



King's Research Portal

DOI:

[10.1177/2053951718786316](https://doi.org/10.1177/2053951718786316)

Document Version

Publisher's PDF, also known as Version of record

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data infrastructure literacy. *Big Data & Society*. Advance online publication. <https://doi.org/10.1177/2053951718786316>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Data infrastructure literacy

Big Data & Society
July–December 2018: 1–13
© The Author(s) 2018
DOI: 10.1177/2053951718786316
journals.sagepub.com/home/bds



Jonathan Gray¹ , Carolin Gerlitz² and Liliana Bounegru^{3,4} 

Abstract

A recent report from the UN makes the case for “global data literacy” in order to realise the opportunities afforded by the “data revolution”. Here and in many other contexts, data literacy is characterised in terms of a combination of numerical, statistical and technical capacities. In this article, we argue for an expansion of the concept to include not just competencies in reading and working with datasets but also the ability to account for, intervene around and participate in the wider socio-technical infrastructures through which data is created, stored and analysed – which we call “data infrastructure literacy”. We illustrate this notion with examples of “inventive data practice” from previous and ongoing research on open data, online platforms, data journalism and data activism. Drawing on these perspectives, we argue that data literacy initiatives might cultivate sensibilities not only for data science but also for data sociology, data politics as well as wider public engagement with digital data infrastructures. The proposed notion of data infrastructure literacy is intended to make space for collective inquiry, experimentation, imagination and intervention around data in educational programmes and beyond, including how data infrastructures can be challenged, contested, reshaped and repurposed to align with interests and publics other than those originally intended.

Keywords

Data infrastructures, information infrastructure studies, science and technology studies, digital methods, data activism, data literacy, data publics, data journalism, critical data studies, data critique, data worlds

Introduction

What is to be done about the apparently ever-increasing volumes of digital data and ever-multiplying processes of “datafication” in society? One common response is data literacy. As we examine below, many data literacy initiatives focus on developing technical, computational and statistical competencies for working with datasets. In this article, we propose and develop the notion of “data infrastructure literacy” in order to both conceptualise and encourage critical inquiry, imagination, intervention and public experimentation around the infrastructures through which data is created, used and shared. Through this notion, we hope to suggest ways in which literacy initiatives might broaden their aspirations beyond data as an informational *resource* to be effectively utilised, by looking at how data infrastructures materially organise and instantiate *relations* between people, things, perspectives and technologies. Data infrastructure literacy programmes aim not only to equip people with data skills and data science but also to cultivate

sensibilities for data sociology, data culture and data politics.

There is a wealth of literature on the social and cultural study of data, information and knowledge infrastructures (see, e.g. Bowker et al., 2009; Edwards et al., 2009; Star, 1999; Star and Ruhleder, 1996). There is also a growing body of literature on “critical data studies” (see, e.g. Dalton et al., 2016; Iliadis and Russo, 2016). How might insights and approaches from these fields be brought to bear on the conceptualisation and practice of data infrastructure literacy? How can they be made relevant for different types of data?

We propose a working vocabulary for how research on data infrastructures might inform literacy initiatives,

¹King’s College London, UK

²Universität Siegen, Germany

³Universiteit Gent, Belgium

⁴Rijksuniversiteit Groningen, Netherlands

Corresponding author:

Jonathan Gray, King’s College London, London WC2R 2LS, UK.

Email: jonathan.gray@kcl.ac.uk

illustrated with a series of empirical vignettes and examples of what we call “inventive data practice”, drawing on previous and ongoing research on open data, online platforms, data journalism and data activism. Just as “inventive methods” are said to “introduce answerability into a problem” in a way which “should not leave that problem untouched” (Lury and Wakeford, 2012: 3), so inventive data practices may question and problematise the default lines of inquiry which are built into data infrastructures, including by re-assembling them in accordance with interests and publics which they were not originally designed for.

We suggest that data infrastructures can be viewed in terms of their alignment and mal-alignment with different kinds of interests, outlooks and concerns. Questions of alignment and mal-alignment may become more prominent as digital technologies are used to redistribute and multiply relations between data infrastructures and their publics, involving new and perhaps unintended actors in making sense with data. When data infrastructures are mal-aligned with particular interests and concerns, they may become an issue for those who wish to use them, leading to various inventive strategies for using and making data differently. Through these vignettes we aim to contribute to a “reflective understanding of the means which have demonstrated their value in practice”, as Weber puts it (2011).

Rather than thinking of data infrastructure literacy in terms of an agenda for the transfer of skills and the extraction of value, we propose that it may be seen as a site for ongoing public involvement and experimentation around infrastructures of datafication. This is particularly pertinent given recent public controversies around digital infrastructures and online platforms in relation to both “fake news” and recent presidential elections in the US which suggest the broader stakes and interests at play (Bounegru et al., 2018). Making digital data infrastructures visible and problematising them is, so we claim, not just possible in situations of breakdown from routine functioning (Star, 1999) but also in cases of mal-alignment with the concerns of the publics that they assemble.

Rethinking data literacy

Advocates suggest that data literacy will be the “most important new skill of the 21st century” (Venture Beat, 2014) and refer to the development of capacities and technologies to help companies, states and citizens make the most of their data. One argues that “competence in finding, manipulating, managing, and interpreting data” must become “an integral aspect of every business function and activity” (Harvard Business Review, 2012). Others warn against a

data literacy deficit, estimating a shortage of millions of “data-savvy managers and analysts” (McKinsey, 2011).

This interest in data literacy is shared by many in the public sector and civil society. A report from the UN makes the case for “global data literacy” in order to catalyse a “data revolution” for sustainable development (Data Revolution Group, 2014). Data literacy is envisaged as that which will enable “change agents” to advance progress towards “the future we want” (United Nations, 2012). Members of the group formerly known as the G8 have argued that data literacy is important in order to “unlock the value of open data” in the service of transparency, accountability and economic growth (G8, 2013).

Data literacy is thus imagined to play a crucial role in different visions of the world, society and the future. But what is it exactly? The UN’s Data Revolution website emphasises capacities to “use and interpret data”, reproducing a graphic depicting data literacy at the intersection of statistical literacy, information literacy and technical skills for working with data (see Figure 1).

Previous research on the topic characterises data literacy in terms of being able to access, analyse, use, interpret, manipulate and argue with datasets in response to the ubiquity of (digital) data in different fields.¹

However, narrower conceptions of data literacy that focus on skills to use data have been met with scepticism. Some have raised concerns about viewing it in terms of “competencies of an extractive and transformative industry” (Letouzé et al., 2015). According to this view, data is presented as a material to extract value from (whether economic, technological, social, democratic or otherwise), a perspective that corresponds with the notion of “information as a resource” (Braman, 2009: 12–15). Letouzé (2016) argues that such conceptions of data literacy may “reinforce and perpetuate, rather than challenge and change, prevailing power structures and dynamics”. Ruppert (2015) and Birchall (2015) argue that public data initiatives can privilege “auditorial” or “entrepreneurial” modes of action, subjectivity or citizenship. Conceptions of data literacy which focus on the value of data risk overlooking questions about the politics of data – including how data is made, how it might be made and used differently and who and what it assembles and attends to. In the following sections, we explore how literacy initiatives may look beyond “data skills” towards cultivating capacities to account for (and reshape) the wider socio-technical infrastructures through which data is created, transformed and circulated.

The concepts of “data infrastructure” and “information infrastructure” have a wide range of different uses

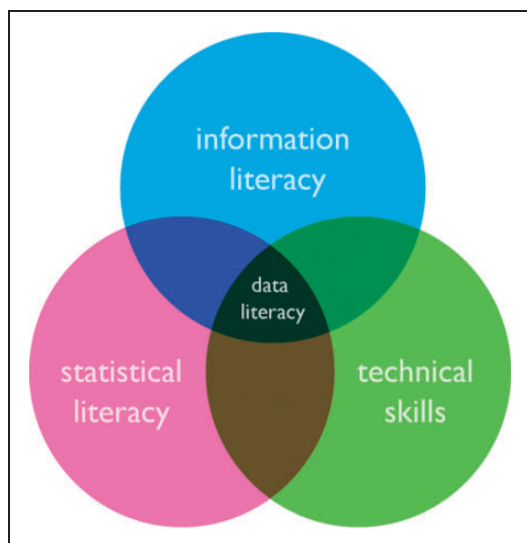


Figure 1. “What is data literacy?” graphic reproduced on UN Data Revolution website. <http://www.undatarevolution.org/data-use-availability/>.

and accompanying “socio-technical imaginaries” (Jasanoff and Kim, 2015). In information policy, these terms are often used to refer to the development of large scale technical systems for the creation, processing and distribution of information. A “National Information Infrastructure” became the centrepiece of US President Clinton’s initiative to support an “information super-highway”, which encompassed the “aggregate of the nation’s networks, computers, software, information resources, developers and producers” (Information Infrastructure Taskforce, 1993; Kahin, 1995). While this project focused on “networking the nation”, over the past few years the same phrase has also been used to describe systems underpinning the creation, processing and distribution of datasets (cf. Cabinet Office, 2015). We further draw on approaches from science and technology studies which evolved in parallel to these developments. This includes Star and Ruhleder’s (1996) proposal to consider information infrastructures in terms of *relations* rather than as “things”. In this view, data infrastructures are comprised of shifting relations of databases, software, standards, classification systems, procedures, committees, processes, coordinates, user interface components and many other elements which are involved in the making and use of data.

Why might one want to move beyond literacies with datasets and towards literacies with infrastructures, relationally conceived? One reason is that datasets do not simply neutrally *designate* aspects of the world, they also *render* the world in accordance with different visions, values and cultures, making it navigable through data. Data infrastructures can

carry a normative force as they produce data formats which prioritise certain ways of knowing over others (Marres and Gerlitz, 2015). At the same time their data are also multivalent and can be used in ways other than intended, by actors other than intended.

Data infrastructure literacy promotes critical inquiry into datafication, into how datasets are created with certain purposes in mind as well as opening up “infrastructural imagination” (Bowker, 2014) about how they might be created, used and organised differently (or not at all) – and the tensions that emerge between these two. It attends to situations of not only inventively repurposing data but also problematising data, gathering alternative data or not gathering data at all (advocating, regulating and designing for gaps, silences and spaces of non-datafication).

Disassembling data infrastructures

Critical engagement with data infrastructures has been central to various interdisciplinary perspectives from the past several decades, including infrastructure studies, data studies, science and technology studies, the history and philosophy of science, human–computer interaction, computer supported cooperative work, ethnomethodology, the history and sociology of quantification, software studies, platform studies, new media studies, critical design studies and associated fields. Our notion of data infrastructure literacy suggests that insights from these fields should be taken seriously by literacy programs.

As many social studies of data have pointed out, data is never “raw” in the epistemological sense of offering transparent, self-evident and unmediated access to phenomena (Bowker, 2005: 184; Gitelman, 2013: 2).² Letting go of the notion that data does nothing more than show us how things are, we can attend to the social, historical, cultural and political settings in which it is created and used and which framings such infrastructures introduce to the data. To this end, Bowker and Star (1999: 34) call for “infrastructural inversion”: bringing the background work involved in the making of data into the foreground and hence we can study the social practices which databases both reflect and enable, such as quantification, classification, commensuration and calculation. Sociologists and historians of quantification outline the links between the development of statistics and statecraft, and the making and governing of populations (see, e.g. Hacking, 1990, Miller, 2001; Porter, 1986, 1996). Desrosières (2002), for instance, shows how scientific and administrative innovations in France, Germany, England and America converged in social conventions for solidifying many aspects of socio-economic life into metrics and

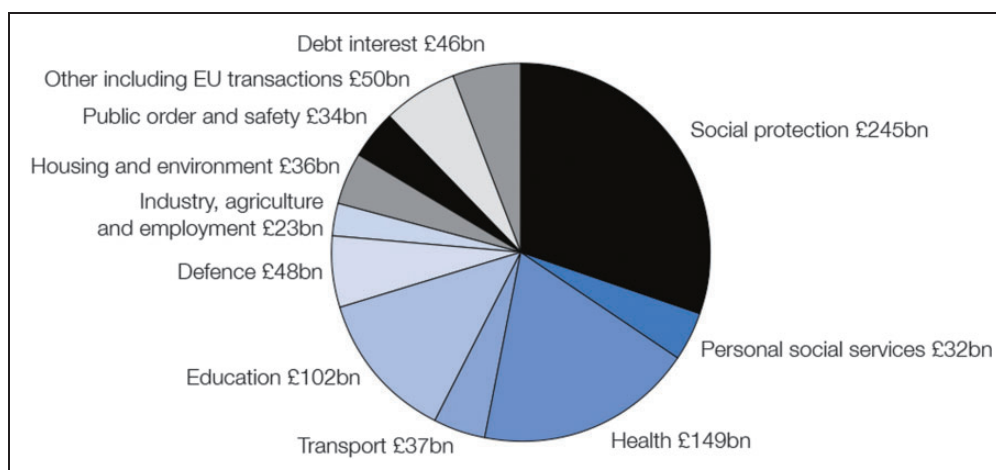


Figure 2. “Public sector spending 2017–2018”, UK Government. <https://www.gov.uk/government/publications/spring-budget-2017-documents/spring-budget-2017>.

measurements that we now take for granted – such as unemployment, inequality, growth and poverty.

Such literatures may serve as a source and a starting point for doing data infrastructure literacy. Agre’s (1997) notion of “critical technical practice” may inspire us to explore how critical, historical and sociological reflection on data infrastructures can be folded back into practical data work as part of what we might call “critical data practice” (Gray, 2018). Another starting point would be in situations of mal-alignment and the “inventive data practices” they may give rise to. To further explore the latter we will turn to two examples of when digital data infrastructures become a “matter of concern” (Latour, 2004) as they are mal-aligned with the interests of some of their publics: (i) open data on public finances and (ii) social media data from Twitter.

Open data infrastructures and fiscal mysteries

Open data has risen to prominence as a way to support transparency, accountability, participation and innovation by enabling citizens, civil society and companies to re-use public data in order to create new apps, analyses, products and services (Gray, 2014). This can change the social life of datasets which were created in relation to specific public sector policy objectives. Take public data about public money in the UK. We might start with a deceptively simple question: what does the UK government spend money on? A cursory search will produce tables and charts with overviews of how total spending is broken down by different areas (Figure 2).

Perhaps curiosity will be sated, deflected or deflated by these big numbers. But if we had more specific issues in mind, we may be disappointed. Where are the millions in IT contracts? How much goes to Deloitte, G4S

or Google? Does the UK spend more on turbines or warheads, education or fossil fuel subsidies?

While the above example lacks granularity, having lots of detail may open up new problems. The UK’s “data.gov.uk” website offers over 1800 datasets, many of which contain thousands of rows of transactions. For example, one document shows every transaction over £500 from the Natural History Museum in February 2017, giving us a peek into the routine affairs of a large museum: post, imaging, public transport. This highlights how infrastructures produce data of varying granularity and scale for different purposes, making distinct kinds of operations possible.

Literacy programmes focusing on data skills may encourage us to think about what we can do with such datasets – to explore and tell stories with them, creating pivot tables, regressions and visualisations. We may perform such operations without knowing much of the life of this data.

In the case of data about public finances, a huge amount of social, political, technical and organisational work goes into the production of a figure such as “social protection: £245bn”. And rather than a seamless and continuous transition whereupon we might “zoom in” from totals of billions to receipts of pennies, we are faced with an array of discontinuous snapshots responding to a barrage of diverse and sometimes conflicting demands – reflecting the colourful social life of public financial data.

Financial transactions are classified in relation to institutional objectives, which are in turn mapped onto relevant financial, statistical and accounting standards. Conventions, norms and standards are encoded through combinations of paper forms, drop-down menus and reconciliation work by accounting and finance teams. Intergovernmental bodies such as

the UN have the task of attempting to align these bureaucratic processes between states to enable transnational comparability – such as the Classification of Functions of Government schema, proposed in the late 1970s.³ Different sets of statistics and accounts must be produced in accordance with various institutional and policy rhythms, and with varying degrees of accuracy and estimation.⁴ Committees confer on methods for making available data fit with desired formats. Resulting conventions simultaneously constrain and enable policy and debate about public money.

The cast of characters involved in the social life of fiscal data is extended and diversified by a combination of access to information laws, public information policies and open data initiatives, which give rise to new “data publics” (Ruppert, 2015). Such publics are assembled through data portals, FOI requests and “civic technology” platforms such as WhatDoTheyKnow.com, and then extract and transform datasets for use in their own projects which may follow different aims from those of institutional data producers. For example, journalists and campaigners associated with the FarmSubsidy project were interested in finding out about how much large companies such as Nestlé receive from European funds. As this data was not published by European Union (EU) or national institutions, they undertook to request, transcribe, compile and align data from documents and spreadsheets in order to generate their own databases, leading to investigative projects and legal cases (Gray et al., 2012) and creating new “enumerated entities” in the process (Verran, 2015). In some cases, such data publics may also take on a role in shaping the data standards, norms and conventions which flow back “upstream” to institutions.⁵

This vignette about open data suggests how data which is initially neatly aligned with specific administrative interests through formatting and stabilisation into particular numerical formats may reach new publics through online and digital technologies which have quite different sets of interests and concerns about public finance. When datasets do not answer questions as hoped or expected, the infrastructures implicated in their creation may become a matter of concern.

Social media data infrastructures and “lively” grammars

Data from social media platforms gives us a different perspective. Platforms offer predefined possibilities for action and interaction such as posting, liking, commenting, sharing, tweeting or friending, which we may consider in terms of what Agre calls “grammars of action” (1994). Whilst it predates social media platforms, Agre’s account of grammatisation is informative when it comes to accounting for platform data

infrastructures. Following Agre, graphical interfaces only allow users to perform previously formalised and software-enabled “unitary actions” (p. 746) which are instantaneously transformed into corresponding data points. Action and datafication are thus designed to be co-constitutive.

A prime example of platform grammars are social buttons such as the Facebook “like” which for years meant that only positive responses were possible until it was opened up to include a slightly more granular grammar of reaction buttons in 2015, including for “love”, “laugh”, “wow”, “sad” and “angry”. Platform grammars do not merely capture actions but also shape what their users can do, delineating horizons of possible engagement and thus possible data points.

What is specific to these infrastructures, however, is that their grammars are simultaneously standardised in form, and also deliberately kept open to partial re-interpretation by various users, developers and other stakeholder groups of platforms (Gillespie, 2010; Rieder and Sire, 2013). To account for data infrastructures in the context of social media data, it is important to attend to a wider cast of characters and practices involved in its making. The “interpretative flexibility” (Bijker et al., 1987: 40–44) of platform data becomes particularly apparent in the case of Twitter’s favourite button (Paßmann and Gerlitz, 2014) which has been treated as both a bookmark and as a popularity measure by its users. Both practices were supported by third-party software that turned button-based activity into either a bookmarking service or a popularity ranking. Most platforms deliberately enable interpretive flexibility around their features and data by opening themselves up to platform interoperability (Bodle, 2010) or third-party developer systems (Rieder and Sire, 2013).

Whilst platforms allow their data to be circulated in new contexts, they are themselves subject to translation and commensuration of data. Recent research on Twitter data suggests that only a fraction of tweets are produced via the official web interface (Gerlitz and Rieder, 2017). Most come from mobile apps and third-party cross-syndication software such as IFTT or dlvr.it, but also custom scripts, professional social media clients as well as (semi-)automated services. These can be considered part of Twitter’s burgeoning infrastructure. They need to conform with Twitter’s platform grammars, but may also support alternative use scenarios including professional, team-based, promotional and spammy tweeting or new functionalities entirely. The heterogeneity of entities to tweet from enables new actors to produce platform data and offer distinct ways of “being on Twitter” (Gerlitz and Rieder, 2017). As in the case of public finance, platform data only appears stable at first glance: platform infrastructures and application programming interface

(API) regulations for instance enable the blending of data from one platform grammar into another, raising questions of commensuration (Espeland and Stevens, 1998): Can data from bots be compared with manually typed tweets? Can hashtags imported from Instagram be analysed together with those originating from Twitter? In addition, external entities not only produce but also promote distinct forms of analysing platform data, adding further levels of interpretation and inscription.

Social media data is thus articulated on several layers: through the platform grammars of user interfaces and platform databases; through the sources where data originates from (including other platforms, websites or algorithms); and through user practices. When working with platform data such as tweets, hashtags or likes, we encounter data at a specific stage in its life, and the work that went into it may not be immediately evident (Baym, 2014). The grammars of data infrastructures may thus be considered “lively” (Marres and Weltevrede, 2013; see also Gerlitz and Rieder, 2017), as they are stable in form, but can take on different meanings and interpretations when taken up by different publics or translated into new contexts.

These “lively grammars” can become apparent when obtaining platform data. Social media data is either retrieved through the extraction of data from media interfaces – which is often called “scraping” (Marres and Weltevrede, 2013) – or through APIs. Whilst the former requires scraping devices or software which needs to be adjusted to the data formats of the respective medium, the latter allows for direct calls to the associated database. Most APIs come with extensive developer documentation, detailing the query formats and limits regarding which data can be accessed in what quantities by whom, and at what cost. API rules are thus central element of the data infrastructures of platforms to manage relations with various stakeholders, developers, clients and data industries, as became visible in the case of Twitter limiting data access to paying partners and policing developers (Puschmann and Burgess, 2014) or Instagram disabling the development of alternative clients (Gerlitz and Rieder, 2017).

The liveliness of platform data does not mean we consider digital data as entirely fluent and adaptable. Digital data remains largely pre-structured in form by the various media devices involved in its creation or translation, and thus may come with a second characteristic: a methodological *bias* (Marres and Gerlitz, 2015). In this context, we do not mean bias simply in the familiar senses of statistical bias or social prejudice, but in a broader sense signalled in Harold Innis’s pioneering studies of communication systems: mediating features which foreground certain aspects of a situation

at the expense of others (Innis, 2008). Whilst some of these biases may be fairly explicit – such as privileging positive affect on Facebook – others are more nuanced and difficult to detect. As Marres and Gerlitz (2015: 2) comment: “when doing network analysis with Facebook, is it really the researcher that here ‘decides’ to use this method, or is this decision rather informed by the object of study with its associated tools and metrics?”. Thus, we must consider the organising capacities and grammars of data infrastructures seriously, without take them as fixed and a priori. Rather we should study how they function in practice, how they are used and adapted and the meaning-making practices of researchers, users and external developers around them – operating between various orders of inscription and interpretive multivalence. For researchers and others working with platform data, the infrastructures implicated in their creation become a matter of concern as a result of these “lively grammars”.

Data infrastructures and their publics

Regimes of measurement, metrification and data collection give rise to cultures of auditing and accountability (Strathern, 2000) as well as the assembly of “data publics” (Ruppert, 2015) with their own interests, capacities and resources. Digital technologies and networks can contribute to the multiplication of these publics. In the case of open data, information generated by institutions may find new publics amongst civic hackers, app developers and data journalists. In the case of social media data, data is used not only by the platform itself but by app developers, data marketers, political campaigns, startups and researchers.

Precisely because data infrastructures are both created with specific purposes in mind yet also multivalent, the relation between data infrastructures and their publics becomes very important – to the extent that the two can be mutually articulating. Following Ruppert (2015), data publics are constituted by dynamic, heterogeneous arrangements of actors mobilised around data infrastructures, sometimes figuring as part of them, sometimes emerging as their effect. Data publics are thus neither subjects subdued by the logics inscribed in data and associated platforms (Birchall, 2015) nor are they sovereign agents empowered by such data infrastructures and computational technologies (Cohen et al., 2011). Instead, as the examples below will show, we can envisage data publics as coming into being around data infrastructures through their activities with data.

Dealing with data publics requires taking into account their specific objectives, needs and capacities. Journalists work with data to find newsworthy stories. Campaigners enlist data to influence policy makers.

Media scholars use data to study the affordances of online platforms. These publics devise methods and tactics to align data with their own interests, concerns, and ways of knowing. Sometimes efforts to achieve alignment will fail. This can constitute an opportunity for data publics to engage with data infrastructures and to attempt to reshape them. How and under what conditions might they succeed in either intervening or inventively aligning data infrastructures with their interests and concerns?

Previous work on “statactivism” explores the creative strategies deployed by different publics to align statistical data with their concerns. In these contexts, the performative capacities of statistics are mobilised in the service of goals that lie on a spectrum between contesting and criticising particular states of affairs (such as existing governance, economic and work regimes) whilst making visible, affirming or legitimising new entities and categories in the service of social and political activism (Bruno et al., 2014). While often conceived as an instrument of governmentality and power, statactivism seeks to exploit the multivalent character of statistical data infrastructures through a series of tactics devised to align them with different visions and objectives.

Statactivism researchers have for instance studied the CompStat performance system started in New York City to reduce crime and achieve other policing goals, which has subsequently been adopted around the world (Bruno et al., 2014; Didier, 2018). CompStat includes leadership meetings aiming to manage police work around principles of “accurate and timely information”, “rapid deployment of resources”, “effective tactics” and “relentless follow-up” (Police Executive Research Forum, 2013: 2). Commentators and critics noted that the focus on crime statistics also shaped police behaviour and the way that crimes were recorded – a phenomenon which researchers describe as the “reactivity” of practices of quantification (Espeland and Sauder, 2007).

Activists and journalists claimed that police were gaming numbers. As one character from the TV series *The Wire* puts it: “Making robberies into larcenies. Making rapes disappear. You juke the stats, and majors become colonels.” Rather than taking crime statistics at face value police officers may be incentivised to develop a cynicism or pragmatism about how numbers are used.

Activists and researchers have sought to align official data with their own purposes in order to analyse whether the CompStat system contributes to discriminatory policing practices. By mobilising and analysing data from “UF-250” forms they have argued that there was a sharp rise in “stop and frisk” practices (which were allegedly used as measures of productivity

in CompStat) which were disproportionately targeting minority groups, thus “bend[ing] the institutional use of the information to show its inner contradictions” (Didier, 2018). The same data infrastructure was inventively repurposed by journalists in order to align with different sets of concerns, shifting the emphasis from identifying and reducing crime to identifying and reducing discriminatory policing (Figure 3).

Other data journalism projects investigate and challenge official data infrastructures, including through inventive strategies of reverse-engineering (Espeland, 2016). Reporters sought to investigate methodological biases in double-voter detection systems in the US.⁶ Their focus was Interstate Crosscheck, a software program which addresses potential voter fraud by detecting double voters and removing them from the voting registration lists. The investigation sought to utilise operations such as sorting, ranking, counting and cross-tabulation of lists in order to understand the methods used to classify people as potential double voters. Journalists found that the program was suggesting potential cases of voter fraud on the basis of first name and last name matches only. Certain minorities were disproportionately threatened with having their names removed from voter rolls, as they were found to be more likely to have common surnames. Reporters worked with legal and advocacy groups to contest the methodologies employed in voting fraud detection. Such tactics of investigating the politics and biases of algorithms has also been described as “algorithmic accountability reporting” (Diakopoulos, 2015). Central to such tactics is an understanding of reactivity biases produced in the situated interplay of the different components of the data infrastructure.

Another inventive response to mal-alignments between data infrastructures and the interests of their publics is not just to appropriate and repurpose them, but to establish different data collection mechanisms. European journalists have developed their own collaborative infrastructures for counting migrant deaths.⁷ While official data collection infrastructures are configured to record migrants who enter EU members states, they are not set up to systematically record cases of migrants who die on the way to the EU. Alternative data collection practices can be viewed as a way to translate anecdotal and disconnected incidents into a more comprehensive picture to bolster political support and policy change (Pécoud, 2016). To this end journalists have set up their own migrant death count operations by aggregating cases of migrant deaths from news media coverage and NGO lists. To achieve an alignment between the analytical capacities of such lists and their own purposes journalists resorted to cleaning, structuring and verifying data to make it amenable to analysis and mapping. Establishing new

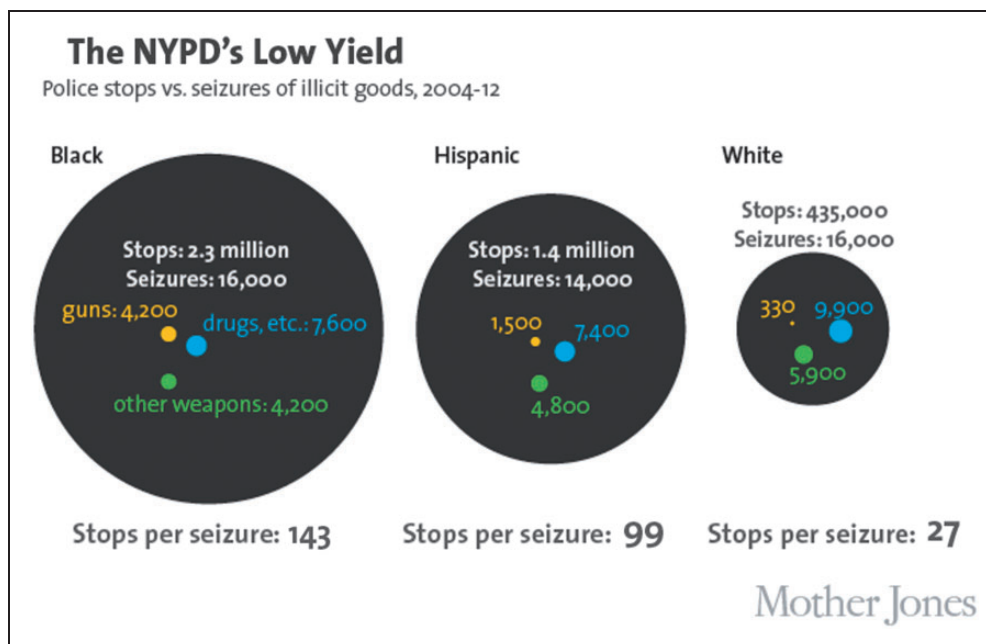


Figure 3. “Are the NYPD’s Stop-and-Frisks Violating the Constitution?”. Chart from Mother Jones, 29 April 2013. <http://www.motherjones.com/politics/2013/04/new-york-nypd-stop-frisk-lawsuit-trial-charts/> (Source: Center for Constitutional Rights).

counting operations in response to the absence or inadequacy of official data has been used to raise awareness around the lack of adequate mechanisms to count homicides by law enforcement officers in the US, killings by US drone strikes and civilian deaths in armed conflicts (Gray et al., 2016).

Researchers also deploy inventive strategies to bring data infrastructures into alignment with their interests, or to exploit mal-alignments. Digital sociologists and digital methods researchers aim to repurpose digital devices and online platforms for social and cultural research (Rogers, 2013; Marres, 2017). This entails a shift from the analytical functions built into platforms to “critical analytics” in order to draw attention to their mediating capacities (Rogers, 2018). For example, researchers use data from edit histories and talk pages on Wikipedia in order to map controversies (Borra et al., 2014; Weltevrede and Borra, 2016). These features of Wikipedia were originally intended to coordinate the improvement of articles, foster consensus and revert spam. Researchers used data generated through these interface features with a different interest in mind: to identify which elements of a page appeared most controversial according to the frequency and character of edits (Figure 4). Wikipedia was thus transformed from article-making to controversy-mapping device. What connects these examples is a sensitivity towards the organisation of data infrastructures and the emergence of infrastructural literacies to re-imagine and re-configure them to align with different interests.

Reassembling data infrastructures

Throughout this article we have argued for an expansion of the concept of data literacy to include not just competencies in reading and working with datasets but also the ability to account for, inventively respond to and intervene around the socio-technical infrastructures involved in the creation, extraction and analysis of data. After noting the rise of conceptions of data literacy focusing on “data as a resource”, we looked at several examples of disassembling data infrastructures, including how to account for their methodological inscriptions, biases and grammars of action. In particular, we looked at when and how these data infrastructures may become a “matter of concern” for their various publics as a result of mal-alignments with their interests. Finally, we looked at the relationships between data infrastructures and their publics, highlighting strategies which are deployed for inventively re-aligning them or creating alternative infrastructures for different objectives, drawing on several examples from previous and ongoing research around data journalism, data activism and digital methods.

Throughout the paper we have sought to emphasise a conception of data infrastructures as distributed accomplishments, constituted by an evolving set of relationships between people and devices, software and standards, words and instruments. Data infrastructures articulate and project social worlds – or “data worlds”

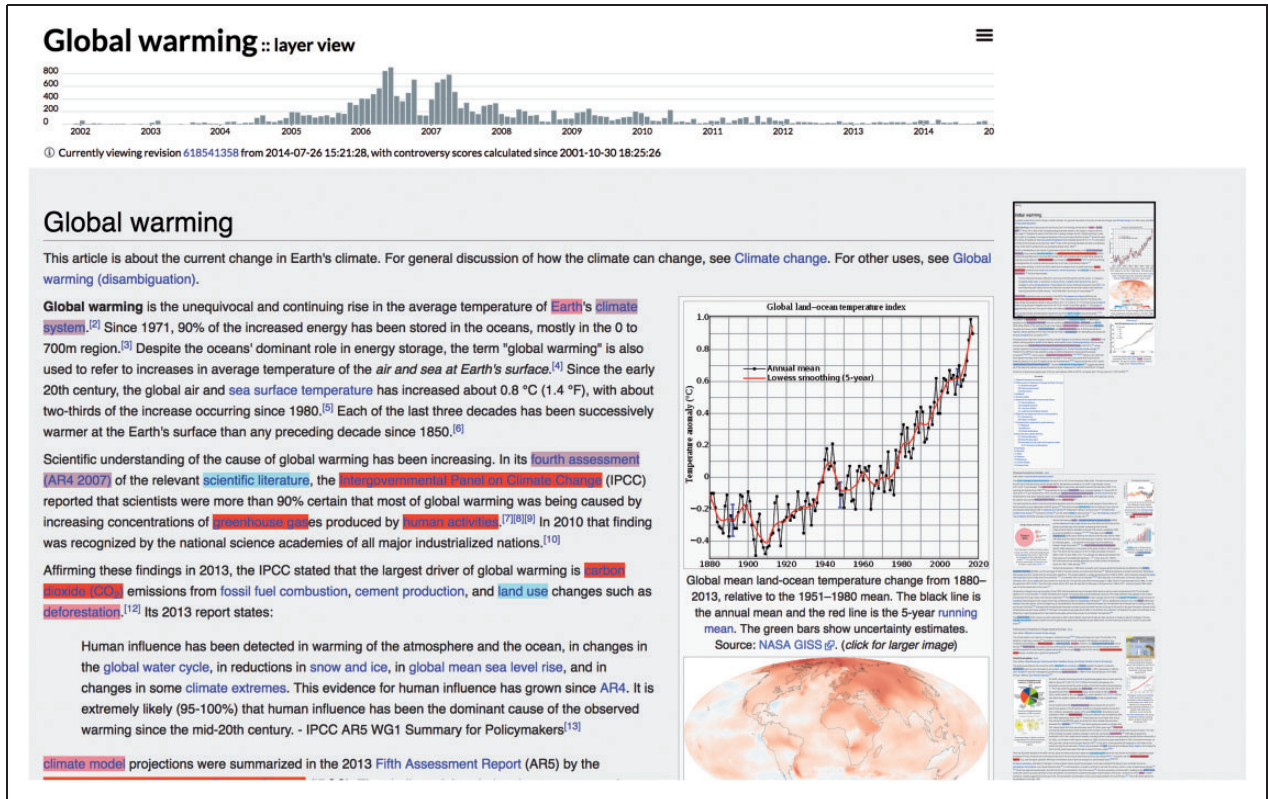


Figure 4. Screenshot of Contropedia project showing controversial elements of the “Global warming” page on Wikipedia. <http://contropedia.net/>.

(Gray, 2018) – which afford their own ways of knowing and possibilities for action. While many previous conceptions of data literacy focus on the effective utilisation of the by-products of these infrastructures as resources for knowing and representing the world, we propose that literacy initiatives should place greater emphasis on developing critical scrutiny, reflexivity, inventiveness and “infrastructural imagination” (Bowker, 2014) with respect to the socio-technical arrangements involved in the making of data. Data infrastructural literacies should cultivate the capacities for reimagining and remaking these data worlds, not just inhabiting them or harvesting their fruits. In this regard, researchers may play a role not only by developing methods and capacities for inventively assembling and reconfiguring data infrastructures to provide different kinds of perspectives but also exploring how they might support experiments in participation and interactivity (Marres, 2017).

John Durham Peters posits the advent of an infrastructural turn – which he calls “infrastructuralism” – in the humanities and social sciences, partly precipitated by the rise of what we might call “infra-structure talk” in politics and public life in the latter part of the 20th century as well as through the work of scholars such as Bowker and Star (Bowker and Star,

1999; Peters, 2015). Drawing on this line of thought, what might be the consequences of a “data infrastructural turn” in relation to data literacy and beyond? Drawing attention to the politics and making of data and data infrastructures could open up new sites of contestation and controversy as well as creating opportunities for new forms of mobilisation, intervention and activism around what they account for and how. This includes not only activist, journalist, professional and research actors that we have discussed in this article but also broader public debates around digital infrastructures, especially following concerns about “fake news”, algorithmic manipulation and bots (Bounegru et al., 2018). Gaining a sense of the diversity of actors involved in the production of digital data (and their interests, which may not align with the providers of infrastructures that they use) is crucial when assessing not only the representational capacities of digital data but also its performative character and role in shaping collective life.

What might be done to support the development of data infrastructure literacy? In the vignettes above, we have examined cases where digital data infrastructures have become a “matter of concern” for various publics. The study of such cases of malalignment, controversy, contestation, breakdown and inventive repurposing

suggests avenues for further inquiry, involvement and experimentation around data infrastructures. In the context of literacy initiatives in universities, schools and training programmes, this might include teaching about data infrastructures as relations rather than simply about datasets as resources. The development of technical and statistical skills may be complemented with field trips, infrastructure ethnography, projects, readings and experiments in participation in order to highlight the various arrangements implicated in the making and social life of data. Policy makers, public institutions, civil society organisations and others may also to take steps to consider their own data infrastructures not only as the means to generate analytical or information resources but as sites of more substantive participation and deliberation about the ways of relating, seeing, doing and being that they engender (Gray et al., 2016). Data initiatives may thus help to support not only the reproduction and translation of what Jasanoff (2017) refers to as “modes of authorised seeing” but also critical reflection on their composition and public debate about possible alternatives (through data or by other means).

Data literacies can serve not only to increase the use and uptake of data but also to multiply the publics who are able to understand and shape infrastructures through which it is created – including exploring infrastructural alternatives to prominent forms such as the platform (Helmond, 2015). Just as Latour (2007: 247) proposes that sociologists should engage in the “reassembling of the collective” through their research, and hence supporters of data infrastructure literacy might make it their task to broaden capacities to reassemble data infrastructures, increasing the visibility of the collectives involved in them and making space for different ways of seeing, knowing and organising the world with (and without) data.

Acknowledgements

We are grateful to the Media of Calculation Research Group and the Digital Methods Initiative at the University of Amsterdam for the opportunities to discuss and develop this work – as well as to participants at the Digital Methods Mini-Conference 2016 and 4S/EASST in Barcelona. We also benefited immensely from comments and reflections on the paper from Bruno Latour and other participants of the atelier d’écriture at the médialab at Sciences Po in Paris during a visit in Spring 2016. Thanks are also due to the editors and three anonymous reviewers at *Big Data & Society* for their time and valuable suggestions.

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

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research for this article was partly funded by the collaborative research centre “Media of Cooperation” (DFG SFB 1187) at the University of Siegen.

Notes

1. See Shields (2004), Carlson et al. (2011), Hogenboom et al. (2011), Calzada and Marzal (2013), Twidale et al. (2013), Martin (2014), Bhargava and D’Ignazio (2015), MacMillan (2015), Koltay (2015a, 2015b), Herzog (2015) and Frank et al. (2016).
2. It is perhaps worth noting that from an ethnomethodological perspective, datasets may indeed be considered to be “raw” in different social settings in contrast to other stages of data management, processing, filtering and imaging. For example, researchers and technicians often talk of “raw data” from cameras, sensors, spectrometers, scanners, servers or surveys to distinguish this from data which have been checked and transformed in preparation for subsequent analytical work. However, we ought not to read these practical distinctions of rawness in terms of more fundamental epistemological claims – just as Wittgenstein (2009: 83–85) cautions against meta-physical interpretations of technical talk of the “different possible states” of a machine.
3. See <https://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=4>. In early meetings and discussions, there were debates about which areas deserved categories of their own, and it was noted that some areas which could be deemed important would either have no category, or would appear in multiple categories – such as environmental spending, space technology and water usage. See <https://unstats.un.org/unsd/statcom/doc79/1979-510-GovFunction-E.pdf>
4. See, for example, discussions at the UK Office of National Statistics about strategies for estimating unmeasured output in relation to the COFOG standard: <https://www.ons.gov.uk/file?uri=/economy/economic-outputandproductivity/productivitymeasures/methodologies/productivityrelatedarticlesandpublications/unmeasuredoutputtcm77263732.pdf>
5. See, for example, <https://github.com/openspending/fiscal-data-package>, <http://standard.open-contracting.org/latest/en/> and <http://iatistandard.org/>
6. <http://projects.aljazeera.com/2014/double-voters/>
7. <http://www.themigrantsfiles.com/>

ORCID iD

Jonathan Gray  <http://orcid.org/0000-0001-6668-5899>
Liliana Bounegru  <http://orcid.org/0000-0003-0198-5158>

References

- Agre PE (1994) Surveillance and capture – Two models of privacy. *The Information Society* 10(2): 101–127.
- Agre PE (1997) Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In: Bowker G, Star SL,

- Turner B, et al. (eds) *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*. Mahwah, NJ: Erlbaum, pp. 130–157.
- Baym NK (2014) Data not seen: The uses and shortcomings of social media metrics. *First Monday* 18(10): 1–15.
- Bhargava R and D'Ignazio C (2015) Designing tools and activities for data literacy learners. Available at: <https://dam-prod.media.mit.edu/x/2016/10/20/Designing-Tools-and-Activities-for-Data-Literacy-Learners.pdf> (accessed 14 June 2018).
- Bijker WE, Hughes TP and Pinch TF (eds) (1987) *The Social Construction of Technological Systems. New Directions in the Sociology and History of Technology*. Cambridge, MA: MIT Press.
- Birchall C (2015) 'Data.gov-in-a-box': Delimiting transparency. *European Journal of Social Theory* 18(2): 185–202.
- Bodle R (2010) Assessing social network sites as international platforms. *The Journal of International Communication* 16(2): 9–24. DOI: 10.1080/13216597.2010.9674765.
- Borra E, Weltevrede E, Ciuccarelli P, et al. (2014) Contropedia – The analysis and visualization of controversies in Wikipedia articles. In: *Proceedings of the international symposium on open collaboration*, p. 34. New York, NY: ACM.
- Bowker GC (2005) *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.
- Bowker GC (2014) The infrastructural imagination. In: Mongili A and Pellegrino G (eds) *Information Infrastructure(s): Boundaries, Ecologies, Multiplicity*. Cambridge, UK: Cambridge Scholars Publishing, pp. xii–xiii.
- Bowker GC, Baker K, Millerand F, et al. (2009) 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.
- Bowker GC and Star SL (1999) *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.
- Bounegru L, Gray J, Venturini T, et al. (eds) (2018) *A Field Guide to "Fake News" and Other Information Disorders*. Amsterdam: Public Data Lab. Available at: <http://fake.news.publicdatalab.org/> (accessed 15 June 2018).
- Boyd D and Crawford K (2012) Critical questions for big data. *Information, Communication & Society*. 15(5): 662–679.
- Braman S (2009) *Change of State: Information, Policy, and Power*. Cambridge, MA: MIT Press.
- Bruno I and Didier E (2013) *Benchmarking. L'État sous pression statistique*. Paris: La Découverte.
- Bruno I, Didier E and Vitale T (2014) Statactivism: Forms of action between disclosure and affirmation. *Partecipazione e Conflitto* 7(2): 198–220.
- Cabinet Office (2015) The national information infrastructure implementation document. Available at: www.gov.uk/government/publications/national-information-infrastructure (accessed 14 June 2018).
- Calzada PJ and Marzal MÁ (2013) Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri* 63(2): 123–134.
- Carlson J, Fosmire M, Miller CC, et al. (2011) Determining data information literacy needs: A study of students and research faculty. *Portal: Libraries and the Academy* 11(2): 629–657.
- Cohen S, Hamilton JT and Turner F (2011) Computational journalism. *Communications of the ACM* 54(10): 66–71.
- Dalton CM, Taylor L and Thatcher J (2016) Critical data studies: A dialog on data and space. *Big Data & Society* 3(1): 1–9.
- Data Revolution Group (2014) *A World That Counts: Mobilising the Data Revolution for Sustainable Development*. New York: United Nations Secretary-General's Independent Expert Advisory Group. Available at: www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf (accessed 15 June 2018).
- Desrosières A (2002) *The Politics of Large Numbers: A History of Statistical Reasoning* (Trans. Naish C). Cambridge, MA: Harvard University Press.
- Diakopoulos N (2015) Algorithmic accountability. *Digital Journalism* 3(3): 398–415.
- Didier E (2018) Globalization of quantitative policing: Between management and statactivism. *Annual Review of Sociology* 44(1). Available at: <https://www.annualreviews.org/doi/10.1146/annurev-soc-060116-053308>.
- Edwards PN, Bowker GC, Jackson SJ, et al. (2009) Introduction: An agenda for infrastructure studies. *Journal of the Association for Information Systems* 10(5): 364–374.
- Espeland W (2016) Reverse engineering and emotional attachments as mechanisms mediating the effects of quantification. *Historical Social Research* 41(2): 280–304.
- Espeland W and Sauder M (2007) Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology* 113(1): 1–40.
- Espeland WN and Stevens ML (1998) Commensuration as a social process. *Annual Review of Sociology* 24(1): 313–343.
- Frank M, Walker J, Attard J, et al. (2016) Data literacy – What is it and how can we make it happen? *The Journal of Community Informatics* 12(3): 4–8.
- G8 (2013) G8 open data charter and technical annex. Cabinet Office. Available at: www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex (accessed 15 June 2018).
- Gerlitz C and Rieder B (2017) Tweets are not created equal. Investigating Twitter's client ecosystem. *International Journal of Communication* 11: 1–27.
- Gillespie T (2010) The politics of 'platforms'. *New Media & Society* 12(3): 347–364.
- Gitelman L (2013) *Raw Data Is an Oxymoron*. Cambridge, MA: MIT Press.
- Gray J (2014) Towards a genealogy of open data. In: *European consortium for political research (ECPR) general conference 2014*. Glasgow: University of Glasgow, pp. 1–38. Available at: <http://dx.doi.org/10.2139/ssrn.2605828> (accessed 15 June 2018).
- Gray J (2018) Three aspects of data worlds. *Krisis: Journal for Contemporary Philosophy*. Available at: <http://krisis.eu/three-aspects-of-data-worlds/> (accessed 15 June 2018).

- Gray J, Bounegru L and Chambers L (eds) (2012) *The Data Journalism Handbook*. Sebastopol, CA: O'Reilly Media.
- Gray J, Lämmerhirt D and Bounegru L (2016). Changing what counts. CIVICUS and Open Knowledge. Available at: <http://dx.doi.org/10.2139/ssrn.2742871> (accessed 15 June 2018).
- Hacking I (1990) *The Taming of Chance*. Cambridge, UK: Cambridge University Press.
- Harvard Business Review (2012) Data is useless without the skills to analyze it. Available at: <https://hbr.org/2012/09/data-is-useless-without-the-skills> (accessed 15 June 2018).
- Helmond A (2015) The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society* 1(2): 1–11. DOI: 10.1177/2056305115603080.
- Herzog D (2015) *Data Literacy: A User's Guide*. London: SAGE Publications.
- Hogenboom K, Phillips CMH, Hensley MK, et al. (2011) Show me the data! Partnering with instructors to teach data literacy. In: *Declaration of interdependence: The proceedings of the ACRL 2011 conference*, Philadelphia, PA, pp. 410–417.
- Iliadis A and Russo F (2016) Critical data studies: An introduction. *Big Data & Society* 3(2): 1–7.
- Information Infrastructure Taskforce (1993) *The National Information Infrastructure: Agenda for Action*. Washington, DC: The White House. Available at: <https://archive.org/details/04Kahle000911> (accessed 15 June 2018).
- Innis HA (2008) *The Bias of Communication*. Toronto, Canada: University of Toronto Press.
- Jasanoff S (2017) Virtual, visible, and actionable: Data assemblages and the sightlines of justice. *Big Data & Society* 4(2): 1–15.
- Jasanoff S and Kim SH (eds) (2015) *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago: University of Chicago Press.
- Kahin B (1995) The internet and the national information infrastructure. In: Kahin B and Keller J (eds) *Public Access to the Internet*. Cambridge, MA: MIT Press, pp. 2–23.
- Koltay T (2015a) Data literacy for researchers and data librarians. *Journal of Librarianship and Information Science* 49(1): 3–14.
- Koltay T (2015b) Data literacy: In search of a name and identity. *Journal of Documentation* 71(2): 401–415.
- Latour B (2004) Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry* 30(2): 225–248.
- Latour B (2007) *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.
- Letouzé E (2016) Should 'data literacy' be promoted? UN World Data Forum. Available at: <https://undataforum.org/WorldDataForum/should-data-literacy-be-promoted/> (accessed 15 June 2018).
- Letouzé E, Bhargava R, Deahl E, et al. (2015) *Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data*. New York, NY: Data Pop Alliance. Available at: <http://datapopalliance.org/wp-content/uploads/2015/11/Beyond-Data-Literacy-2015.pdf> (accessed 15 June 2018).
- Lury C and Wakeford N (eds) (2012) *Inventive Methods: The Happening of the Social*. New York: Routledge.
- McKinsey Global Institute (2011) Big data: The next frontier for innovation, competition, and productivity. Available at: www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed 15 June 2018).
- MacMillan D (2015) Developing data literacy competencies to enhance faculty collaborations. *Liber Quarterly* 24(3): 140–160.
- Marres N (2017) *Digital Sociology: The Reinvention of Social Research*. London: Polity Press.
- Marres N and Gerlitz C (2015) Interface methods: Renegotiating relations between digital social research, STS and sociology. *The Sociological Review* 64: 21–46.
- Marres N and Weltevrede E (2013) Scraping the social? Issues in real-time social research. *Journal of Cultural Economy* 6(3): 313–335.
- Martin E (2014) What is data literacy? *Journal of eScience Librarianship* 3(1): 1–2.
- Miller P (2001) Governing by numbers: Why calculative practices matter. *Social Research* 68(2): 379–396.
- Paßmann J and Gerlitz C (2014) 'Good' platform-political reasons for 'bad' platform-data. Zur Sozio-technischen Geschichte der Plattformaktivitäten Fav, Retweet und Like. *Mediale Kontrolle unter Beobachtung* 3(1): 1–40.
- Pécoud A (2016) The politics of counting migrant deaths in the Mediterranean. *Political Power & Social Theory: The Blogpages*. Available at: www.politicalpowerandsocialtheory.com/the-refugee-crisis (accessed 15 June 2018).
- Peters JD (2015) *The Marvellous Clouds: Toward a Philosophy of Elemental Media*. Chicago: University of Chicago Press.
- Police Executive Research Forum (2013) *Compstat: Its Origins, Evolution, and Future in Law Enforcement Agencies*. Washington, DC: Bureau of Justice Assistance and Police Executive Research Forum. Available at: <https://www.bja.gov/publications/perf-compstat.pdf>.
- Porter TM (1986) *The Rise of Statistical Thinking, 1820–1900*. Princeton, NJ: Princeton University Press.
- Porter TM (1996) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
- Puschmann C and Burgess J (2014) The politics of Twitter data. In: Weller K, Bruns A, Burgess J, et al. (eds) *Twitter and Society*. New York: Peter Lang, pp. 43–54.
- Rieder B and Sire G (2013) Conflicts of interest and incentives to bias: A microeconomic critique of Google's tangled position on the web. *New Media & Society* 16(2): 195–211.
- Rogers R (2013) *Digital Methods*. Cambridge, MA: MIT Press.
- Rogers R (2018) Otherwise engaged: Social media from vanity metrics to critical analytics. *International Journal of Communication* 12: 450–472.
- Ruppert E (2015) Doing the transparent state: Open government data as performance indicators. In: Rottenburg R, Merry SE, Park SJ, et al. (eds) *A World of Indicators: The Making of Governmental Knowledge through Quantification*. Cambridge: Cambridge University Press, pp. 1–18.

- Shields M (2004) Information literacy, statistical literacy, data literacy. *IASSIST Quarterly* 28(Summer/Fall): 6.
- Star SL (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3): 377–391.
- Star SL and Ruhleder K (1996) Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7(1): 111–134.
- Strathern M (2000) *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy*. New York: Routledge.
- Twidale MB, Blake C and Gant JP (2013) Towards a data literate citizenry. In: *iConference 2013*. Urbana, IL: University of Illinois, pp. 247–257. Available at: <http://hdl.handle.net/2142/42563>.
- United Nations (2012) The future we want. Available at: <https://sustainabledevelopment.un.org/content/documents/733FutureWeWant.pdf> (accessed 15 June 2018).
- Venture Beat (2014) Democratizing data: Why data literacy will be the most important new skill of the 21st century. Available at: <https://venturebeat.com/2014/11/07/democratizing-data-why-data-literacy-will-be-the-most-important-new-skill-of-the-21st-century/> (accessed 15 June 2018).
- Verran H (2015) Enumerated entities in public policy and governance. In: Davis E and Davis PJ (eds) *Mathematics, Substance and Surmise*. New York: Springer, pp. 365–379.
- Weber M (2011) In: Shils EA and Finch HA (eds), *Methodology of Social Sciences*. New Brunswick, NJ: Transaction Publishers.
- Weltevrede E and Borra E (2016) Platform affordances and data practices: The value of dispute on Wikipedia. *Big Data & Society* 3(1): 1–16.
- Wittgenstein L (2009) *Philosophical Investigations* (Trans. Anscombe GEM, Hacker PMS and Schulte J). Chichester, UK: Blackwell Publishing.