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."

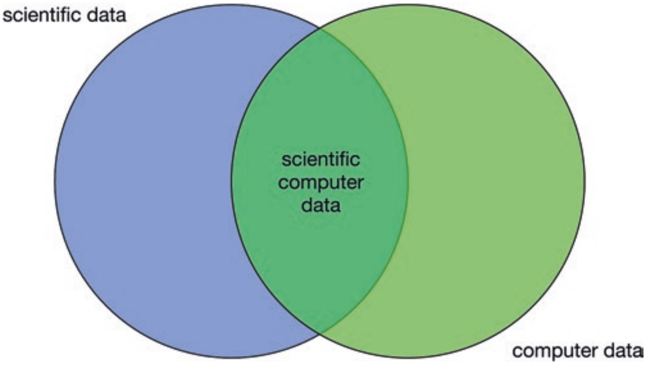

Fig. 1 Scientific data, computer data, and scientific computer data

- 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 re-assembly, 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).