, but the available data sets measured different things and used different scales of measurement. A break came in the form of data collected for commercial purposes: the Real Property Index collected for the real estate industry. This was a nationwide survey and classification of real property that made possible meaningful comparisons of residential housing. It included information on availability of utilities (gas, electricity), running water, bathing and toilet facilities). Its utility for Wood was its broad geographical reach and the consistency of items catalogued or measured. Wood was then able to find enough data on various indices of ill health (mortality, morbidity, delinquency) that she could map areas of more intense social ill health onto transparencies and overlay those on maps of housing quality created with the RPI. With correlations established in enough municipalities, the housing data could come to signify the intensity of social ills without need for further study. Wood's purpose was to advocate for minimum building standards as essential to a healthy society. Once her stress on comparability of data was accepted, the data became more fine-grained and the standards more demanding. Ramsden traces the trajectory from increased amounts of data, an increase that eventually made the data unusable, to the selection of key elements that could be taken as informative of a range of qualities. More is not always better.

What Ramsden also shows is that the impact of information depends on the aims of those with the power to use the information. The impact of uniform housing codes depended on the attitudes and aims of those for whom the data were produced. In Wood's time the goal was to improve the quality of life of the poor and indeed rates of disease and crime were (initially) lower in those areas where housing stock conformed to the new standards. But it takes more than square footage and running water to create a good life. In an urban context, it takes transportation, jobs, shops, spaces for social life. In the 1950s and 1960s the social reformers' goal of public health was replaced by urban renewal which came to mean the wholesale destruction of neighborhoods perceived as plagued with substandard housing (and the associated

social ills) and their replacement with high-rise, uniform apartment blocks. In some cases the residents of neighborhoods classified as substandard were moved wholesale into new areas and placed in housing that conformed to the physical requirements validated by the extensive data collection and analysis Ramsden chronicles. But, as the Boston example shows, the uprooting of families and neighbors destroyed social bonds that had come to flourish in less than ideal physical circumstances. Those social bonds, the relations of friendship and acquaintance as well as of commerce, are just as important in overall health as the physical requirements codified by housing advocates. New kinds of data sought and generated in the wake of the West End project in Boston signaled the lack of transportability of quality of life indicators that had to do with the interrelationships of residents with one another. The conclusion seems to be that these indicators are local, and the measurable indices too variable to be applicable across municipalities and states. It might be possible to read the lesson as that the conditions of "healthy" neighborhoods are plural: there is more than one way to make for good quality of life. Such a lesson is not generally welcome in a context looking for uniform and universal standards. And so, the story of data on housing and public health comes in some ways full circle, but through paths that demonstrated the importance of minimal standards, the relation between housing conditions and disease (especially communicable diseases such as tuberculosis) and what from one perspective one could call crime and from another, security.

Mueller-Wille's study of Franz Boas's anthropometric methods is, like Ramsden's, a story of efforts to integrate physical with social information. Boas was a staunch environmentalist, believing that both environment and biology contributed to the expression of specific physical characteristics in individuals. To this end his anthropometric effort was directed at matching continuity and discontinuity of physical characteristics with genealogical relations and tribal identities. Boas selected a small number of physical variables but measured them on a large number of individuals. In particular he conducted a major study of the Chickasaw, a people of the southeastern United States that was relocated together with the Choctaw as Americans of European descent pushed into Indian lands. Mueller-Wille stresses the hybrid character of the data Boas developed. Height and cranial features could be measured with standardized measuring instruments. Genealogical information, including tribal affiliation of parents, was, however, provided by the subjects themselves. Tribal affiliation was, in turn, determined by the informants according to internal standards of kinship and belonging (whether mother or father was of the tribe, etc.). Boas was interested in these data for theoretical purposes. His interest in pursuing this line of research among American Indians flagged when the data came to be seen as relevant to the highly political issues of tribal membership and entitlement. In spite of Boas's loss of enthusiasm for his study, the data have been preserved and have now been used to ground longitudinal studies of the (descendants of) the populations Boas studied and in statistical reanalyses of the data themselves. Mueller-Wille alludes to some of the recent debates about Boas's study of immigrants to the United States, a study he took up after abandoning the research on native Americans. This work, too, was caught up in political controversy, as Boas (and many after him) argued that his data showed the role of environment in the production of physical traits like cranial size. Such findings are not at all to the lik-

ing of those who urge restricting immigration of certain ethnic groups based on the assumed immutability of certain of their physical characteristics.

The essays of both Ramsden and Mueller-Wille, like those of Cambrosio and colleagues, Bechtel, Boumans and Leonelli, and Ankeny, make evident the struggle to identify and provide usable data. In both cases, the data gathered in one context needed to be comparable with data taken in another context. This required identifying what data could be found in all the contexts that needed to be compared, and selecting what among those variables would be most informative. Uniform measuring tools were necessary as well as universally available targets of measurement. Observers needed to be trained to use the tools, such as calipers, and to perform the measurements of, for example, the right cranial dimensions. And, in both cases, techniques of visualization are also required to make the import of the data evident. Overlaying transparencies marked by frequency of socially undesirable phenomena over municipal maps marked according to quality of housing stock enables even the non-statistically literate to see the connection between quality of housing and health. Drawing up kinship trees and putting information in tabular form again enables a "reading" of the anthropometric data not facilitated by mere reporting without attention to presentation. The first journey is the journey to comparability and association, facilitated by the selection of categories of data and of tools and the training required to use them; the second to visibility, facilitated by techniques of presentation. Furthermore, in both cases, the data are enlisted for further purposes. These purposes also leave their mark on the character of the data. So, the conceptions of public health current in mid-century support a focus on internal features of a home – square footage, running water, availability of a toilet – but not the elements of social glue that bind the inhabitants of those homes into a community. The need to have all members of a population represented for a comprehensive kinship mapping of the physical traits requires that obtaining the data involve no violation of modesty (disrobing) thus limiting what measurements can be performed. Finally, there is also a sense in which the data break free (or are broken free) from the contexts of their production and get deployed towards aims that the original research may not have envisioned and might not even endorse. The emphasis on the internal, physical requirements for healthy living determined the kinds of data that were available about neighborhoods and when urban planning shifted its emphasis slightly from public health, the welfare of the inhabitants, to urban renewal, a more comprehensive and impersonal value, the role of the data expanded from supporting building standards and encouraging or enforcing renovation to supporting destruction of neighborhoods whose housing stock seemed unamenable to updating and displacing their former inhabitants. Boas could see how his data could be used not just for the theoretical environmentalism he advocated but also in debates about pure-bloodedness, tribal membership and access to the benefits associated with both. When data are comprehensive (that is, include information on selected variables for an entire or very large segment of a population) they lose some of the apparent mooredness to their context and take on an independent life that makes them available for reuse in other contexts. As Mueller-Wille reminds us however, in spite of appearances, a crucial part of the data Boas gathered on the Chickasaw was indissolubly rooted to its context, as tribal identity was recorded based on the testimony of the subjects.

Koray Karaca takes us inside the black box of a particle detector. This is a research world characterized by very different challenges than the social worlds explored by Boas and by Edith Wood and her followers. His detailed description of the data preparation process in the ATLAS experiment at the CERN LHC nevertheless reveals some similar patterns as characterize the housing and anthropometric data generation, but at a vastly greater scale. The similarity is in the need to find the telling data among quantities of events, just as in the previous two cases the challenge was to identify the relevant variables. But whereas the issue in the previous two cases was to identify the particular variables that were both universal and variable enough to enable data to exit the contexts of their production for comparison with similarly produced data, in the case of the High Energy Particle Physics world, the challenge is to winnow down the unmanageable amounts of data produced in any given run of the collider. A succession of triggers selects events that will be informative given the aims of any given experiment. Collision events produce masses of different kinds of particle at various energy levels. Many, perhaps most, of these are already well understood. The point of the experiments is to find the rare events that are evidence of predicted but not yet detected particles, like the Higgs boson was for ATLAS, or particles or events that indicate physics processes not foreseen in current physical theory, the Standard Model. At the first selection level, the point is to thin or prune the data to a more manageable size, so that events of interest, that is, potentially informative events, will be more salient. What remains as data are the products of (a very small proportion of) the collision events and at the next stage these are further reduced while the remaining products are amplified by adding information about the trajectories known to produce those particular products or signatures. So, the collision data is reintroduced based on theoretical understandings of the particle and energy properties. Finally, at the third level selection triggers reflecting just what it is the researchers hope to find are applied to the products of the second level of selection. The products of this third selection are the data that will be subject to analysis.

Clearly a great deal of theory is required to design the triggers. So, while the data are not theory-laden in the old Kuhnian sense they are certainly theory-mediated (to adopt a phrase of George Smith's). Karaka describes a process of transforming the "blooming, buzzing confusion"[1] of collision events into data suitable for analysis and for use as the basis of inferences about particles. While on a first reading it may seem that the sequence of triggers renders the data hopelessly theory-dependent, it is important to remember that their "provenance" remains available. Unless one wants to call into question the entire enterprise of High Energy Physics, the relation of the final data to the wild unmanageable dance in the collider is readable through the technical specifications of the series of triggers. Seen this way, on a second reading, what Karaka has described at the LHC is in some ways a better documented production of data than most. James's phrase refers first to the human infant's experience of the sensible world, but can also apply to the indefinite number of ways we can individuate the contents of our sense perception. That we humans perceive as we do (three-dimensional medium sized objects perceived via wave lengths within

[1] To repurpose William James' famous description of infant perception (James 1890).

a small part of the full spectrum) is a product of our evolutionary history, but even within what we could perceive, we attend only to a small portion. While some of the selections we make can be reconstructed by reference to the interests we have in extracting information from a given perceptual experience, many of the criteria by which selections are made are buried in pre-conscious neural processes. Unlike the processes of the detector, our processes are opaque to us, revealed only by researchers studying mammalian perceptual systems. Whatever the world out there is like, its signals must be processed in order that we can begin to make some sense of them. In both the constructed detector and the human detector, what is important is that the selection processes, however much they may transform the inputs, retain the relationships (of relative magnitude, of time, of extension) as the data travel from input to cognitively accessible.

Finally, Coopmans and Rappert take us into the exotic (or, per their coda, perhaps not so exotic) world of art and art forgery. Here the stakes are complicated. In the three previous contemplated papers, the stakes are faithfulness and accessibility in service to pragmatic or theoretical values. This paper makes clear that, in the art context, faithfulness and accessibility have more than an epistemic interest. Conceptually speaking, the appraisal of a (putative) work of art has (at least) two dimensions, not always distinguished, and often conflated. One is the identification, perception, and communication of the aesthetic values exhibited by a particular object: the interactions of color, form, figure, and meaning. The other is the attribution to a particular hand, the hand of Leonardo, Rembrandt, or skipping ahead a few centuries, Pollock. In the former case the aesthetic properties of a work and one's abilities as a connoisseur (whether professional or amateur) who can detect them are at stake. In the latter, millions of dollars. The art world has always been a world of mixed motives and mixed values. Artists need to live, after all. But the twentieth century, in particular, has seen a boon in the secondary art market of dealers and auction houses, where works attributable to the hand of a Picasso, a van Gogh, a Monet, a Rembrandt, fetch sums the makers themselves never dreamed of. In such a world, the temptation to produce look-alikes, forgeries, is overwhelming. What is relevant to determining the real from the fake?

Coopmans and Rappert draw on the memoirs of various players in the art world to bring out the contested nature of evidence in this world. Hoving, the connoisseur, "knows" at a glance both whether a work possesses aesthetic value *and* whether it is a genuine work by the artist to whom it is attributed. Indeed, for him, there is no difference between these judgments. Like Boyle summoning credibility for his air-pump by inviting gentlemen onlookers, Hoving tries to enlist us as witnesses by pointing to the telltale details of the work that give away a forgery as fake. We, the non-connoisseurs, ought to be able to see as well as he, once tutored (and given enough study and tutelage might even come to be able to make such judgments on our own). Is this a real X? Here, let me show you. But there is another kind of detail: the physical trace, especially the fingerprint, but chemical analysis, too. Chemical analysis can help identify a fake as a fake, but the fingerprint links the object to a body in a way that does not depend on our being able to appreciate what the connoisseur asks us to "see.". Here the presumption is a causal chain from the print on the object to the hand of the painter and here we can see the orthogonal system of

value at work. A not very good Pollock or Picasso can still fetch higher sums at auction than a (judged by aesthetic criteria) better painting by an unknown. Does the worth of the work lie in its intrinsic aesthetic qualities or in its origin? If someone has been clever enough to produce a painting that looks every inch a Matisse or a Renoir (but is not), should we care? Would we put it up in our living room or entrance way? Would we enjoy it less? This depends on the nature of the pleasure it gives us, the pleasure of responding to aesthetic qualities or the pleasure of pride of ownership. Of course, there are complicated issues of originality, as well as of quality, that could be pursued in a fuller discussion of the relative value of originals versus forgeries, but this paper draws our attention to the wealth of data extractable from a work of art, as from "nature." What data are relevant, what data will travel well and what data will not depend on the purposes for which data are sought and the contexts in which they are to be deployed.

These papers reveal a variety of forms of travel. Data can be made visible through juxtaposition with other data, can be repurposed (from commercial valuation purposes to public health purposes, from theoretical to political), can be rearranged, can be reduced and recreated, can replace and be displaced. Other papers in the volume reveal how data from one source must be manipulated and adjusted in order to be compared or integrated with other data from different sources. Much discussion centers on the importance of preserving metadata in order to preserve the integrity and meaning of the data.

In all the tracings of data *journeys* we go back to the origin of the data, the start of the journey, and here the papers reveal the variety of means by which data are gathered, stored, and deployed. So, no, there is no such thing as "raw data."[2] There are the phenomena of the world, some perceptible to us, some not, always filtered by our means of perception. The resulting data, however closely linked to the phenomena, are always symbolic representations in some medium. Even the samples collected by dragging a receptacle through the sea become symbolic of what is left behind. To be informative about those phenomena, a single datum must be set in the context of other data: as discussed by several chapters, and particularly Mary Morgan's and James Griesemer's, we must have a data set or a singularity set against a background of other data. The data are always selected and produced, a function of the techniques of observation, of measurement, of recording. The expression "raw" is, however, trying to get at some aspect of the process. As we study the journeys we need to go back to some originary point where the data have been subjected to the least processing, a point closest to their context of generation. There are different ways to convey this notion. Niccolò Tempini suggested "source" versus "derivative." In another context we might think of less versus more defeasible. If data travel, there is a place or a status from which they travel. So we might think in terms of base level measurements that can be used to plot against other base level or against higher level measurements to engage in comparative analysis or to tease out the significance of the measurements. What counts as base level will depend on the context and what counts as a higher-level set of correlations in one context may be base level in another.

[2] A point also emphasized by the title and contents of Gitelman, ed. (2013).

But even if there is no such thing as self-announcing data, it doesn't follow that all data are "just interpretations." A naturalistic versus an interpretive approach may be too coarse a distinction for data, although does reflect a difference in how science studies has approached data. The naturalistic approach (found among science practitioners and some philosophers) is accused of naiveté, not understanding the work that goes into generating data that can be used to support scientific inference. The interpretive approach (found among constructivist science studies scholars) is accused of undermining the trust we rightly place in science. While data need other data and sometimes also need theory in order to "speak", to focus on the interpretative dimensions does not mean that the data are unreliable or fake, but that data must be selected, must be classified, must be set in relation to other data. What requires attention are the methods for obtaining, recording, and storing data and how well those methods serve the purposes for which data was sought in the first place. This doesn't place data practices above criticism, but helps us see the multiple places where criticism can be directed and therefore the multiple places where data practices may require defense. The perspective of science practice reveals a wealth of epistemologically relevant moves in the research context. Our understanding both of the trustworthiness of science and of its limits will be enhanced by the attention to data practices manifested in these essays.

References

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.

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.

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.

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.

Gitelman, Lisa, ed. 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.

James, William. 1890. *Principles of Psychology*, 488. New York: Henry Holt.

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.

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.

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.

Tempini, Niccolò. this volume. 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.

Helen E. Longino received her PhD in Philosophy from the Johns Hopkins University in 1973. Her teaching and research interests are in philosophy of science, social epistemology and feminist philosophy. She is particularly interested in the relations between scientific inquiry and its social, cultural and economic contexts. In addition to *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality* (University of Chicago Press, 2013*),* she is the Author of *Science as Social Knowledge* (Princeton University Press, 1990), *The Fate of Knowledge* (Princeton University Press, 2001) and many articles in the philosophy of science, feminist philosophy and epistemology and Coeditor of *Scientific Pluralism* (University of Minnesota Press, 2007). She has taught and lectured at universities in Europe, Asia and South America, as well as in the United States. She is currently C.I. Lewis Professor in Philosophy at Stanford University.

Visual Metaphors: Howardena Pindell, Video Drawings, 1975

Niccolò Tempini

Abstract Closing reflections on data journeys through the contemplative reading of a visual work by Howardena Pindell.

As Grace Deveney reports (2018),[1] the series *Video Drawings* came together when feminist African American artist, curator and educator Howardena Pindell took a sheet of acetate and drew signs, arrows and numbers upon it. Then she hung it on her television screen, taking advantage of the static that glues the light plastic to the screen glass. Using a camera, she took photographs of the transient juxtapositions that formed in front of her. Finally, she selected the set of most interesting combinations from the lot, including "those that had a "weird" sense of movement" (153).

[1] Deveney, Grace. 2018. Interrupting the Broadcast: Howardena Pindell's Video Drawings. In *Howardena Pindell: What Remains to Be Seen*, Beckwith, Naomi and Cassel Oliver, Valerie (eds.), 151–168. Munich: DelMonico Books-Prestel & Museum of Contemporary Art Chicago.

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

Alan Turing Institute, London, UK
e-mail: n.tempini@exeter.ac.uk

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_21

Untitled (Video Drawing: Baseball Series), August, 1975. Chromogenic color print (Pindell, Howardena (b. 1943): Untitled (Video Drawing: Baseball Series), August, 1975. Chromogenic color print, 4 5/8 × 6 7/8″ (11.7 × 16.7 cm). Purchase. Acc. no.: 276.1976. New York, Museum of Modern Art (MoMA). © 2019. Digital image, The Museum of Modern Art, New York/Scala, Florence. Courtesy the artist and Garth Greenan Gallery, New York.)

This *Video Drawings* work might speak as any other picture to the reflections about data and science practices that have been shared in this book. So many lines could be pursued that my few words here might sound arbitrary. I share them to invite, not to foreclose.

By juxtaposing notations to image Pindell prefigures something of data practices and analytics – "They resemble weather movement notations, dance notations, and particle tracks" (162). But just as with data, there is nothing obvious about them.

Let me first point out a few parallels, starting from the most immediate – *Video Drawings* is a curated collection of stills, selecting the compelling and the surprising. Highlighting some at the expense of others, the artist had to make certain assumptions consequential that shape the nature of what is shared with others and add explicit intention to an experiment. More interestingly, the works organised in this set are visual juxtapositions. They exploit the predilections of some materials to work together, and in so doing they facilitate the linking with one another of two heterogeneous records. The TV image becomes something else, as it is transformed by Pindell's re-purposing practices. Through the stilling of the moving image in a

photograph, a record of bodily and material configurations has been extracted (cfr. The digital 'scraped') from material released for public consumption rather than careful analysis – the broadcast of a sport performance. Like data, these are objects that can be repeatedly discovered to reveal something new. Their circulation can restart.

Deveney (2018) points out how the *Video Drawings* juxtaposition eventually collapses the original image, erasing distinctions between foreground and background, body and field. The low resolution of the TV screenshot makes us think of instruments pushed to the limits. It makes us wonder about the fragility of the processes with which data can be put to previously unimagined uses. The visual record is rescued from a rather adventurous path of transmission and yet, for all the loss of detail in the grainy still, other, new information seems to emerge in the juxtaposition with the acetate sheet. The artist, presumed spectator of the TV stream, reclaims a role and intervenes, highlighting the creative opportunities opened by constraints and omissions. New links are suggested, various relations between data points can be postulated. But the juxtaposition remains ultimately open-ended as to its most important message. Consensus remains controvertible, and several narratives might successfully navigate the visual space. The acetate sheet, with its pre-recorded content, seems to suggest the existence of frameworks and assumptions enabling and structuring ways of reading the record, which are now lost history. The drawn notations are not analysis of the picture themselves, but could they be abstractions obtained from other analyses? At least they seem available to participate in one.

Yet, for all the stimulation it offers, the composition refuses to promise that any new meaning can be found. As Deveney observes (2018), the composition hangs between order and chaos, predictability and randomness. It might not be what it seems. It remains unclear what its final use or destination might be. In this suspension, the work reminds us of many steps of data journeys that we have discussed in this book – somewhere between the past moment of generation and the future moment of definite interpretation.

Niccolò Tempini is Senior Lecturer in Data Studies at the University of Exeter, Department of Sociology, Philosophy and Anthropology, and a Turing Fellow at the Alan Turing Institute. He is an interdisciplinary social scientist interested in questions of information, data, technology, organization, value and knowledge. He researches Big Data research and digital infrastructures, investigating the specific knowledge production economies, organization forms and data management innovations that these projects engender with a focus in their social and epistemic consequences. He studies the practices of data scientists, software developers, researchers and nonprofessionalised experts to understand how different forms of knowledge and value intersect with each other when different actors come to grips with new methods and new forms of data, information technology and organization. His research has been published in international journals across science and technology studies, information systems, sociology and philosophy (more information at www.tempini.info).

Index

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