Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2016: 1–12

DOI: 10.1177/2053951716654502
bds.sagepub.com
SAGE

Data journeys: Capturing the socio-material constitution of data objects and flows

Jo Bates[1], Yu-Wei Lin[2] and Paula Goodale[3]

Abstract

In this paper, we discuss the development and piloting of a new methodology for illuminating the socio-material constitution of data objects and flows as data move between different sites of practice. The data journeys approach contributes to the development of critical, qualitative methodologies that can address the geographic and temporal scale of emerging knowledge infrastructures, and capture the 'life of data' from their initial generation through to re-use in different contexts. We discuss the theoretical development of the data journeys methodology and the application of the approach on a project examining meteorological data on their journey from initial production through to being re-used in climate science and financial markets. We then discuss three key conceptual findings from this project about: (1) the socio-material constitution of digital data objects, (2) 'friction' in the movement of data through space and time and (3) the mutability of digital data as a material property that contributes to driving the movement of data between different sites of practice.

Keywords

Critical data studies, data practices, data cultures, data materiality, Big Data, data flow

At 09:00 UTC (10:00 British Summer Time) on 24 June 2014, two different instruments located at Sheffield Weston Park weather station observed a temperature of 18.5°C. One of these instruments, owned by Museums Sheffield, generated a data point that was recorded in a CSV file. Later that day, a curator at Weston Park Museum uploaded it to an Access database. From here, the curator circulated it in the local community via Twitter, and at the end of the month, it was automatically emailed to the Met Office as part of the station's climate record for June 2014. The second datum was generated by equipment owned by the Met Office and was automatically transmitted by electromagnetic signal to the World Meteorological Organisation's Global Telecommunication System. From here the datum replicated as meteorological organisations around the world, including the Met Office data centre in Edinburgh, downloaded and ingested it into their systems. At this point, the datum replicated again and began to travel down a series of different paths. In almost real time, Met Office weather forecasters in Exeter incorporated the datum into their numerical weather prediction (NWP) system. Simultaneously, the Met Office distributed the datum, for a fee, to commercial actors such as firms that supply meteorological data to the weather derivatives industry. At a slower pace, the datum also travelled to the MetDB synoptic database, from where climate scientists calculated climate averages for June 2014 before saving the data to the MIDAS database to be used by other climate scientists.

The life of data

In this article, we discuss our development and piloting of a new methodology for illuminating the *life of data*

[1]Information School, The University of Sheffield, UK
[2]University for the Creative Arts, UK
[3]University of Sheffield, UK

Corresponding author:
Jo Bates, Information School, The University of Sheffield, Room 236, Regent Court, 211 Portobello, Sheffield S1 4DP, UK.
Email: jo.bates@sheffield.ac.uk

as they journey between and through such sites of data practice (places where people are engaged in practices of data production, processing, distribution and use). The *data journeys* approach that we introduce – piloted on the Secret Life of a Weather Datum project – aims to focus attention on the *life of data* as they move through space and time, through different sites and cultures of data practice, on their *journey* through the 'information production chain' (Braman, 2006) from initial production through to re-use in different contexts. The demand to unpack the 'life of data' (Beer and Burrows, 2013; Ruppert et al., 2012) has emerged from diverse positions across the social sciences and humanities. In many cases, this call has taken aim at deeply reductionist and technocentric forms of data practice observable in the new academic fields of Data Science, Computational Social Science and Social Physics, as well as in parts of business and government, as researchers and organisations respond to the promise of 'Big Data'. In response to such trends, Kitchin (2014a) demands a 'situated, reflexive and contextually nuanced epistemology' to counter disruptive 'data-driven' methods within the academy, Dalton and Thatcher (2014) call for critical data studies that recognise the 'contingent and contested social practices' that shape the production and interpretation of all data, and others have drawn attention to the ways in which data are 'cooked' (Gitelman and Jackson, 2013) and 'made' (Vis, 2013).

The observation that data are socially constituted objects is not new. Earlier work in the philosophy of science has explored data as something that are produced within a social context (Jensen, 1950 in Kitchin, 2014b). More recently, social scientists researching scientific knowledge infrastructures have pointed to the ways in which the conceptualisation of data as neutral and objective is mistaken (Bowker and Star, 2000). For Bowker (2008), '"raw data" is both an oxymoron and a bad idea'. Through 'inverting' (Bowker, 1994; Bowker and Star, 2000) and looking beneath the surface of knowledge infrastructures and recognising them as social and relational, scholars have explored some of the complex and often invisible political, cultural and ethical processes that contribute to their development (see Bowker et al., 2010; Edwards, 2010; Star, 1999). Gitelman and Jackson (2013: 4) argue for the adoption of this 'infrastructural inversion' approach for understanding the socially situated production of 'Big Data', calling on researchers to 'look under data to consider their root assumptions' and to question the material conditions of their production.

Yet, for the most part, critical research on emergent 'Big Data' practices and infrastructures has remained at the conceptual and theoretical level (Kitchin, 2014b). Whilst various calls have been made for critical engagement with the philosophical and methodological assumptions surrounding 'Big Data' (boyd and Crawford, 2012; Dalton and Thatcher, 2014; Gitelman and Jackson, 2013), relatively few scholars have conducted empirical work on specific 'Big Data' practices. Amongst those that have, many have remained external to sites of data practices, relying upon documentary analysis to inform empirical investigation (Hogan, 2015; van der Vlist, 2016; Williamson, 2015). Yet, in order to contribute to the development of alternative futures in which 'publics might be said to have greater agency and reflexivity vis-à-vis data power' (Kennedy and Moss, 2015), it is important that critical 'Big Data' research gets 'under the hood' to grasp how local and situated 'Big Data' practices structure how data work in the world, and thus how particular practices, and their social consequences, might be ameliorated. There is therefore a growing need for methodological approaches that are able to capture detailed empirical understanding about 'Big Data' in practice, including how socio-material factors influence the constitution of data objects and shape how they move through space and time connecting different sites of practice across vast data infrastructures.

Outside of this body of work on 'Big Data', the published research that empirically examines data practices from a sociological perspective tends to be rooted in the field of infrastructural studies. Influenced by actor-network theory (ANT) and similar Science and Technology Studies (STS) methodologies, this field of research emphasises the production of detailed accounts of specific knowledge or data infrastructures and the intra-network politics of their development (e.g., Edwards, 2010; Leonelli, 2013; Ruppert et al., 2015). The *data journeys* approach is related to this body of work in terms of its interest in empirically examining people's practices of data production, processing, distribution and use. However, the concept of a *data journey* aims to better situate data across interconnected sites of practice distributed through time and space, drawing attention to the movement of data between these sites, and beginning to respond to the call from key thinkers in the field of knowledge infrastructure studies for the development of 'a methodological repertoire that can match the geographic and temporal scale of emerging knowledge infrastructures' (Edwards et al., 2013). In so doing, the approach places emphasis on the diverse social worlds that are interconnected, in part, by the *journey* of data through and between different sites of data practice, with the intention of illuminating the concrete ways in which evolving socio-cultural values and material factors cohere over time to create the socio-material conditions that frame activities of *data production, processing and distribution* and resultantly influence the form and use of data and their movement across infrastructures.

The structure of the paper is organised as follows. We begin by discussing the materiality of digital data, and how this relates to the social world. We then present a theoretical rationale for the *data journeys* approach, followed by the research design for our empirical work on meteorological data journeys. We then discuss three key conceptual findings from our research about the socio-material constitution of digital data objects and flows. The first of these focuses on the socio-material constitution of key data points and datasets. The second on the movement of data through and across infrastructures, drawing upon Edwards' (2010) concept of 'data friction' to examine some of the socio-material conditions and power dynamics that constrain the mobility of data as they move between sites of practice. The third observation explores the mutability of weather data as they move through and between sites, reflecting on the ways in which this mutability contributes to driving data between different sites of practice.

The materiality of data

In order to address questions relating to the *life of data*, it is important first to consider what digital data are and how they relate to the social and material world. As well as having symbolic properties that differentiate them from other forms of informational resource (see, for example, Borgman, 2015; Buckland, 1991; Rowley, 2007), digital data, similar to other informational artefacts, can also be recognised as 'material objects' (Dourish and Mazmanian, 2011). In this sense, data can be understood as 'material-semiotic ''things''' (Wilson, 2011). The materiality of digital data can be understood and theorised in multiple ways. We can observe the physicality of digital data stored in the magnetic atoms of a hard drive, and when they are transmitted wirelessly as electromagnetic signals – as Edwards (2010: 84) argues 'data are *things*...with dimensionality, weight, and texture'. The materiality of data can also be recognised in the sense that they are the product of a particular set of practices (Wilson, 2011), through which cultural values materialise in the form data take, in a similar way to how ideologies and values become visible in the built environment (Harvey, 1991). Data also have material consequences, and we can pay attention to the ways in which they have 'practical instantiation and...significance' in the world (Leonardi, 2013). Perhaps more subtly, we can illuminate the material factors that cause data to have consequences. Dourish and Mazmanian's (2011) work on the materiality of digital representation points to four such factors: (1) the material conditions of their production that impact significantly upon data generation, processing, distribution and use; (2) the physical materiality of data infrastructures which impacts how space is used and imagined; (3) the 'material properties' of data objects (e.g., size, durability, mutability) that impact what data represent and how people encounter, use and transform them and, (4) how data enable us to view things through an informational lens which impacts how we might perceive those things. We recognise all these forms of materiality in our conceptualisation of digital data. However, for clarity, we employ the terms: *physical infrastructure* to refer to material artefacts such as computers and instruments, the *physicality of data* when referring to their atomic and electromagnetic form and *crystallisation* when discussing how socio-cultural values gain substance in the materiality of data and their infrastructures. We limit our use of the word material to refer, firstly, to the *material conditions of production*, and secondly, to the '*material properties*' of data – factors such as their 'mutability, persistence, robustness, spatiality, size, durability, flexibility, and mobility' (Dourish and Mazmanian, 2011: 4).

In foregrounding the materiality of data, we draw upon Harvey's (1991) work on the 'internal relation of the material structure of ideas' to recognise that ideas and values do not work as an external controlling force over data practices, rather they act as a framework and justification for the activities that practitioners are engaged in (Bieler and Morton, 2008: 118). Rather than externally shaping material forms, ideas and values 'tak[e] on substance through practical activity bound up with systems of meaning' that are often embedded in the economy (Bieler and Morton, 2008: 119). Whilst acknowledging this interrelationship between the socio-cultural and the material, we also recognise the necessity for some analytical separation of the two categories. In so doing, we aim to avoid imagining the 'socio-material' as a constitutive entanglement (Orlikowski, 2007) pre-existing perception, and instead recognise socio-material structures as being historically constituted through the actions of both historic and present-day human actors (Bieler and Morton, 2001; Leonardi, 2013).

Data journeys: Theoretical framework

Drawing upon the above framework, we developed and piloted the *data journeys* methodology as an approach for illuminating the socio-material *life of data* as they travel between and through different sites of data practice. The approach is based on following data through multiple interconnected organisations and projects within and across knowledge infrastructures.

Interest in the movement of data through space is seen in a number of research areas. For example, Beer and Burrows (2013) draw upon Mackenzie's (2005) concept of the 'performativity of circulation' to explore the role of popular culture in the accumulation and

flow of new forms of social data. Similarly, researchers in the interdisciplinary field of mobilities studies have examined the ways in which the movement of people, objects, capital and information impacts social and economic life (Sheller and Urry, 2006). In the more empiricist traditions of Information Science, analysis of the flow of data across their lifecycle within information systems and knowledge infrastructures is relatively common. Sands et al. (2012), for example, develop a 'Follow the Data interview protocol' to study the flow of data leading into and out of astronomers' research publications in order to understand the people and infrastructures responsible for the development of large astronomy sky surveys. Similarly, McNally et al. (2012) use the 'data flow' concept in research design to produce detailed accounts of the durability, replicability and metrology of flows of data within data intensive research contexts, and examine how 'people, infrastructures, practices, things, knowledge and institutions' work together to shape the flow of data through these spaces. These bodies of research have produced detailed pictures of data flows and practices across a range of contexts. However, in general, they have tended to emphasise the internal dynamics of specific knowledge infrastructures and information systems, and neglected to situate these practices in relation to the wider socio-material contexts and power dynamics shaping their development.

Whilst academic research in this field has tended to refer to the 'flow' of data within a given context, the term 'flow' tends to suggest a disconnect of data from physical sites of data practice. The concept of a *data journey* aims to better locate data in physical space; places which should not be imagined as 'self-contained' units, but as sites constituted in part by social relations external to their particular locale (Massey, 1994: 5). 'Flow' also suggests the smooth movement of a liquid. However, as Borgman (2015) observes, 'Data do not flow like oil'. Here, *journey* may better symbolise the disjointed breaks, pauses, start points, end points – and 'friction' (Edwards, 2010) – that occur as data move, via different forms of 'transportation' (wires, electromagnetic waves, etc.), between different sites of data practice across knowledge infrastructures. Further, to conceptualise a data object as something that journeys, rather than flows, helps draw attention to particular moments of 'mutability' (Manovich, 2001) of data objects, potentially illuminating some of the diverse ways in which data – as mutable, programmable objects – are adapted for different ends by practitioners situated at different sites across the infrastructure.

The term *journey* also reminds the researcher that their role is not simply to map the movement of data through space, but also to be a traveller – to stop off, take in their surroundings and absorb the culture. This process of journeying as method is observable in early work in the field of cultural studies, for example, Raymond Williams' (1958) bus journey through the Welsh towns and villages of his youth, recounting stories that point to the shape of the culture and its transformation over the years. More recently, it is visible in research that engages what Sheller and Urry (2006: 217) describe as 'mobile ethnography'.

Drawing on these ideas, we began to imagine a research design in which the researcher moves through space following data on their *journey* through interrelated sites of data practice. However, it was not only the movement of data from point A to point B that interested us. We were also interested in paying attention to potential movement, blocked movement and lack of movement (Sheller, 2011: 6); the temporality of these movements – their speed and timing; forked journeys as data were replicated and re-used in different ways in different places; intersecting journeys as data from different sources were linked or combined in some way; if and how data mutate as they travel from site to site and, end points in data journeys – for example, as a result of data deletion, corruption and obsolescence.

These spatial dynamics of digital data are shaped by the historically constituted socio-material conditions that human actors encounter, reproduce, subvert and ameliorate as they engage in practices of data production, processing, use and distribution at different sites. In order to fully grasp the spatial dynamics of data journeys, it was therefore important also to explore their historic development, the way they have evolved over time and how the transforming shape of *data journeys* relates to the evolution of the broader socio-material conditions of which they form a part. As Massey (1994) observed, there is an 'inherent dynamism of the spatial' (p. 4). The ways in which data move between sites of data practice are not static; the places and connections are dynamic, evolving over time, and emergent socio-material conditions can open up new possibilities for *data journeys* through space and time.

Data journeys: Research design

Taking these theoretical observations into consideration, the design of the Secret Life of a Weather Datum project aimed to illuminate the socio-material constitution of meteorological data objects and flows. We began by conducting an initial mapping of key data journeys. Drawing upon initial desk research, we began by identifying UK-based sites of weather data production and use across state, science, market and civil society. We then mapped the journeys of data between relevant organisations, projects, datasets and individuals using post-it notes on flipchart paper (Figure 1).

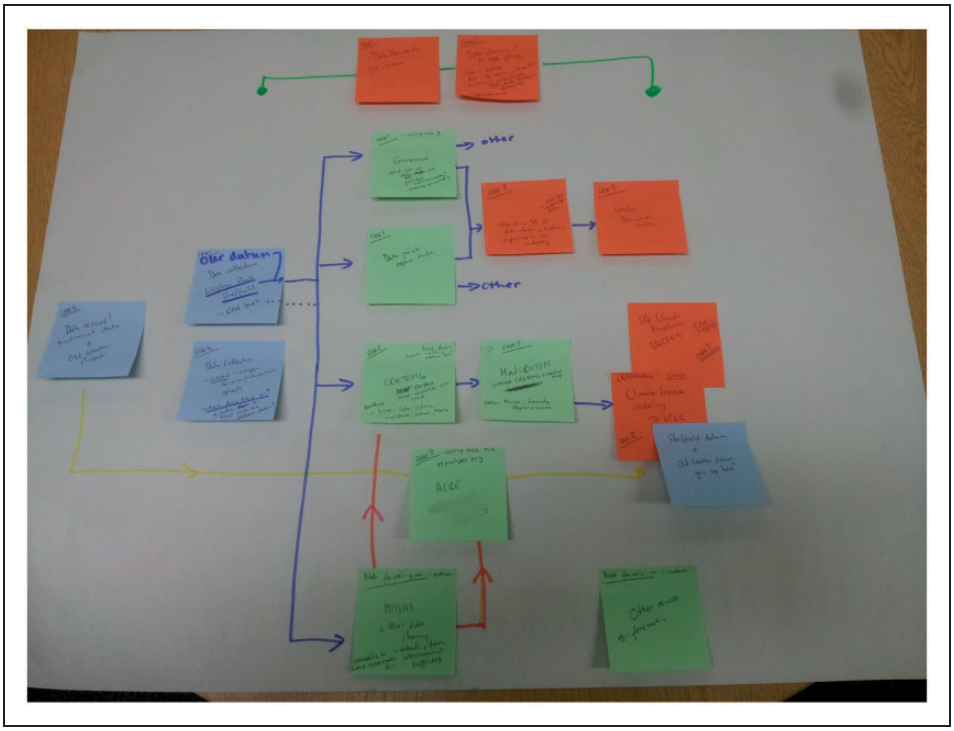

Figure 1. Initial Rich Picture Mapping of Data Journey (anonymised).

This visual representation was then adapted as new information was gathered about the detail of the *data journeys* we uncovered.

Our initial mappings allowed us to identify a number of potential data journeys to explore, and after making initial enquiries regarding access to research sites, we decided to focus on the journeys of data produced at our local weather station (Sheffield Weston Park) and data produced by amateur weather observers and citizen scientists. We then followed these data on their journeys from sites of production on into processing by the UK's Met Office, and on into re-use in climate science and financial markets. We also explored the intersecting journeys of data generated by amateur observers and citizen scientists. As well as identifying key informants through desk research, snowball sampling techniques were also adopted once we were in the field. In total, primary data were gathered in relation to eight sites of data practice: Sheffield's Weston Park weather station, Met Office headquarters in Exeter, the Climatic Research Unit at the University of East Anglia, the Inter-governmental Panel on Climate Change (IPCC), archives that store historical weather observations, the Old Weather citizen science project, amateur weather observers in distributed locations, and a firm that supplies weather data to the weather derivatives market. At each of these sites, primary data were generated including, as appropriate to each site, in-depth interviews incorporating an oral history component with data practitioners and other relevant individuals, field observations involving reflective field notes and photography, digital ethnography of selected forums and Twitter hashtags and documentary analysis of policies, legislation and other relevant sources. Through adopting an element of oral history interviewing in our conversations with participants, we were able to draw upon their memories of the development of the infrastructure in order to construct an evolutionary and dynamic picture of the *life of data* which emphasises key moments in the development of data journeys and practices.

The primary data we generated were used to illuminate the journey of data through and between each site, the specific data practices that people were engaged in at each site, the socio-cultural values that framed and were used to justify participants' data practices, and the varying material conditions and institutional contexts of these practices, including an analysis of the public policies and legislation that shaped the movement of data between sites. We also aimed to uncover tensions and changes in the socio-cultural constructs that practitioners were bringing to their data work at different sites, and explore how these constructs are interrelated with the broader socio-material context.

Our initial findings have been published on a public facing website – http://lifeofdata.org.uk – that was developed as part of the project. The interactive website draws upon a tube map metaphor in order to represent

visually the journey of data as they move between the different sites of data practice that we explored. Each of these sites is represented by a 'clickable' station on the tube map, and within each station the user is invited to explore the different data practices, cultures, and public policy frameworks that contribute to the production of digital data, and their movement between, and use across, different sites. Where permissions from research participants were, granted original research data including audio interviews and photographic images are embedded into the website to bring the story to life. The dynamic nature of the infrastructures we explored is also manifest in the design of the website through our efforts to represent participants' memories of particular moments during the evolution of data prac tices over time (see Figure 2).

Data journeys: Insights and reflections

Through adopting the *data journeys* methodology in our empirical work, we have developed further our understanding of the ways in which evolving socio-cultural values and material factors cohere over time to create the socio-material conditions that frame activities of *data production, processing* and *distribution*, and resultantly influence the form and use of data and their movement across infrastructures. This section will discuss three key conceptual observations we made through our empirical work about: (1) the socio-material constitution of digital data objects, (2) different forms of socio-material 'friction' (Edwards, 2010) experienced by data as they move (or not) across space and time between different sites and (3) the mutability of digital data as a material property which contributes to driving the movement of data between different sites. The intention of this section is to present some key conceptual findings. Empirical findings can be explored further at: http://lifeofdata.org.uk, and more detailed empirical analyses will be published in further papers.

The socio-material constitution of digital data objects

Our analysis demonstrated the ways in which the practices of those who produce meteorological data are bound up in complex systems of meaning that crystalise

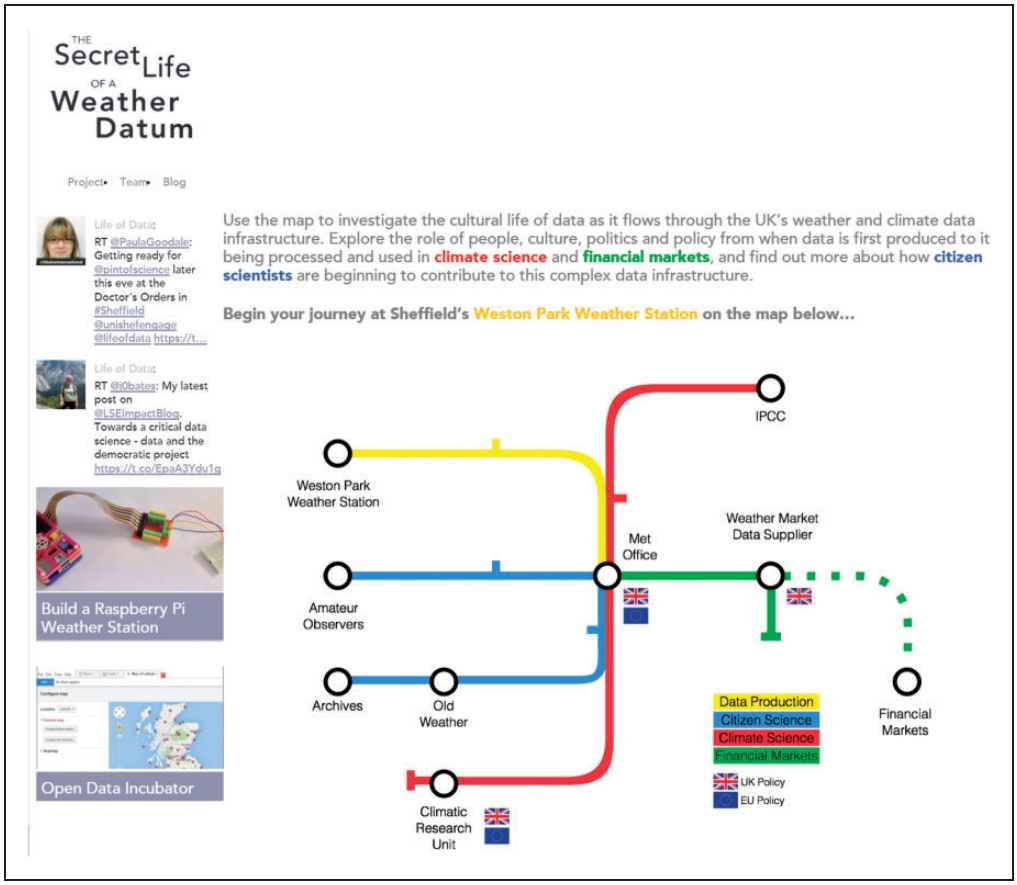

Figure 2. A screenshot of the project website http://lifeofdata.org.uk (as of 30 June 2016).

in the *material form* and *properties* of data (Dourish and Mazmanian, 2011). Even at the micro level of a digital datum – for example, the two data points recording a temperature of 18.5°C on 24 June 2014 mentioned in the opening vignette – we observed how inter-related socio-cultural values, practices and material conditions of production, evolving over time, came to take on substance at a particular moment in the form of two single data points.

The Weston Park weather station was founded and first began recording data in the 1880s in response to a fatal outbreak of diarrhoea in the city. It was suspected that the cause of the outbreak was related to weather temperature, but doctors needed data so they could view the problem through an informational lens, and ultimately predict future outbreaks and improve the public health of the city. The local Corporation, under pressure from the Department of Health, asked the museum curator Elijah Howarth to build and run a weather station. He decided to locate it at Weston Park, which was conveniently next to his place of work. The weather station has been active since this time and has produced one of the longest and most complete climate datasets on record.

Since the 1880s, responsibility for the station has passed down through generations of curators. The museum curator who currently looks after the station told us stories about how these individuals – barring a short period of variability in standards during the early years of the station – took great pride in looking after the station and the data it has generated (both digital and paper records), contributing to the durability and persistence of the climate dataset over the years. Data generated by the station are produced to international climate data standards. Prior to automation this was not an easy task, and over the years curators have been found at the station on Christmas day in order to reset instruments to ensure accurate data generation. The curator perceived the station as part of the local fabric of Sheffield and remembered how local people came to its rescue – fundraising £3000 in a 'matter of weeks' – when an expensive piece of equipment was stolen from the roof of the museum.

Over the years, the physical infrastructure of the station has survived other threats. It came unscathed through a bombing raid on the museum in World War Two – during which the curator braved high winds on the roof of the museum to capture hourly wind readings for the military. More recently, the station has been threatened by funding cuts to museums in the wake of the economic crisis, which have led to staffing cuts. In order to adapt to some of these pressures and ensure the continuity of the weather station, the previous curator allowed the national Met Office to add its own weather observation instruments to the weather station compound in 2010. This new equipment generates data alongside the museum equipment. It feeds its data in real time to the Met Office via the WMO Global Telecommunication System and is part of the national synoptic network. As the curator describes, this adaption in the physical infrastructure, whilst likely necessary to ensure the continued presence of a weather station at Weston Park, means the Met Office no longer depends on the climate data generated by the museum equipment. This development has resulted in a reduction in power for the local museum in its relationship with the national Met Office, and the emergence of a tension between the value system of the curator who looks after the weather station and perceives it and its data as part of the cultural heritage of Sheffield, and the more technocratic values of the distant national Met Office that emphasise efficiency, speed and volume of data.

In discussing his work and the history of the station, the curator expressed strong values of public service, civic duty, resilience, pride in his contribution, and responsibility to his forbearers, local community and data users. It was clear that the historical mix of cultural values, practices and material conditions of production outlined above inform and frame the values and practices of the current curator. We observed how, enabled and empowered by this history, the activity of the curator to protect, maintain and run the museum weather station and look after the Met Office equipment in the context of current socio-material conditions resulted in (1) the production of these two specific data points at that particular moment in time, influencing their accuracy, timing, unit of measurement and so on, (2) the specific *material forms* in which the data points were represented, for example, as digital objects stored in a CSV file, Access database, Twitter and Met Office databases such as the NWP system and MetDB synoptic database and (3) the *material properties* of the data points, for example, their mobility, persistence, durability and spatiality. In the form and properties of these data, we can observe how historically constituted values, practices and material conditions continue to have meaning and take on substance in the world.

Similar observations can be made of all other digital data that we observed across sites of weather and climate data practice. For example, the specific form of the CRUTEM4 global climate temperature dataset (Jones et al., 2012), which is derived in part from data generated at Sheffield Weston Park, has been shaped by struggles between climate scientists and climate change sceptics. After a prolonged and complex struggle involving the Climatic Research unit's (University of East Anglia, UK) email systems being hacked and a government inquiry into the Unit's scientific practice – which

found accusations of misconduct 'patently false' (House of Commons, 2010) – climate scientists accepted the sceptics' demand for full transparency of the underlying weather station data feeding into the CRUTEM datasets. However, this means that some data series cannot be included in the latest version, CRUTEM4, because of the socio-material conditions of their production – that is, for a variety of economic, socio-cultural and political reasons, their source country prohibit the data being made publicly available. These struggles around the publication of underlying station data have therefore crystallised in the specific form that the CRUTEM4 dataset takes, and gained substance as a result of the practical activity of climate scientists' decision making and negotiating around which specific data series can, and cannot, be incorporated into the global dataset.

Friction in data movement

Through examining the journey of data between different sites, we were able to identify factors that enable and restrict the movement of data across infrastructures, and observe sites of potential movement, blocked movement and lack of movement (Sheller, 2011: 6). It was evident that whilst data are often mobile between sites, they do not necessarily move smoothly or easily from one place to another – they experience 'friction' (Edwards, 2010) as a result of the complex socio-material contexts they exist within.

Most obviously, we can observe the diverse forms and levels of friction experienced by data as they move along different types of path through space and time from historical ships into the International Comprehensive Ocean Data Set (ICOADS). The journey begins with the slow movement of handwritten data points inscribed in the log books of Royal Navy ships. The data in these log books spent time slowly traversing the oceans, before being removed from ships and deposited in archives around the world where they were boxed up and left untouched for years in varying states of decay. We can then observe the 'friction' reduce as climate scientists began to recognise the *potential* for data movement, and acquired pockets of funding to 'recover and rescue' data from the log books. The log books were extracted from the archives and transported to facilities where they were digitised, before digital copies were transmitted via the internet into the living rooms of citizen scientists working on the Old Weather project. These volunteers transcribed the digital copies of these handwritten data points via a 'cloud' platform, through which they were written to a server, before being captured by climate scientists and integrated into the ICOADS. Nevertheless, despite this clear reduction in 'friction' enabled by developments in both the physical and social dimensions of the infrastructure, significant amounts of data remain locked away in archives: their paths into the ICOADS database blocked primarily by the *material conditions of production* – namely a lack of public funding for 'unsexy' data recovery projects.

We can also observe how policy makers, commercial actors and civil society campaigners have recognised *blockages* to *potential* data movements between public bodies such as the Met Office and third parties such as commercial re-users of meteorological data. In some cases, efforts to overcome these *blockages* have led to new policies and legislation, for example, Open Data policies and Re-use of Public Sector Information regulations. However, whilst Open Data policies that allow anybody to access and freely re-use data apply to some data, for example, some data produced by the Met Office, Open Data policies are not the norm across the infrastructure.

The weather observation data produced by Museums Sheffield at Weston Park, for instance, was shared readily with the Met Office, as well as with students and researchers at the local universities. The curator was also responsible for fostering a rich local data ecology in which weather station updates were shared with the public on Twitter and in the local newspaper. However, whilst strongly in favour of the idea that these data belonged to the public, the curator was wary about Open Data policies given that the sustainability of the weather station was dependent upon the small-scale commercialisation of the weather data it produced; a factor that seemed unlikely to change soon given the financial challenges posed by recent cuts to public funding for museums. These material conditions that shape the production of the museum's meteorological data have a significant impact upon the curator's understanding and practices regarding opening data generated by the museum equipment, highlighting how 'friction' in the movement of data can reflect, and be shaped by, power dynamics at play in the wider context. In this case, the curator's efforts to keep the museum weather station going in the face of significant reductions in public spending, and a reduction in the importance of the museum's data since the Met Office installed its own equipment at the site, is dependent upon creating and maintaining some data 'friction'.

Elsewhere at different sites across the infrastructure we can observe some policy makers, politicians and financial market actors pushing to reduce 'friction' in data movement by calling for the removal of charges for commercial re-use of Met Office historic and real-time bulk data in order to reduce costs and spur innovation in the weather derivatives industry. As a public sector Trading Fund, the Met Office is institutionally

obliged to generate revenue through the commercial exploitation of the goods and services it produces, although in the case of the Met Office revenue comes primarily from the services it is contracted to provide to the UK government and public sector. These material conditions of Met Office data production and processing mean that, similar to the curator at Weston Park, there has been some resistance to Open Data policies that are perceived to risk the financial stability of the organisation during an era of deep public sector restructuring. Despite these issues, some meteorological data have been opened. However, even when there is the will to make high volumes of frequently updated, highly detailed data available for others to re-use, material barriers exist to making that happen in practice, for example, as a result of datasets being updated four or more times a day, and models getting bigger and more detailed, the volume of data being processed is significant and growing. The sheer volume of data therefore presents a material challenge when the Met Office wants to make data available to third parties.

Overall, our findings demonstrate that where data do and do not end up on their journey from production through to re-use is influenced by a range of interrelated socio-material factors, including: the material conditions of their production and efforts to sustain physical infrastructures and institutions given this context; the material form and properties of data such as their size and speed; the relative power and influence of different actors who desire to shape the movement of data; and the socio-cultural values framing beliefs and practices around the role of publicly funded infrastructure, data sharing and valued forms of data re-use. Illuminating the causes behind these shifting patterns of 'friction' as data move, or not, between sites of data practice has the potential to provide a fascinating insight into the power dynamics that are shaping emergent material conditions of production.

Data objects as mutable mobiles

Observations of what happens to digital data when they do move indicate that 'mutability' is an important *material property* of data (Dourish and Mazmanian, 2011). Digital data, as objects that embody Manovich's (2001) five principles of numerical representation, modularity, automation, variability and transcoding, are easily manipulable and hence mutable. Similar to other forms of digital documentation (Borgman, 2010) and new media (Manovich, 2001), digital data are not 'fixed once and for all, but something that can exist in a myriad of forms and copies' (Manovich, 2001: 36). As Law and Mol (2001) argue, this mutability refers not only to the shape of the object itself but also extends to variation in what it means for the object 'to *work*' (pp. 5–6) at different sites.

The mutability of objects as they move through Euclidean space, for example, between two sites located in different geographical locations, has been explored within the field of ANT and, more broadly, STS. The *data journeys* approach, which emphasises the movement of data between different sites, allowed us to unpack the mutability of data within and across infrastructures. It allowed us to question, as data move through space and time: does the socio-material context force them to hold their original shape, or are they adapted based on the needs of different sites? We observed that digital data are a form of 'mutable mobile' (Law and Mol, 2001) – as they move between sites, practitioners remix, repurpose, and adapt them in different ways for different ends. Similar to de Laet and Mol's (2000) 'Zimbabwean bush pump', both the data and the socio-cultural values and relations that crystallise within them mutate as practitioners process data across diverse sites of practice. This high level of mutability also contributes directly to their usefulness for the different practitioner groups who work with and shape them and can therefore be recognised as a key material factor driving the movement of data between different organisations and projects.

Significant examples of data being adapted as they move through different sites are practices of data cleaning and homogenisation. Every day data arrive from meteorological offices around the world at the firm that supplies data to the financial markets. A team of 3 to 4 people then spend all day, every day, analysing the data, looking for unusual readings and missing data points for stations including Sheffield Weston Park. If errors or gaps are found in the data, data points are adjusted and filled in based upon the team's climatological knowledge and readings from surrounding stations. Data points are also altered based upon knowledge of particular weather stations gained from historic and present-day station metadata; for example, if the weather station location or instruments have changed over the years, incoming data are 'homogenised' in order to correct for the resulting divergence in the observations. The changes made to particular data points enable incoming data from different sources to be aggregated in a way that generates a more uniform representation of the weather appropriate to the context of use. A detailed audit trail of any changes made is saved in the database. A very similar process happens when data arrive at different sites across the infrastructure, for example, the Met Office and the Climatic Research Unit. However, the temporal dynamics of these mutations differ; for example, the pace at which data are cleaned and homogenised is much slower for climate data processing, where

accuracy is more important than speed, than for data being fed into the weather derivatives industry and forecasting where immediacy is vital. Further, whilst the mutability of digital data is always technically present, social factors intersect to restrict this mutability at particular points. When a particular version of a dataset is used in a weather derivatives trade, for example, data become immutable – they will not undergo any further alterations, as this would impact upon the validity of the contract. Similarly, when a new version of a climate dataset is published, for example, CRUTEM4.4.0.0 (Jones et al., 2012), data points are made temporarily immutable. However, unlike in the case of the weather derivatives contract, they may change again at a future point when the next version of the dataset is published. In many cases, these practices of cleaning and homogenisation of data that take place at different sites are undertaken in order to generate datasets that are accurate and complete enough for the purposes to which they are to be put, whether that be climate research or financial market trades.

Our research only touched on a few examples of data mutability across a small number of sites. However, the reproducibility and reconfigurability of digital data allows these mutations to happen simultaneously, at scale, and largely independently of one another, factors which contribute substantially to the complexity and scale of data in existence. Their mutable nature enables digital data to be re-used, re-purposed, and put to work for different purposes in different places and contexts, factors which contribute to driving the movement of data between different sites across infrastructures. Whilst the value systems of some of these different sites of data practice may conflict, for example, those of Weston Park and the data supplier to the financial markets, the data as 'mutable mobiles' connect these different human actors in complex, often invisible, relations that together form a key component of emergent socio-material conditions.

Conclusions

As it becomes increasingly clear that emergent 'Big Data' practices across a variety of domains are contributing to the re-constitution of socio-spatial relations, it is crucial that methodologies are developed and research conducted that help to illuminate the concrete ways in which 'Big Data' are constituted through complex socio-material practices, and how they contribute to the ongoing reconstruction of socio-spatial relations. The *data journeys* methodology aims to contribute to this endeavour, shifting the gaze of critical 'Big Data' research from that of an external onlooker to one in which the researcher becomes embedded alongside data as they journey through 'Big Data' infrastructures from sites of production through to diverse sites of re-use; and from an emphasis on the 'bigness' of 'Big Data' to a focus on the specific ways in which the small and local make up 'Big Data' infrastructures. Through identifying particular moments in the socio-material constitution of data objects and flows as meteorological data move between different sites of practice, the *data journeys* methodology allowed us to begin to capture some of the ways in which diverse social worlds are becoming increasingly interconnected and interdependent as they contribute to, and are impacted by, emergent practices of data production, distribution and use.

We identified the importance of historically constituted socio-cultural values in shaping practices of data production. The dedication and pride taken in the production of data at Sheffield Weston Park over the years contributes directly to their scientific and economic value at other sites including the Met Office, climate science and financial markets, as well as to the data and the weather station being valued as an important part of the cultural heritage of Sheffield. We also identified how the material properties of digital data – for example, their volume, mutability and durability – impact on how data move between different sites. Increasing volumes of quality data may contribute to their desirability for potential re-users, but can also pose challenges for how best to distribute data between sites. Their mutability – the ways in which they can be cleaned, homogenised, and otherwise re-configured, linked and aggregated with other data – increases the re-usability of data, and therefore contributes to generating demand and driving the movement of data between sites. Through drawing attention to the broader power dynamics influencing the evolution of these practices – particularly material conditions of production such as a lack of public investment and funding for data recovery projects and sites such as Weston Park, as well as broader questions in the UK about the structure and governance of public institutions such as the Met Office in the context of a deep restructuring of the state – the approach also shed light on the ways in which data practices and journeys are deeply politicised. We observed that different actors were working to influence the distribution of data between sites for a range of political and economic ends from the restructuring of public institutions, to the protection of local infrastructure and cultural heritage in this context, to efforts to deepen the financialisation of climate uncertainty through pushing to open data used by the weather derivatives industry. Overall, the *data journeys* methodology illuminated the ways in which data are produced, processed and used across diverse sites of practice that are interconnected by the movement of

data across space and time, the ways in which socio-cultural values and material factors come together to frame and give justification for these practices, and how together these contribute to the production of emergent socio-material conditions.

Acknowledgements

We would like to thank the three anonymous reviewers for their suggestions for improvements to the article, which contributed significantly to its development.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was supported by the Arts and Humanities Research Council AH/L009978/1.

References

Beer D and Burrows R (2013) Popular culture, digital archives and the new social life of data. *Theory, Culture & Society* 30(4): 47–71.

Bieler A and Morton AD (2001) The Gordian Knot of agency—structure in international relations: A neo-Gramscian perspective. *European Journal of International Relations* 7(1): 5–35.

Bieler A and Morton AD (2008) The deficits of discourse in IPE: Turning base metal into gold? *International Studies Quarterly* 52: 103–128.

Borgman C (2010) *Scholarship in a Digital Age*. Cambridge, MA: MIT Press.

Borgman C (2015) *Big Data, Little Data, No Data*. Cambridge, MA: MIT Press.

Bowker G (1994) Information mythology: The world of/as information. In: Bud-Frierman L (ed.) *Information Acumen: The Understanding and Use of Knowledge in Modern Business*. London: Routledge, pp. 231–247.

Bowker G (2008) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker G and Star S (2000) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Bowker G, Baker K, Millerand F, et al. (2010) Toward information infrastructure studies: Ways of knowing in a networked environment. In: Hunsinger J, Klastrup L and Allen M (eds) *International Handbook of Internet Research*. Dordrecht: Springer, pp. 97–117.

boyd D and Crawford K (2012) Critical questions for Big Data. *Information, Communication & Society* 15(5): 662–679.

Braman S (2006) *Change of State: Information, Policy, and Power*. Cambridge, MA: MIT Press.

Buckland MK (1991) Information as thing. *Journal of the American Society for Information Science* 42(5): 351–360.

Dalton J and Thatcher J (2014) What does a critical data studies look like, and why do we care? Seven points for a critical approach to ''Big Data''. *Society and Space Open Site*. Available at: http://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data (accessed 11 June 2016).

De Laet M and Mol A (2000) The Zimbabwe bush pump mechanics of a fluid technology. *Social Studies of Science* 30(2): 225–263.

Dourish P and Mazmanian M (2011) Media as material: Information representations as material foundations for organizational practice. In: *Working paper for the third international symposium on process organizational studies*, Corfu, Greece, June 2011. Available at: http://www.douri.sh/publications/2011/materiality-process.pdf (accessed 11 June 2016).

Edwards P (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Edwards P, Jackson SJ, Chalmers MK, et al. (2013) *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor: Deep Blue. Available at: http://pne.people.si.umich.edu/PDF/Edwards_etal_2013_Knowledge_Infrastructures.pdf (accessed 30 June 2016).

Gitelman L and Jackson V (2013) Introduction. In: Gitelman L (ed.) *Raw Data Is an Oxymoron*. Cambridge, MA: MIT Press, pp. 1–14.

Harvey D (1991) *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Oxford: Blackwell.

Hogan M (2015) Data flows and water woes: The Utah Data Center. *Big Data and Society* 2(2): 1–12.

House of Commons (2010) *The Disclosure of Climate Data from the Climatic Research Unit at the University of East Anglia*. London: The Stationery Office Limited. Available at: http://www.publications.parliament.uk/pa/cm200910/cmselect/cmsctech/387/387i.pdf (accessed 11 June 2016).

Jensen HE (1950) Editorial note. In: Becker H (1952) *Through Values to Social Interpretation*. Durham, NC: Duke University Press, cited in Kitchin R (2014) *The Data Revolution*. London: Sage.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., & Morice, C. P. (2012) Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres*. 117, D05127, pp. 1–29.

Kennedy H and Moss G (2015) Known or knowing publics? Social media data mining and the question of public agency. *Big Data and Society* 2(2): 1–11.

Kitchin R (2014a) Big Data, new epistemologies and paradigm shifts. *Big Data and Society* 1(1): 1–12.

Kitchin R (2014b) *The Data Revolution*. London: Sage.

Law J and Mol A (2001) Situating technoscience: An inquiry into spatialities. *Environment and Planning D: Society and Space* 19(5): 609–621.

Leonardi PM (2013) Theoretical foundations for the study of sociomateriality. *Information and Organization* 23(2): 59–76.

Leonelli S (2013) Global data for local science: Assessing the scale of data infrastructures in biological and biomedical research. *BioSocieties* 8(4): 449–465.

Mackenzie A (2005) The performativity of code: Software and cultures of circulation. *Theory, Culture & Society* 22(1): 71–92.

McNally, R., Mackenzie, A., Hui, Al, Tomomitsu, J. (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): 81–94.

Manovich L (2001) *The Language of New Media*. Cambridge, MA: MIT Press.

Massey D (1994) *Space, Place and Gender*. Minneapolis, MN: University of Minnesota Press.

Orlikowski WJ (2007) Sociomaterial practices: Exploring technology at work. *Organization Studies* 28(9): 1435–1448.

Rowley J (2007) The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 33(2): 163–180.

Ruppert, E., Harvey, P., Lury, C., Mackenzie, A., McNally, R., Baker, S. A., Kallianos, Y. & Lewis, C. (2015) *Socialising Big Data: From concept to practice*. Available at: http://www.cresc.ac.uk/medialibrary/workingpapers/wp138.pdf (accessed 30 June 2016).

Ruppert E, et al. (2015) Socialising Big Data: From concept to practice. In: *CRESC working paper no. 138* (February). Available at: http://www.cresc.ac.uk/medialibrary/workingpapers/wp138.pdf (accessed 11 June 2016).

Sands, A., Borgman, C. L., Wynholds, L., & Traweek, S. (2012) Follow the data: How astronomers use and reuse data. *Proceedings of the American Society for Information Science and Technology* 49(1): 1–3.

Sheller M (2011) Mobility. *Sociopedia*. Available at: www.sagepub.net/isa/resources/pdf/mobility.pdf (accessed 11 June 2016).

Sheller M and Urry J (2006) The new mobilities paradigm. *Environment and Planning A* 38(2): 207–226.

Star SL (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3): 377–391.

van der Vlist FN (2016) Accounting for the social: Investigating commensuration and Big Data practices at Facebook. *Big Data and Society* 3(1): 1–16.

Vis F (2013) A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday* 18(10). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4878/3755 (accessed 11 June 2016).

Williams R (1958) Culture is ordinary. In: Szeman I and Kaposky T (eds) *Cultural Theory: An Anthology*. London: John Wiley and Sons, pp. 53–59.

Williamson B (2015) Educating the smart city: Schooling smart citizens through computational urbanism. *Big Data and Society* 2(2): 1–13.

Wilson M (2011) Data matter(s): Legitimacy, coding, and qualifications-of-life. *Environment and Planning D: Society and Space* 29: 857–872.