What difference do data make? Data management and social change

What Difference Do Data Make? Data Management and Social Change

Abstract:
This article expands on emergent data activism literature to draw distinctions between different types of data management practices undertaken by groups of data activists. We build upon extant literature on data management infrastructure, which primarily discusses how these practices manifest in scientific and institutional research settings, to analyse how data management infrastructure is often crucial to social movements that rely on data to surface political issues. We offer three case studies that illuminate the data management strategies of these groups. Each group discussed in the case studies is devoted to representing a contentious political issue through data, but their data management practices differ in meaningful ways. The project Making Sense produces their own data on pollution in Kosovo. Fatal Encounters collects 'missing data' on police homicides in the United States. The Environmental Data Governance Initiative hopes to keep vulnerable U.S. data on climate change and environmental injustices in the public domain. In analyzing our three case studies, the authors surface how temporal dimensions, geographic scale, and sociotechnical politics influence their differing data management strategies.

Keywords: statactivism; data activism; data infrastructure; environmental data, civic data, police data

Acknowledgements: Special thanks to volunteers from the Environmental Data and Governance Initiative, Fatal Encounters, and Making Sense.

Introduction

What difference do data make? Over the last decade, scholars have struggled to understand the rapid informationalization of society, as the data that drive this trend grow larger by the moment, are increasingly commodified for private profit, and are used to control populations through both finely targeted advertisements and surveillance architectures that link corporate and government data streams. A rigorous set of scholarship now addresses this issue of data quantity by focusing on the sociotechnical dimensions of data infrastructures – how 'big data' is collected, categorised, analysed, stored, controlled, and accessed, and how these practices produce widely uneven distributions of political and economic power (Eubanks, 2018; Bates et al., 2016; Ribes and Jackson, 2013).

Simultaneously, a growing set of literature examines social and activist movements that organise in response to data collection by corporations and the state (Milan and Van der Velden, 2016; Liboiron, 2015; Currie et al, 2016; Bruno et al., 2014; Dalton and Stallmann, 2018; Dalton and Thatcher, 2014; Gieseking 2018). In some cases, these political activists achieve some agency by avoiding data capture through the use of encryption devices or by designing alternative, non-commercial and collectively owned platforms. Other activists respond by creating their own representations and framings of an issue through community data collection, visualisation, and analysis. An example of the latter are data activist projects that collect and publish data on policing in the United States. Here, the best data generated from the Federal Bureau of Investigation (FBI) consistently underestimate the body count of those killed by police, so activists and journalists approached the issue through various methodologies to improve accuracy.

Whether working with data to make an issue more visible or to contest its 'official' representations, the results are often only a small part of these activists' actual work. Behind the visualisations or public-facing databases are a suite of data management infrastructures and organisational norms that form a considerable part of activists' mundane practice. For activists who rely on data to represent and politicise an issue, the acts of standardising data, anonymising them and making them robust over time are essential group strategies. In many cases groups will be concerned to keep track of who contributes and accesses the data, to guarantee they remain safe from tampering, and to ensure their longevity in electronic storage media. Yet these broader infrastructural practices of data activists are an often-underappreciated area of scholarly attention in literature on these projects.

This article's central focus is on the data management practices of data activists; it argues for more research devoted to the infrastructural needs of social movements that rely on data to politicise issues. To make this argument we begin by defining *data activism*, drawing on a rich and growing set of literature on the topic; through this research we make some conceptual distinctions between types of data activists that we draw on for our case study selection. In the second section we examine literature on data management infrastructure, most of which we find focuses on large institutional projects generally within government or research universities; less research is devoted to projects arising from the level of grassroots social movements.

In the third section of the paper we offer three case studies that provide a window into the data management strategies of groups devoted to representing a contentious political issue through data, either by producing their own data, collecting 'missing data', or keeping vulnerable data in the public domain. Methodologically, evidence for our three case studies is drawn primarily from personal interviews with a member from each organisation, selected because he or she played a relevant role in some aspect of these projects' data infrastructural work, whether selection, design, implementation, governance, or use, or a combination of several of these activities. We combined these interviews with news reports about issues of concern to these activists, media coverage of the activist projects, and several primary documents in the form of procedural reports and press releases produced by these projects. Finally, in our conclusion we discuss some of the implications revealed in the case study analysis. In particular, we find lessons to be learned regarding the role that data management plays in shaping the governance structures of data activist projects, as well as the need to attend to the political-economic dimensions of information infrastructure itself.

The Nuances of Data Activism

In the past decade scholars within Information Studies, Geography, and Science and Technology Studies (STS) have sought new terms to describe data practices as novel forms of activism and resistance. These scholars, and the activists they analyse, understand that data are not benign, neutral information resources underpinning our representations of the world but are a major source of political power and social critique. While in some instances the practices that fall under the term 'data activism' may not be new, the literature largely positions data activism as a response to several forces specific to the past several decades: the use of indexes and benchmarking indicators by the modern nation state, numbers that are so routine that they produce reality “through an irreversible ratchet-effect” (Desrosieres, 2008, p. 12); the capturing and monetisation of online user data and the consequent rise of finely tuned microtargeting that sways commercial consumption and political votes alike (Srnicek, 2016; Tufekci, 2014a); the often covert exchange of data between powerful commercial platforms and government surveillance apparatuses; and the fact that online surveys, digital mapping, and digital sensors are now widely accessible and relatively affordable to lay publics.

Given this backdrop, literature on data activism documents examples of citizens who use data to address the vast informational asymmetry in democratic societies. Scholars of citizen science, for instance, offer rich case studies that expose the power imbalance between scientists and laypeople in making scientific claims about climate change, air pollution, and the design of urban space (Irwin, 2001). Scientists have at their disposal highly technical “inscription devices,” such as microscopes and lab protocols, that make their objects of study stable, authoritative sources of information (Latour, 1987). The general public, in contrast, often has very few and highly unreliable devices with which to make claims (Priest, 2013). Nevertheless, as these scholars have shown, lay publics can use their own data gathering techniques to intervene in scientific debates and enact more democratic forms of environmental policy making as a result. Citizens have used data, for instance, to contest chemical weapons disposal in the United States (Futrell, 2003), map the exposure to toxins and pollutants in buildings (Murphy, 2006), and take lo-fi aerial photographs to document evidence of sewage flow into protected sites (Wylie et al., 2014).

Bruno et al. (2014) use the term “statactivism” to pinpoint more precisely activists who use “numbers, measurements, and indicators as a means of denunciation and criticism” (p. 199). Statactivism acknowledges “the double role of statistics in representing as well as criticizing reality” and of revealing the political and negotiated dimensions of statistical work (p. 200). The authors describe how statactivists can use statistics to make a community, social category, or cause more visible; they illustrate the force of a new social category, for instance, when citizens championed the need for the government to recognize a new class, the ‘*intellos precaires*,’ or precarious workers, who have higher education degrees but no reliable long-term employment. Once established as a stable category, individuals who fall in this group can start putting forward collective demands. Statactivism can also cast doubt on or rejects official state indicators and benchmarks. Scholars from critical GIS (geographic information systems) similarly call such acts of resistance to institutional and commercial datasets “counter-data actions” (Dalton and Thatcher 2014, n.p.). The concept of counter-data draws from longstanding work in critical GIS (geographic information systems) to create alternative cartographies that privilege the geographic knowledge of individuals, such as indigenous groups or LGBTQ communities, often left out of mainstream political discourse, science, industry, and technological practice (Dalton and Stallmann, 2018; Gieseking 2018).

To widen their scope of analysis beyond statistical or numerical representations, Stefania Milan and Lonneke Van der Velden (2016) use the term “data activism”, which includes making use of visual and qualitative data as well as tactics to *avoid* data capture. What the authors term “pro-active” data activism uses data to create or contest representations of an issue, while “re-active” data activism avoids data collection and surveillance through encryption tools, obfuscation and anonymity. This stark binary can easily fall apart – re-active tactics can also entail very pro-active, creative design strategies that offer alternatives to data extraction platforms – but it does begin to make important analytical distinctions between various types of data activism.

This article focuses its analysis on what Stefan and Van der Velden term the "pro-active" type – those activists who work with data as a tactic to challenge authoritative accounts that are either inadequate, politically vulnerable, or misleading, and who must generally consider some form of data management practices to put forward their case. We want to spotlight the practices of political movements devoted to issue visibility specifically through data collection and maintenance. In these cases, the acts of creating and managing data are not ancillary to movement building but are the adhesion that tie activists together and make their political movement cohere.

We also selected examples that further nuance this literature, by showing how representational work can manifest in at least three ways. The first two cases offer examples of counter-data, though responding to slightly different deficiencies of government accountability. In the first case we describe citizens' response to data that has not been gathered by the state in any comprehensive manner to fully assess a phenomenon. We use an example of data activism that sought to correct and augment the statistical work of the U.S government, which fails to account for all deaths caused by police. For activists opposing police brutality, such as Black Lives Matter, recourse to reliable and accurate statistics on deaths in custody has been a crucial strategy to call for policing reforms. We illustrate how one organization, Fatal Encounters, uses a collective database to produce this missing data and so critique the data that does exist. In our second case study, we examine activism that produced a new dataset to make visible an issue that was being deliberately obscured by their government. In this case, activists collected air quality data around Kosovo and in its capital city, Prishtina; they used the data within media campaigns to force a public debate around a health crisis that the government had largely kept invisible. Our final example focuses on visibilising an issue by means other than critiquing existing institutional data. In our third case study, we look at activism undertaken to keep politically vulnerable data in the public domain. We look at the DataRescue work led by the Environmental Data Governance Initiative (EDGI), comprised of networks of scholar-activists that formed in reaction to Donald Trump's election. EDGI's goal was to archive data created by U.S. federal scientists that documented evidence of climate change and human-induced ecological violence.

Controlling representations entails careful epistemic work. Open, collaborative networks of contributors in particular will want to ensure that the data are reliable and withstand scrutiny in the public sphere. Yet outside of some more detailed descriptions of civic science and sensing projects (Jalbert et al., 2017; Kinchy et al. 2014; Wylie et al. 2014) literature on data activism has rarely examined data management infrastructures of activists with close scrutiny. Literature on data activism typically describes the project of *generating* data to create new statistical representations or to challenge official ones, and it often looks at how these representations circulate; it has widely ignored issues of stewardship and the dynamics of data management among the activists themselves. Moreover, the ways in which various groups practicing data activism think about the politics of the infrastructures they engage with to collect and maintain their data is under-represented in this scholarly work. In the following section we briefly go over scholarship focused on data management infrastructures to clarify some heuristics for examining these practices, before moving on to our case studies.

Particulars of Activist Data Management

When we talk about infrastructures of data management, what, exactly, does this encompass and what does it mean in the particular cases of grassroots activists? Just as literature on data activism has devoted less focus on everyday data management infrastructures, so professional literature on data management often has very little to say about its manifestation in grassroots settings outside of government and scientific contexts. Within archival science and information studies, the concern is almost solely on the design, adoption, and use of data repositories in university, corporate, and government settings (Lauriault et al., 2007; Borgman, 2007; Frost et al., 2015; Borgman 2015).[1] For instance, 'Elements of a Data Management Plan' is a list provided by the Inter-university Consortium for Political and Social Research (ICPSR), which maintains one of the oldest research data archives in the world.[2] The Elements list factors of data maintenance relevant to projects that rely on managing datasets: the need to consider data formatting, back-up, access and sharing rights, ethics and privacy, and quality assurance, among several others. This list provides a standard rubric or heuristic for thinking about those practices that groups involved in data collection and maintenance for evidence collection may need to consider. However, this professional literature is limited because it is concerned with establishing good practices for institutional settings only.

This professional literature also does not offer a framework for analyzing data management practice within broader, sociotechnical relations. Drawing on STS literature, we think of data management not only as a set of best practices but as relational *infrastructures*. Infrastructures in this sense are not a set of connected technological artifacts programmed to be useful but largely unnoticed; instead infrastructures emerge through relations of users and within the context of the other social, institutional, and economic systems that they are necessarily part of (Star and Ruhleder, 1996; Dodge and Kitchin, 2011; Ruppert 2012). This framing helps us analyse infrastructures of data management as a non-linear, non-routinised set of relations between people and technologies working in complex organizational contexts. In these settings, difficulties will inevitably emerge; these may not be resolved by gaining more information or skills, such as a new user who learns to onboard a system, but may encompass more complicated cultural dynamics that arise around access privileges, software selection, privacy concerns, and economic trade-offs, to name just a few. Infrastructures shape these dynamics as much as they are shaped by them. Yet even this STS literature is largely absent of cases that look at data infrastructures at grassroots levels; instead it offers rich descriptions of government censuses, scientific data models, and systems of corporate data capture and dissemination (Bowker and Star, 1999; Edwards, 2010; Kitchin, 2014; Gillespie, 2018). This literature therefore raises a set of questions about the dynamics of these practices particular to activist projects working *outside* of institutions.

Take, for instance, that literature on data management in a research context often assumes colleagues who share some institutional, disciplinary, or technical knowledge. Data activist projects, on the other hand, will coordinate people with widely different backgrounds and skillsets, some of whom may never meet face to face. Many data activist projects will need a structure to manage multiple contributors and low barriers of entry for participation while maintaining data integrity. How this coordination takes shape, whether through gatekeeper and hierarchies of access permissions or by more radically decentralized and participatory methods, also shapes the relations among the participants themselves.

A related question concerns storage, backup, and security: does someone take ultimate responsibility of keeping an electronic dataset secure in storage media so that it

[1] Data management is an interdisciplinary subject that is also widely treated in literature from data and computer sciences, which focuses mostly on technological dimensions.

[2] ICPSR's repository includes 250,000 files of research in the social and behavioural sciences and 21 specialised collections in in education, aging, criminal justice, substance abuse, terrorism, and more. https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/elements.html

is reliable over time, or can this role be federated or even outsourced? Who designs the formatting and description protocols of the dataset? In university or government settings the roles regarding data management tasks may be set by traditional institutional hierarchies (senior researchers on down to postdoc and PhD students), but for grassroots activists in the civic sphere, defining these roles may be ad hoc as the data collecting or maintenance unfolds; it can, for instance, fall to the work of one or a few people with technical and professional skills or take shape through more democratic decision-making procedures.

Finally, we can ask about the politics of the technologies used to store and maintain data. Literature on data management does address the wider political economy of software, particularly by arguing for the economic virtues of open source software over closed licenses or commercial platforms (Frost et al., 2015; Fry et al., 2008; Strasser 2013). In the institutional research context, digital data repositories that are not tethered to expensive licensing contracts can have greater longevity and make their content free to users. Yet data management literature has less to say about the activists' wide use of social media to publicise, galvanise, and organise contentious politics. For grassroots projects with little to no business plan or funding, off-the-shelf "free" platforms may be the best tool to get the job done. Activists often use social media to locate each other and narrate their causes to wide audiences; these platforms boost organizational capacity of people working outside of traditional institutions (Tufekci, 2014b).

Yet in other cases, platforms such as Google or Facebook could pose problems in the long term should the policies of these opaque companies change or if activists want more control over their data, particularly when privacy becomes a concern. Writing in 2011, a year after Facebook went public, Geert Lovink lamented that "Activists organize transnational campaigns online, and Web 2.0 companies profit from the free labor and attention provided by networks of users" (p. 167-168). Awareness of "platform capitalism" has only since grown (Srnicek, 2016). As a result of the surveillance and data capture by these corporate platforms, some activists are now building their own alternative communication and networking systems to mobilise and exchange goods and transactions using open source software, platform cooperatives, mesh networks, and alternative internet protocols. An analysis of data activists and their management tools, therefore, should look at both the use and repurposing of networked media and at alternative technological platforms created to sustain their work. We can ask whether projects build alternative infrastructures that give owners more control over their data, and we can ask to what extent these technologies might shape the data in terms of formats and automated metadata.

In sum, while literature on data management infrastructure falls short in discussing the practices of grassroots projects operating outside of institutional contexts, it focuses our attention on a few dimensions that we can ask of data activist projects:

1. How did the coordination of the data management take shape – that is, how does the data's collection and maintenance become distributed between participants over time?
2. How are the platforms that collect, store and update the data managed over time – by one or few individuals or collectively?
3. How are access and sharing privileges across data infrastructures determined and distributed?
4. Who designs the data formatting and description protocols, and who makes sure these are maintained over time?

5. How do these data infrastructures play a role, in turn, in shaping the relations among participants?
6. What software has been chosen to maintain the data, how was it chosen, and what is its wider political economy – i.e. open source software vs. freemium platforms vs. commercial software that must be licensed?

Though these are no means exhaustive, we draw on these six factors now to analyse our three case studies looking at missing data, vulnerable data, and data created to make an issue newly visible in the public sphere.

Case Study 1: Filling in the Gaps

The accuracy of statistics on policing is one of the most persistent problems facing activists in the Black Lives Matter Movement. After the death of Michael Brown in 2014, widespread protests across the U.S. called for increased accountability and oversight of policing practices. Multiple organizations took up the charge of gathering statistics on the number of people killed by law-enforcement. Similarly, activists in the early 1990s sought the very same data on policing after the Rodney King trial and the acquittal of officers from the Los Angeles Police Department. In 1994, the Attorney General was mandated to collect data on the "use of excessive force," and the Bureau of Justice Statistics (BJS) was to issue an annual report on this (McEwen, 1996). However, the law never required state law agencies to report to BJS. As a result, year after year, the BJS continuously admonished police departments for failing to provide accurate - and in some cases, any - counts of deaths in custody or provide consistent metadata so that trends could be charted over time (Smith and Austin, 2015; Mumola, 2007). Relying only on police departments for this information, the BJS faced a persistent problem of missing data.

As the public sought ways to hold police accountable, grassroots groups and data journalists built public databases of police killings by sourcing materials across a wide array of public records requests, media reports, obituaries, and social media. Many of these databases relied on crowd-work to harvest missing metadata and fill in gaps related to information about the officer and the person who was murdered. Methods of data collection and categorization vary across all projects (Currie et al., 2016). Here we focus on the group Fatal Encounters. The project was volunteer-run and directed by its founder journalist, Brian Burghart, from its inception until 2014, when it started to receive grants and crowdfunding. Until 2017, Burghart received grant funding to direct the project and received help from both upaid and paid volunteers.[3] The website charts deaths by fire, vehicle, stabbing, choking, suffocation, pepper spray, and more. Fatal Encounters also uses Freedom of Information Act requests, which provide data on deaths that may never get reported on or appear in newspaper obituaries (Burghart, 2017b). This citizen data came up with numbers that were in many instances larger than those published by the U.S. Department of Justice.

Organisationally Fatal Encounters is a hierarchy run by Burghart, who delegates access privileges, manages the data's backend, and designed the Google document that formats and standardises all entries. Data contribution, however, is federated: anyone who wants to contribute to the dataset can do so through a Google Form vetted by Burghart.[4] The data then populates a private Google Sheet that feeds into a CSV file that streamlines the metadata and is backed up and maintained by Burghart. To collect

[3] From email correspondence with D. Brian Burghart, June 25, 2017.

[4] Email correspondence with D. Brian Burghart, June 25, 2017.

the data, Fatal Encounters' volunteers draw from a multitude of sources: FOIA requests, public records, police records, media reports, coroner reports, social media submissions, photographs, original reporting, and crowdsourced verification. [5] Fatal Encounters' founder double checks all the reports received against local news stories before publication. Fatal Encounters, however, tends to track and record more data than other groups such as the counts main *The Guardian* and *The Washington Post*). Because Fatal Encounters values exhaustive and comprehensive verification, some cases may stay in the database until such time that the cause of death has been clarified, either through public records, FOIA requests, or updated media reports. This data is finally made public on a Google spreadsheet that is accessible and downloadable, but not editable, online.

In the interest of time and financial resources, the group uses free Google software – Google Forms and Google Sheets – to solicit data from the public and manage the data they collect from FOIA requests, independent investigation and public reports (Burghart, 2017a). Using free software works well for this group, since it creates a very low bar for participating in terms of technological skill.[6] In cases of more politically vulnerable data, however, this practice of using free corporate software may not suffice.

Problem 2: Creating New Data

Kosovo is one of the EU's most polluted countries, and by 2015 the consequences on citizens were becoming apparent through rising rates of cancer, respiratory tract infections and cardiovascular diseases (Making Sense, n.d.; McQuillan, 2015). The government response was to remain silent amidst this health crisis, and its Environmental Protection Agency refused to release air quality data that could stir public outrage. In this context, the country became one of the pilot sites for Making Sense, a citizen sensing project that determined to make the country's environmental problems visible.

Making Sense is a European project involving five partners, with research for policy and action led by faculty at the University of Dundee. One of the project's critical outputs is the Making Sense Toolkit, a collection of resources for communities who want to deploy citizen-led campaigns to capture and share open data about the environment.[7] The Toolkit offers detailed case study reports, documentation of technological requirements, and a sensor onboarding guide, among other documents, and it describes how citizens can come together in open, collaborative settings to set data collection strategies, learn how to use sensors, and coordinate publicity campaigns around their findings.

In Kosovo, Making Sense joined forces with the Peer Educators Network and Science for Change Kosovo Movement, a grassroots collective devoted to breaking the government's silence on pollution. As one of the organisers describes it, the 30 participants, many of them youth, built Making Sense on radical democratic participatory approaches of non-exclusion and semi-horizontal structures where decision-making took place through weekly general assemblies. Participants held a three-day training workshop, then self-selected into groups in charge of either communications or sensing and devising protocols.[8] Three campaigns followed, from April 2014 till June 2017 (Making Sense, n.d.).

[5] Email correspondence with D. Brian Burghart, June 25, 2017.

[6] Email correspondence with D. Brian Burghart, June 25, 2017.

[7] http://making-sense.eu/publication_categories/toolkit/

[8] Interview with Professor Mel Wood conducted 15 August 2018.

In the first campaign organisers focused on using sensors in locations scattered around the country to find areas with the highest concentration of pollution; members generated 73 sessions of data from every Kosovo region (Ibid). The second campaign narrowed to one of its most polluted cities, the capital Prishtina, where volunteers concentrated much of their efforts on a primary school that they monitored for two months. The 3rd campaign looked largely at areas that had proximity to coal powered plants, which were some of the most significantly polluted sites in the country.

The first phase used analog diffusion tubes that were not connected to the internet but provided a meaningful baseline for analysis in a lab. Participants had to collect this data, which measured nitrogen oxide levels, then share it manually (McQuillan, 2015). The sensors were calibrated with equipment provided by the U.S. Embassy, whose instruments were considered more reliable than the Kosovo EPA.[9] After data collection, the first campaign was able to demonstrate that levels of nitrous oxide at hotspots exceeded EU limits by large margins. In the second campaign Dylos DC 1700 Sensors measured for PM2.5 particles, micro-particles that increase a person's chance for respiratory diseases and lung cancer. Volunteers found these dangerous micro particles prevalent on the primary school grounds most days (PEN, 2017).

In terms of data management for these campaigns, participants were focused on aggregating data from the sensors and then interpreting the data for immediate publicity. Participants who collected data could, via their sensors, see the peaks and troughs of their measurements, their walking route, and hot spots of poor air quality that showed up in red on mobile app.[10] After completing the measurements, participants sent their geotagged data to a single member who processed it for aggregated longitudinal data collection; our interviewee called this person "the black box of Kosovan data collection." From there, the participant uploaded the data to GitHub and indexed it on a free, open source platform called Smart Citizen.[11]

The Smart Citizen platform, created by the Fab Lab Barcelona, provides a data management tool for citizen sensing projects; it stores sensor data and showcases it through a dashboard and a map of sensing data uploaded by all the registered users worldwide. On the map a user can select specific sensors for more detailed analysis and to see how the sensed phenomena changes over time (Making Sense, 2016). Smart Citizen has a distributed version control system, allowing decentralised control and ownership of the data, so the Making Sense member who uploaded the data could access and add to it; the account could either be shared with other members or members could create their own accounts and upload data. Making Sense therefore didn't operate as a networking tool for the Kosovo activists; instead it put their data in the context of the world-wide community sensing movement.

Once the data were aggregated, another Making Sense team member interpreted the findings by providing a short overview of the values, air pollution levels and possible health impact. These details formed the basis for articles sent to mainstream and social media and drove their campaigns around the issue. Participants designed a media campaign that entailed taking slogans and dummies with masks to the street to open up the conversation around results of data collected. Because of the participants' foundation of radical democratic, participatory approaches, said our interviewee, they

[9] Interview with Professor Mel Wood 15 August 2018.

[10] Interview with Professor Mel Wood 15 August 2018.

[11] The platform was developed in Java and HTML5, and it allows developers to build new features on top of existing applications (Diez and Posada 2013). Certain digital sensors can send data directly to the Smart Citizen dashboard, but that wasn't the case here.

had the competencies to make their collective action highly effective.[12] One of the primary results was that air quality conditions became publicly visible and tied to local health problems, and this visibility pressured the KEPA to publish its environmental data for public use (ibid.). Even more significantly, the government wrote a citizens' right to clean air into Kosovo's Constitutions thanks in part to the pressures of the campaign.

Problem 3: Preserving the Public Domain

The Environmental Data Governance Initiative (EDGI) began in November 2016, soon after Donald Trump was elected to the presidency of the United States. Internationally, scholars shared concerns that Trump's ideological position on climate change would result in the removal of already-existing public resources on this and related topics. Some of EDGI's founding members are from Canada, where they remember former Prime Minister Stephen Harper's administration physically destroying scientific libraries and archives and silencing government climate scientists (Glass, 2016; Kupferman, 2016). EDGI members similarly feared the Trump administration would reduce the capacity for scientists to produce knowledge and leverage science-based calls for reform and regulation (Paris et al., 2017; Rinberg et al., 2018).

EDGI is a primarily volunteer network[13] that investigates potential threats to the scientific research infrastructure necessary to create and enforce environmental and energy policy. It includes 160 active members, with a volunteer community of over 1,100 who identify broadly, from community organizing to web development and academics (Knutson et al., 2018). New collaborators must be nominated and voted in by existing members; consensus must be reached before any decision is made regarding the tasks to be executed. A steering committee governs the activities of the organization (EDGI, n.d.a).

One essential part of EDGI's earliest work involved the coordination of DataRescue events around the United States (along with the organization DataRefuge). DataRescues, which typically took place at university campuses, invited members of the public to gather to flag and copy federal scientific datasets, documents, and webpages into a patchwork of repositories (InternetArchive, n.d.). This process involved coordination with the Internet Archive's (IA) end-of-term (EoT) crawler that routinely archives .gov webpages in periods of executive agency transition (Data Rescue, 2017, InternetArchive, n.d.). However, in some cases, webpages, data sets or other elements within the volunteer-flagged websites could not be crawled and archived by the IA's EoT crawler. In this case, participants built bespoke tools to scrape and archive the uncrawlable data sets (Data Rescue, 2017).

To effectively manage the uncrawlable datasets, volunteers designed an open source web application called Archivers.space, a project management tool that uses archival principles to manage the dataset's full lifecycle (EDGI, n.d.b). The tool tracks the dataset from its uploading to an Amazon server through multiple stages of research and vetting by participants. The vetting entails providing checksums to confirm data integrity and creating a .zip file that includes descriptions of the dataset's chain of custody, context, and provenance. A subset of EDGI volunteers focused on archiving governed Archivers.space; other volunteers who work in Archivers.space would get permissions from event organisers to participate.[14] To take part in the checksums and

[12] Interview with Professor Mel Wood contucted 15 August 2018.

[13] A handful of EDGI members are employed through grant funding to keep the organization running (Knutson et al, 2018).

[14] From DataRescue Workflow. http://datarefuge.github.io/workflow/researching/

describing phases, volunteers needed a background in library science or to have participated in Data Rescues and other EDGI events.

EDGI is currently part of a collaboration called Data Together (DT), which, along with hosting events and conducting research, is developing a way to preserve and make EDGI's data accessible to broader interested publics (EDGI, n.d.a; Knutson et al., 2018). The collaboration includes Protocol Labs, a project devoted to creating a distributed file system, also called an Interplanetary File System (IPFS), and with qri.io, which allows collaborative data sharing on the distributed web. DT is hoping to address the needs of members of the scientific community and grassroots advocates and organisations who may not have the data infrastructure expertise necessary to extract meaningful information from government data portals. The Data Together collaboration is unique in that it seeks to conceptualise and practice distributed, community-driven data stewardship. Each organization participating in the DT partnership grants institutional approval on major decisions, such as how sharing privileges will be determined, and the mode of consensus used for decision-making.[15] The DT team, comprised of members from EDGI, Protocol Labs and qri.io, uses EDGI's model described above for garnering consensus to adjudicate the further construction of the platform, including technical decisions about data formatting and description.[16]

EDGI suggests the potential for the open data movement to be interrogated through activist practices, by critiquing the inequitable power relationships between citizens, government, and the private sector to access information and use it to shape society. The open data movement, much like the open software and open access movements, advocates for placing research, administrative, and civic data into the public domain, often with the stated goal of improving governance (Sánchez and Viejo, 2017; Kitchin, 2014; Obama, 2009). The case of EDGI shows that providing and promoting open data can also be an activist project. EDGI also addresses fundamental issues of ownership and control at the level of infrastructure, which concern any project with long-term preservation goals. EDGI's DataRescue work not only supplements scientists' research but also pays close attention to the politics of technological infrastructures by designing a low-barrier, distributed, participatory platform with traditional archival protocols.

Scale, Temporality, Governance, and Values in Design

In all three case studies above, data management infrastructure plays a critical role in shaping the tactics and political formations of data activists. Whether the data is being collected, shared, or archived, the placement of the data into a publicly accessible repository in all three cases is a crucial part of mobilising collective action, creating accountability, building community, and exposing an issue to the public. Yet when comparing these projects, they differ in their temporal dimension, geographic scale, governance structures, and sociotechnical politics.

Matters of data management can be affected by geographic scale. In the first and third cases, contributors to the project are geographically dispersed, so their data management strategies included an accessible interface to elicit and inventory contributions from far-flung volunteers. For Fatal Encounters and EDGI, the web interface both networks the participants and manages the data, with various levels of access privileges to participants depending on different phases of data management in

[15] From Paris' participation in conversations with Data Together through the Spring of 2018, including a Data Together Community meeting held 12 March 2018. Stream of the meeting accessible at: www.youtube.com/watch?v=zeY_fYknpM8&t=587s&index=14&list=PLtsP3g9 LafVul1gCctMYGm9sz5FUWr5bu

[16] Conversation with Data Together, March 12, 2018.

each project. In Kosovo, on the other hand, data contribution was neither web-generated nor web-based; the group did not rely on networked platforms to carry out their work but met face to face to make decisions and share knowledge and outcomes. Aggregating and uploading the data to the Smart Citizen platform was the work of one person; rather than organizing and mobilizing distributed volunteers, the online platform was put to use for media campaigning in the project's publicity phase and to connect the project to the wider citizen sensing movement. So geographical scale in part determines how much data management infrastructures also act as networking platforms, and therefore shape the relations among participants.

Temporality plays another factor in data management practices. Urgent issue-oriented campaigns need data as evidence to make claims to the public and authorities. In many cases, data activist projects do not require a long-term data strategy, especially when the data can be transformed into political communication and immediately put to use. Each of these three projects deploys a different temporal approach their data. In Kosovo, the project was more concerned with capturing an immediate snapshots of air quality in the country and its capital and putting these to use in existing environmental campaigns. Currently the data are static, and there are no signs that the activists plan to contribute beyond the datasets collected during the original three campaigns. Fatal Encounters, on the other hand, does rely on software that can easily facilitate long term, ongoing data capture. The project uses on off-the-shelf, corporately-owned freeware to maintain the data over the long-term, as is common in many volunteer-led projects with no formal institutional or technological support; these tools get the work done of mobilizing volunteers and publishing data for others, such as journalists, to access easily. EDGI, in our third example, designed archival principles into its bespoke open source data management software. Such a strategy makes sense for ensuring control over scientific data that must be highly reliable and available to establish long-term environmental trends such as climate change.

The governance structures of the three projects also played a role in data management and access. Burghardt centrally controls Fatal Encounters by maintaining the data over time and vetting all contributors, who are largely anonymous to each other. For Making Sense, contributors came together face-to-face to decide collectively on data collection and publicity strategies, but one person largely took control of the data aggregation and publishing steps. EDGI's processes show how all aspects of data management can be collaborative and federated – including the design of software itself – while still maintaining some access restrictions to maintain data integrity.

Yet while Fatal Encounters has the most top-down management structure, it is also the easiest for participants to contribute to both in terms of technological know-how and access permissions. To add to Fatal Encounters database, one fills out a Google form and waits for vetting by Burghart. EDGI's Archivers.space, on the other hand, requires participants to learn the system and have some specialised knowledge to take part in certain aspects of the archival process. Making Sense, as well, required training before participants could use the sensors, and the Smart Citizen platform requires an account and understanding of how to sync data to its platform. None of the projects, therefore, were entirely horizontal but had various asymmetries in terms of governance and access depending on the stage of the data handling. Again, data management processes in this way shape, as much as are shaped by, projects' governance and access structures.

Finally, in two of the cases, the data management strategies and software used reflect the political structures and ideologies of the collective action projects themselves. The use of open source software was a deliberate choice for EDGI and

Making Sense, groups devoted to openness, semi-horizontal governance and inclusivity. Archivers.space reflects a commitment to collaborative but federated and decentralized contributions that still leave room for various levels of access. To publish their data, Making Sense selected an open source platform that contextualises their political activity in relation to hundreds of other citizen sensing projects around the world. This custom-build software reflects citizens' political choice to support public domain resources and to remain autonomous from corporate data capture, a decision that can be especially important for activists collecting sensitive personal data. That said, open source software can also can create greater technical barriers for participants who do not have the skills or luxury time to design or learn custom software for their needs. Fully bespoke software custom-built for a project, such as Archivers.space, can be a difficult bar for most activist projects. Instead activists can try to seek out not-for-profit data management software alternatives to corporate platforms, such as Smart Citizens, that can be used by many activist projects at once.

"Back Up! Back Up! We Want Freedom! Freedom!"

Data activism makes use of powerful tools for constituents to voice their perspective, whether it be through holding law enforcement accountable for poor and dangerous policing practices, or the empowerment that comes from shared scientific evidence. This article argues for scholars to give more attention to the data management practices of activists; it provides some heuristics for analysing activists' data management infrastructures, primarily asking about the ability of participants to take part in aspects of the data management and the politics of the technical platforms used. Through a comparative case study analysis, we show how these infrastructures can relate to a project's temporal goals, its governance structures among participants, the project's geographic scale, and the need for activists in some cases to consider the political economy of their management tools.

While Black Lives Matter protesters in Ferguson chanted loudly, "Back Up! Back Up! We Want Freedom! Freedom!" at lines of riot police, we can see how this mantra also applies to data activists in the face of enormous power asymmetries in terms of data ownership and control. While data activism may not appear as valiant an act as a street protest, managing and maintaining grassroots data promotes immeasurable public good in the long-term.

References

Bates, J., Lin, Y.-W., & Goodale, P. (2016). Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society*, *3*(2), 2053951716654502. https://doi.org/10.1177/2053951716654502

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT Press. Retrieved from https://www.jstor.org/stable/j.ctt5hhbk7

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.

Star, S.L. and Bowker, G.C. (1999), Sorting Things out: Classification and its Consequences, MIT Press, Cambridge, MA.

Bruno, I., Didier, E., & Vitale, T. (2014). Statactivism: Forms of Action between Disclosure and Affirmation. *Partecipazione E Conflitto. The Open Journal of Sociopolitical Studies*, *7*(2), 198–220.

Burghart, D. B. (2017a). Fatal Encounters. Retrieved March 31, 2017, from http://www.fatalencounters.org/2013/09/

Burghart, D. B. (2017b). Methodology |Fatal Encounters. Retrieved March 31, 2017, from http://www.fatalencounters.org/methodology/

Currie, M., Paris, B. S., Pasquetto, I., & Pierre, J. (2016). The conundrum of police officer-involved homicides: Counter-data in Los Angeles County. *Big Data & Society*, *3*(2), 2053951716663566. https://doi.org/10.1177/2053951716663566

Dalton, C. M., & Stallmann, T. (2018). Counter-mapping data science. *The Canadian Geographer / Le Géographe Canadien*, *62*(1), 93–101. https://doi.org/10.1111/cag.12398

Dalton, C., & 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 & Space*. Retrieved from 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/

DataRescue. (2017). DataRescue Workflow. Retrieved February 9, 2018, from https://datarefuge.github.io/workflow/harvesting/

Desrosières, A. (2002). *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge: Harvard University Press.

Diez, T. and Posada, A. (2013), “Smart Citizen”, available at: https://iaac.net/research-projects/ intelligent-cities/smart-citizen/ (accessed 5 August 2018).

Dodge, M., & Kitchin, R. (2007). The automatic management of drivers and driving spaces. *Geoforum*, *38*(2), 264–275. https://doi.org/10.1016/j.geoforum.2006.08.004

EDGI. (n.d.a). About. Retrieved August 16, 2018, from https://envirodatagov.org/about/

EDGI. (n.d.b). Archiver’s Space. Retrieved August 15, 2018, from https://www.archivers.space/

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Mit Press.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin’s Press.

Frost, D., Collins, S., & Kitchin, R. (2015). Funding models for Open Access digital data repositories. *Online Information Review*, *39*(5), 664–681. https://doi.org/10.1108/OIR-01-2015-0031

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J., & Rasmussen, B. (2009, January 20). Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes [Programme/Project deposit]. Retrieved August 13, 2018, from http://repository.jisc.ac.uk/279/

Futrell, R. (2003). Technical Adversarialism and Participatory Collaboration in the U.S. Chemical Weapons Disposal Program. *Science, Technology, & Human Values*, *28*(4), 451–482. https://doi.org/10.1177/0162243903252762

Gieseking, J. J. (2018). Operating anew: Queering GIS with good enough software. *The Canadian Geographer / Le Géographe Canadien*, *62*(1), 55–66. https://doi.org/10.1111/cag.12397

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Glass, H. (2016, November 22). Attack on climate action under Trump? It happened in Canada. *Christian Science Monitor*. Retrieved from https://www.csmonitor.com/Environment/Inhabit/2016/1122/Attack-on-climate-action-under-Trump-It-happened-in-Canada

InternetArchive. (n.d.). End of Term Web Archive: U.S. Government Websites. Retrieved August 16, 2018, from http://eotarchive.cdlib.org/

Irwin, A. (2001). Constructing the Scientific Citizen: Science and Democracy in the Biosciences. *Public Understanding of Science*, *10*(1), 1–18. https://doi.org/10.1088/0963-6625/10/1/301

Jalbert, K., Rubright, S. M., & Edelstein, K. (2017). The Civic Informatics of FracTracker Alliance: Working with Communities to Understand the Unconventional Oil and Gas Industry. *Engaging Science, Technology, and Society*, *3*(0), 528–559.

Kinchy, A., Jalbert, K., & Lyons, J. (2014). What is Volunteer Water Monitoring Good for? Fracking and the Plural Logics of Participatory Science. In *Fields of Knowledge: Science, Politics and Publics in the Neoliberal Age* (Vol. 27, pp. 259–289). Emerald Group Publishing Limited. https://doi.org/10.1108/S0198-871920140000027017

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Thousand Oaks, CA: Sage.

Knutson, S., & et al. (2018). *EDGI Annual Report 2018*. EDGI. Retrieved from https://envirodatagov.org/publication/edgi-annual-report-2018/

Kupferman, S. (2016, December 16). Q&A: Michelle Murphy, the U of T professor who’s racing to preserve climate-change data before Donald Trump takes office. *Toronto Life*. Retrieved from https://torontolife.com/city/toronto-politics/qa-michelle-murphy-u-t-professor-whos-racing-preserve-climate-change-data-donald-trump-takes-office/

Lauriault, T. P., Craig, B. L., Taylor, D. R. F., & Pulsifer, P. L. (2007). Today’s Data are Part of Tomorrow’s Research: Archival Issues in the Sciences. *Archivaria*, *64*(0), 123–179.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Liboiron, M. (2015). Disaster Data, Data Activism: Grassroots Responses to Representations of Superstorm Sandy. In J. Leyda & D. Negra (Eds.), *Extreme Weather and Global Media* (1 edition). New York: Routledge.

Lovink, G. (2011). *Networks Without a Cause: A Critique of Social Media*. Wiley.

Making Sense (n.d.). *Making Sense: The Toolkit.*

Making Sense (2016). *Online Technical Toolkit.*

McEwen, T. (1996). *National data collection on police use of force*. U.S. Dept. of Justice, Office of Justice Programs, Bureau of Justice Statistics.

McQuillan, D. (2015, December 14). Science for Change Kosovo Year 1. Retrieved August 21, 2018, from http://danmcquillan.doc.gold.ac.uk/scienceforchangekosovo-year1.html

Milan, S., & van der Velden, L. (2016). *The Alternative Epistemologies of Data Activism* (SSRN Scholarly Paper No. ID 2850470). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2850470

Mumola, C. J. (2007). *Arrest-related deaths in the United States, 2003-2005*. Washington, D.C.: Bureau of Justice Statistics.

Murphy, M. (2006). *Sick Building Syndrome and the Problem of Uncertainty: Environmental Politics, Technoscience, and Women Workers* (1 edition). Durham N.C.: Duke University Press.

Obama, B. (2009). Transparency and Open Government. Memorandum for the Heads of Executive Departments and Agencies. *Federal Register*, *74*(15). Retrieved from https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment.

Paris, B. S., Dillon, L., Wylie, S. A., Pierre, J., Pasquetto, I., Marquez, E., … EDGI. (2017, September). Pursuing a Toxic Agenda. Retrieved February 5, 2018, from https://100days.envirodatagov.org/pursuing-toxic-agenda/

PEN, S. for C. N. (2017). Children poisoned by pollution - Kosovo 2.0Kosovo 2.0. February 10. Retrieved August 21, 2018, from http://kosovotwopointzero.com/en/femijet-po-helmohen-nga-ndotja-e-ajrit/

Priest, S. (2013). Critical Science Literacy: What Citizens and Journalists Need to Know to Make Sense of Science. *Bulletin of Science, Technology & Society*, *33*(5–6), 138–145. https://doi.org/10.1177/0270467614529707

Ribes, D., & Jackson, S. J. (2013). Data Bite Man: The Work of Sustaining a Long-Term Study. In *Raw Data is an Oxymoron* (pp. 147–167). Cambriadge, MA: MIT Press.

Rinberg, T., Anjur-Dietrich, M., Beck, M., Bergman, A., Derry, J., Dillon, L., … EDGI. (2018). *Changing the Digital Climate* (100 Days and Counting No. 2). EDGI. Retrieved from https://100days.envirodatagov.org/changing-digital-climate/

Ruppert, E. (2012). The Governmental Topologies of Database Devices. *Theory, Culture & Society*, *29*(4–5), 116–136. https://doi.org/10.1177/0263276412439428

Sánchez, D., & Viejo, A. (2017). Personalized privacy in open data sharing scenarios. *Online Information Review*, *41*(3), 298–310. https://doi.org/10.1108/OIR-01-2016-0011

Smith, M., & Austin, R. L., Jr. (2015, May). Launching The Police Data Initiative [blog]. Retrieved from https://www.whitehouse.gov/blog/2015/05/18/launching-police-data-initiative

Srnicek, N. (2016). *Platform Capitalism* (1 edition). Cambridge Malden, MA: Polity.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Strasser, C. (2013, April 24). Closed Data… Excuses, Excuses. Retrieved August 13, 2018, from http://uc3.cdlib.org/2013/04/24/closed-data-excuses-excuses/

Tufekci, Z. (2014a). Engineering the public: Big data, surveillance and computational politics. *First Monday*, *19*(7). Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/4901

Tufekci, Z. (2014b). Social Movements and Governments in the Digital Age: Evaluating a Complex Landscape. *Journal of International Affairs*, *68*(1), 1–18.

Wylie, S. A., Jalbert, K., Dosemagen, S., & Ratto, M. (2014). Institutions for Civic Technoscience: How Critical Making is Transforming Environmental Research. *The Information Society*, *30*(2), 116–126. https://doi.org/10.1080/01972243.2014.875783