Urban Data Analytics as Research Topic, Method and Ethical Concern[1]

Daniel Trottier, Ju-Sung (Jay) Lee and John Boy

Abstract

Local and global business interests assemble images of neighbourhoods from localised knowledge, including disparate forms of public data such as reviews, blog posts, and open data from municipalities and other organisations. (In)visible forms of working with and worrying about neighbourhood data can be understood as an engagement with the neighbourhood's reputation, or rather its symbolic trajectory: a set of tangible and intangible indicators through which an urban space is known and treated accordingly over time. This paper addresses ethical concerns that emerge from contemporary datafied urban ethnography. We consider a combination of large-scale and bespoke, quantitative, and qualitative analyses of available sources with sustained ethnographic engagement with a Dutch neighbourhood coping with a troubled reputation. While the latter activities can mitigate ethical issues stemming from the former, ethnography in turn raises further concerns of exploitation and risk exposure and should not be treated as a kind of 'ethical panacea' for big, open, or public data projects. A multifaceted and interrogative approach to data collection may offer a more rounded account of contemporary urban data practices by drawing upon distinct and possibly conflicting accounts of social life. The challenge is to prioritise under-represented and otherwise marginalised voices in both the design and the dissemination of research on urban data analytics.

Keywords

Data Analytics; Urban Data; Big Data; Digital Ethnography; Surveillance of Neighbourhood

1 This research was funded by a grant from the Leiden-Delft-Erasmus Centre for BOLD Cities.

DOI 10.14361/dcs-2021-0215

Introduction

At the time of writing, a Google News search of the Dutch neighbourhood Moerwijk revealed terms like 'rattenoverlast', 'rattenplaag', and 'ratten zo groot als katten' (*rat nuisance, rat plague* and *rats as big as cats*). For a prospective resident, this may be one of many encounters with digital content that paints an unflattering account of the community. While the rat problem may be a demonstrable fact (as would be the case elsewhere), its prominence online is evidence of this neighbourhood's embattled reputation. Public records and newspapers have long remained a means to render urban spaces meaningful, generating knowledge that serves to exclude disadvantaged communities (Gutsche 2015). The domestication of the internet has augmented this knowledge production, notably through user-generated and geo-located content on social platforms. A potential democratisation of open source dataveillance (van Dijck 2014) is shaped by the uneven adoption of data services. Many actors play a role in producing, circulating, and exploiting data, even in neighbourhoods that are not considered 'hot' or otherwise relevant. Neighbourhoods far removed from flashy 'smart city' initiatives are nevertheless heavily datafied.

This paper is informed by ongoing research on Moerwijk, a residential and partly post-war neighbourhood located in the south-western reaches of The Hague. Our case study of this neighbourhood considers both beneficial and harmful uses of (personal) information technologies in making neighbourhoods openly legible. As researchers, we must consider the potential impact of our own data gathering activities on the individuals and communities who are implicated. Our datafied urban ethnography aims to contextualise aggregate data gathered from platforms. Addressing ethical and data protection concerns is increasingly an administrative and legal obligation for scholarly research, and one which we also address in other ongoing analyses (Lee et al. 2021). Beyond filling in forms when seeking approval from a governing body, ethical concerns should orient researchers in relation to their subject of analysis, ideally as an ongoing dialogue. Ethical reflection in this context accounts for how data is collected, the scholarly and broader public exposition of findings, and how they may be translated into other contexts and practices.

Our case study makes use of various sources of data about Moerwijk – notably public and openly accessible data – as a means to understand how the neighbourhood is generally represented. We take an open-ended and multi-sectoral approach to data collection. This includes statistical data from the CBS (Het Centraal Bureau voor de Statistiek), the Dutch census agency, digital platforms like Twitter and Instagram, and real estate data, including publicly available data on platforms like Funda.nl. These sources provide a multi-contextual account of Moerwijk as an urban space that generates data. Critical analysis driven by surveillance studies literature provides insights into the socio-cultural and political implications of this data, notably on the long-term symbolic trajectory of the neighbourhood. Yet a reasonable understanding of this trajectory – and its impact on those who depend

on the neighbourhood – can only be acquired through sustained ethnographic engagement. Not only are field work and interviews necessary counterpoints to large scale data analysis; they are a guiding source of insights on their own. Ethnography serves to follow up on hunches generated from the above analysis, while centring the experiences and perceptions of stakeholders.

This paper aims to acknowledge and begin to address the ethical concerns that emerge from contemporary datafied urban ethnography, in which so-called 'public' data is critically interrogated and contextualised through qualitative engagements with local stakeholders. After localising our ethical concerns below, the following section situates urban data in terms of mediated visibility, with geodemographics as a relevant antecedent. This is followed by methodological and ethical dilemmas in researching these data practices, focusing on various data sources and possible interactions between them. Here, we consider necessary steps that might not otherwise arise during ethical screening following conventional protocols. These considerations stem from our ongoing multimodal research on Moerwijk. Our focus aligns with broader concerns about the reproduction of stratification through the widespread adoption of digital platforms, the cumulative effects of these practices for disadvantaged communities (Gandy 2016), and the risk of gentrification of newly desirable neighbourhoods in times of mounting wealth disparity.

Some of these ethical concerns are easy to recognise. There is the ongoing risk of processing and publishing data that may identify or even stigmatise individuals. There is also the more aggregate and less tangible concern of bringing harm to disadvantaged communities. Beyond these risks, we can question the appropriate and desirable use of large pools of public, semi-public, and non-public data that shape our analyses (cf. Zimmer 2010). This paper is written in response to a rise of open data initiatives among research communities that advocate free access to scientific data, including data pertaining to urban environments and populations. There may be a net social benefit if certain forms of aggregate data are rendered accessible and legible to all. Yet making a broad range of data available to anybody with sufficient resources inevitably allows for abuse and exploitation, much of which we may only learn about when coping with its impact. While as-yet-unknown abuses of aggregate data may not be reasonable grounds to opt out of open data initiatives, they speak to the importance of assessing the ethical and broader societal impact of information-based infrastructures.

While such concerns may be distinct from digital media technologies and data, the latter form a surveillant infrastructure through which managers and owners can monitor essential services, labour, and other economic and cultural activities. Public discourse often states that particular neighbourhoods, residents and other stakeholders may be left behind in these developments. Presumably this means that existing disadvantages experienced by residents of such neighbourhoods would be exacerbated as surrounding – and competing – communities undergo digitalisation. There is also a potential, more tangible fear of abuses or

instrumentalisation of digital platforms within neighbourhoods like Moerwijk. Residents and local merchants may turn to Google, Amazon, and Meta-affiliated platforms to generate positive attention and revenue, even if residents and most stakeholders are in no position to own or fully grasp these tools. Yet in practice, these are clearly not equally allocated or equally exploitable services, especially when addressing profit-driven social media platforms that bear no responsibility for the communities that use them.

Urban Data and Symbolic Trajectories

Discrediting Google news alerts are part of a broader process by which a neighbourhood's reputation is reshaped through digital platforms. Local and global business interests integrate localised knowledge from disparate forms of public data. These include reviews of amenities such as restaurants and lodgings, blog posts, as well as open data from municipalities and other organisations. Urban data collection and analysis by a loose assembly of local actors and global platforms thus expand in terms of data sources and practitioners who generate cultural capital that "enhances the flow of visitors and developers" (Zukin et al. 2017: 475). This data may be framed as assisting residents and others who interact with Moerwijk, a residential neighbourhood in The Hague. Yet as seen above, its often-disproportionate focus on spectacular and stigmatising events and characteristics harms those who are dependent upon the neighbourhood, or who may otherwise be preoccupied with its standing. Data-based scrutiny and assessment of neighbourhoods – whether conducted by residents, investors, or academics – offer a disproportionate account of events, reducing multi-contextual accounts to specific measures by obscuring the interpretation that "always prefigures data analysis" (van Dijck 2014: 201). Simultaneously, an uneven uptake of digital media across stakeholders[2] may facilitate misrepresentations of marginal communities and the reproduction of systemic biases (O'Neil 2016). Yet even accurate and flattering portrayals are data practices that can either exclude or marginalise local participation in image construction.

Working with and worrying about data, considering surveillance capitalism, can be understood as an engagement with the neighbourhood's image, or rather its symbolic trajectories: a set of tangible and intangible measures through which an urban space is known and treated accordingly over time (Lee et al. 2021). These measures embody a collective process, as many actors are involved in the ongoing public standing of a neighbourhood. These include residents, elected officials and

2 For the purpose of our research, we define stakeholder as any individual or organisation who has a vested interest in and/or depends on the well-being of Moerwijk as a neighbourhood.

local businesses, but also global platforms – particularly social media – that enable those actors to embed their discourse in (and connect it to) similar concerns across the globe. Public discourse about the digitisation of urban spaces might centre on corporations like Alphabet's own endeavours (de Vynck/Wong 2020). Yet these platforms also enable others to engage in information collection, analysis, and exploitation, with social and organisational consequences.

Neighbourhood reputations are a form of "contemporary image ensembles", which local and global stakeholders routinely contribute to and draw upon (Mackenzie 2019: 5). These efforts are mobilised through "observation events distributed throughout and across devices, hardware, human agents" in addition to "artificial networked architectures such as deep learning networks" (ibid.). Image ensembles are mediated in a way that fuses routine social media practices with wider-reaching speculation about the value and commodification of neighbourhoods. Urban data analytics is a means for local actors and global platforms to generate capital, mobilising others to provide pervasive yet reductive accounts of services and spaces. Digital media is thus pervasively (ab)used in the symbolic trajectory of urban neighbourhoods. Our concern with the misuse of data compels us to address the ethical dimensions in mediated relations between residents, businesses, and global digital platforms. This includes identifying harmful uses of digital media, especially for those who may suffer from either gentrification or living in a stagnant and stigmatised neighbourhood. Freely available tools that pervade social and digital life but may remain restricted to residents in terms of the required media literacy and other access issues function to either improve or maintain the reputation of a neighbourhood. As urban data analytics rely extensively on (and thus advance) monitoring practices, particularly of disadvantaged communities in urban spaces, surveillance and visual studies can help address the mobilisation of data about neighbourhoods on digital platforms.

A mixed method approach is necessary because symbolic trajectories are composed of numerous data practices, diverging in terms of context and purpose. Large-scale data analysis by researchers in media, sociology, geography, and informatics (among others) begs for *qualitative* interrogation to critically question the arrangement of data as "'natural' phenomena" (van Dijck 2014: 202; emphasis in original). Moreover, researchers who study urban data from a surveillance studies lens are a (perhaps modest) part of this image ensemble and should shape their own data collection and analysis along often-conflicting ethical concerns. These include the potential misuse of available data, as well as the impact we may have in shaping the way data about these neighbourhoods (and neighbourhoods like them) are understood and utilised. As van Dijck remarks, academics risk being complicit in transferring "the power over data-collection and interpretation from the public to the corporate sector" (ibid.: 203). In observing and reporting on digital media practices in this context, researchers may impact those under study. They may strive to make their findings open to all, and these interventions distribute data that could reinforce inequalities within the spaces under study.

Researchers may be critical of surveillance practices in urban spaces while reproducing their harms, including assumptions about data and representation that carry through their analysis. Scholarship should aim to locate embodied, situated knowledge in a manner that does not facilitate its colonisation by the very practices we may critique, for instance in advancing new ways to measure and eventually commodify local (digital) cultural practices. The challenge is to refine mixed methods protocols that neither neglect situated knowledge and experiences, nor make them ripe for exploitation.

Surveillance of Neighbourhoods: Precedents of Data Technologies as Methodology

Surveillance technologies reshape relations among those who inhabit and/or invest in urban spaces, often by making aspects of a neighbourhood visible at the expense of its reputation. Public discourse frames these technologies as empowering for individuals, who may monitor and transmit information about their neighbourhoods. Readers could even expect that digital media may disrupt relations with employers, landlords and governments. Such visions suggest the possibility of residents' being visible on their own terms (cf. Albrechtslund 2008), but also collecting and disseminating information about more powerful actors (Mann et al. 2003; Reilly 2015). Under conditions of new and ubiquitous visibility (Thompson 2005), virtually any social phenomenon may be more easily monitored and scrutinised. Yet while data-led surveillance practices persist through the growth of platforms like Google, they can be characterised as volatile, as they are "not permanent but temporary, not equally imposed on everyone, and oscillating between appearing and disappearing" (Bucher 2012: 1177). Residents of marginalised neighbourhoods may cope with asymmetrical relations of surveillance, as well as asymmetrical relations regarding their knowledge of how this surveillance operates (Brighenti 2007). These concerns compel residents and others to engage in digital self-preservation of locational reputation (Kien 2020).

These demands on residents, merchants and other potential urban data stewards fuel a shared belief "in the objective quantification and potential tracking of all kinds of human behavior and sociality through online media technologies" (van Dijck 2014: 198), a tacit endorsement of continuous forms of data-based surveillance. Shared beliefs in objective urban data analysis may reproduce existing inequities (Gandy 2016) including racial, economic, and virtually any grounds upon which communities may be marginalised. While in principle there are minimal barriers to adopting localised online platforms and mobile devices, not all forms of access are equal, notably in contrasting local businesses and individual residents against global platforms and multinationals that purchase their services. Premium and paid versions of platforms typically offer greater analytic insights as well as greater bandwidth, compared to free and 'open-to-all' services.

An unpaid volunteer and a successful realtor may use the same tools in name, but in practice their affordances and outcomes may differ tremendously. Both free and premium services pose privacy concerns. No-cost versions of services like Google or Facebook typically generate and retain more data about neighbourhoods than they disclose to users. The continued ubiquity of these and other platforms in urban spaces raises persistent questions about privacy for stakeholders (van Zoonen 2016). While privacy may be framed as an individual concern, we can also interrogate how particular communities may be (dis)advantaged when profiled and assessed through persistent data collection.

As surveillance technologies with tangible research offerings for much social scientific research, geodemographic services coupled with GIS serve as a precedent to how neighbourhoods are monitored in diverse sectors (Burrows/Ellison 2004). GIS may still be in use, yet it has been supplemented with (a) social platforms on mobile devices in the domestic sphere, and (b) other localised data collection practices in other established sectors, including public services, real estate, and the hospitality sector. Geodemographics can be understood as a "social profile for the postal code" (Dalton/Thacker 2015: 3) which in turn interacts with a more abstract reputation of an area that is both ephemeral (difficult to 'pin down' and operationalise) and persistent (equally difficult to alter when more objective measures of reputation improve). Neighbourhood reputation is thus reproduced through datafied profiles. It is tempting to distinguish these in terms of quantitative (in the case of datafication) versus qualitative (in the case of a generalised reputation), while any single knowledge practice alone may be understood as holistic and multi-modal.

Many of these technologies are also effectively black boxes (ibid: 6), meaning that their functioning – including how they may assist in allocating outcomes to people – is not sufficiently understood, either as an abstract concept or in applied practice. The concept of the black box in this case is scalable. We can refer to any single platform, algorithm, or sensor as potentially unknowable to those who use or are otherwise impacted by it. At an aggregate level, the way in which these various tools are interoperable (or may be temporarily aligned) can also be rendered opaque as it gains prominence:

> Spatial Big Data is the logical outcome of long-running attempts to resolve these two built-in uncertainties of geodemographics. It does so through the promise of representing a fully measured, quantified, geolocated individual, rather than the homogenized, quantified areal units of geodemographics. (ibid.: 7)

These technologies work simultaneously with data from people and locations (both with comparable epistemological salience). GIS is assembled from disparate sources, but also framed as a singular tool that an administrator or public servant can wield (Haque 2003) – even while framed as "not a single system or database organizer as found in traditional statistical software" (ibid.: 43). There are other

actors in the GIS public servant scenario, yet it is typically restricted to public servants and private service providers. In comparison, the current data landscape in locations like Moerwijk is assembled even more in terms of the actors involved. In framing this phenomenon broadly, we include several different organisations that share conflicting concerns about Moerwijk, or concerns that happen to touch on Moerwijk.

Prior research also identifies "[s]ources of ethical misconduct" in practitioner use of GIS:

> (1) actions of the GIS user that could be deemed unethical due to the technical incompetence leading to biased decisions (limited knowledge in cartography); (2) misinterpretation of results due to lack of information or understanding of the true nature of the real-world phenomenon; and (3) concerns regarding the quality of data as public managers, because of high demand, are increasingly relying on private vendors and other alternative data sources. (ibid.)

While these are indeed examples of abuse of geodemographic data, there is an implication that greater technical competence, a greater volume of data, greater processing power, or fine-tuned algorithms will grant access to the "true nature" of urban life processes, which could in turn absolve practitioners from ethical concerns. Focusing on these risks overlooks concerns over whether data should be processed in the first place, alongside social harms that may arise from ever more refined surveillance of urban spaces.

Data Practices and Ethical Concerns: Tensions Between Open Data and Ensuring Anonymity

Datafied engagements of neighbourhoods like Moerwijk are informed by several principles in terms of accessing and handling public data. While commendable on paper, in practice they may provide conflicting or even contradictory guidance. Prominent calls by research clusters, disciplinary organisations and funding agencies promote open research data (e.g., NWO 2021). This involves not only making research findings open to the public, but also the datasets that underpins these findings. Such calls are especially salient when the data is about public spaces, and when it can potentially be of service to 'the public', however they may be defined. Open research data can benefit urban initiatives and may even directly benefit those who reside in otherwise underserved neighbourhoods. Yet extending the shelf life of this data also extends the possibility of misuse and unanticipated harms stemming from this collection. Calls for open data must be tempered by the prioritisation of privacy and data protection. As a starting point, this involves ensuring the anonymity (or pseudonymity) of those who may otherwise be identified in our data collection. Yet after addressing what we may consider the 'low

hanging fruit' of privacy concerns, there may be a lingering risk of reidentification, through the combination and processing of multiple data sources, as well as the aggregate identification of clusters of people through (as an example) postal codes. Setting aside these specific concerns, making a neighbourhood 'knowable' in the aggregate may also bring unanticipated harm, for instance when providing empirical accounts of the grounds for its troubled reputation.

Datafied urban ethnography aims to be "as open as possible, as closed as necessary" (Landi et al., 2020) by providing and prescribing a socio-cultural context in which urban data is understood. This involves positioning digital data about neighbourhoods as a point of inquiry for ethnographic research, rather than strictly as research data that can simply be extracted. It also demands critical inquiry and assessment for data collection and dissemination, all the while remaining open to the possibility of accessing new data sources, which would warrant a reconsideration of these concerns. This potential conflict between open and protected data raises questions about what right we have to collect and mobilise data about Moerwijk, as well as alternative arrangements. By providing the context (and narratives) of those whose lives are implicated in this data, we aim to allow those who depend on the neighbourhood to speak to the data, rather than the other way around. Yet researchers must acknowledge that they cannot control how open research data is later utilised, and qualitative contextualization can be overlooked by stakeholders who do not share our concerns about humanising data. We struggle with these dilemmas in the context of the ongoing datafication of urban spaces, and social scientists otherwise risk being left at the margins of these transformations if they adopt blanket measures of self-exclusion on ethical grounds. It is therefore perhaps best to see transparency not as a binary decision – to disclose or to mask – but as a "toolkit" (Reyes 2017).

Digital Data Sources

Any identifiable details from digital content, from Twitter or other data sources, could obviously be omitted in processing raw data. While we analyse and report on public tweets from various stakeholders, any details that would identify a private (non-public) individual are not included by default in reports or public datasets. These would be user accounts that, based on biographical or profile metadata, do not represent an organization, business, or public figure. If researchers wish to identify an individual via their Twitter content, they should seek permission from that individual directly. Depending on the context of the tweet, we employ a range of measures to ensure that non-public individuals are not identifiable (e.g., quoting only a portion of the tweet; paraphrasing; translating the tweet from Dutch to English). As for non-private actors (user accounts), because these are affiliated with the marginalised neighbourhood, not identifying them may incur a disservice by, for example, robbing them of (sympathetic and potentially benefi-

cial) public attention that they do not typically receive. This point is further elaborated upon below in a section that discusses explicit identification of neighbourhoods themselves.

We stress that our digital research activities, including the above privacy considerations, endeavour to maintain "contextual integrity" (Nissenbaum 2010). We, and other internet researchers, would argue that adhering to this standard should neutralise concerns surrounding some digital practices that are legal but nevertheless debated, such as data scraping that does not fully observe a platform's terms of service (AoIR 2019). The ethical consideration of such practices always needs to consider contextual factors (Fiesler 2020).

One ongoing challenge involves how to manage the possibility of as-yet unknown data sources during the course of the project (e.g., wearable data, new type of or newly emerged access to an app or platform, new forms of interoperability between existing data sources). These may provide novel insights into the symbolic trajectories of neighbourhoods, as expressed through the most recent iterations of embedded urban surveillance. Yet an unreflexive incorporation of additional sources risks displacing residents' accounts further, especially if they are unlikely to be early adopters of these technologies. In fact, our research has already had to contend with a shift in data sources. Prior to the availability of Twitter Academic, tweets were obtained through the standard and limited Search API and the Twitter web interface, both of which yield less than comprehensive data. In this case, the new data source can be considered an improvement over the status quo ante since Twitter as a well-known and popular platform provides insights into how multiple stakeholders frame the neighbourhood.

Ethnographic Field Work

As stated above, stakeholders comprise of residents, civil servants, employees and volunteers for local civic organisations, visitors to the community, representatives of small businesses, and public officials. Such engagements centre their experiences with how digital media and technologies shape and assess their neighbourhood, alongside their perceptions and judgements of these practices and their impacts on their lives, professions, and on the neighbourhood. Our preliminary impressions suggest that short-term engagement with communities (e.g., one-time visits to collect data, followed by complete disengagement) are not well received, and that a more prolonged dialogue with residents and other stakeholders is preferred. While this may be an uncontroversial statement, it also brings considerable staffing costs and possible delays in terms of publication targets. Moreover, a nearby neighbourhood (Schilderswijk) is dealing with what may be described as 'participant fatigue' after being the subject of several research projects (cf. Mandel 2003). Residents may want to share their opinions about the stigmatisation of their neighbourhoods; yet a disproportionate emphasis on these

themes by well-meaning ethnographers can be dispiriting to the subjects, who may come to internalise the very same stereotypes about their homes that the research set out to subvert.

On a global scale, neighbourhoods like Moerwijk are currently gripped by the COVID-19 pandemic. In the short term, this has limited urban mobility as well as the ability to congregate, furthered economic anxieties, and compounded other existential needs such as access to food, health care and other essential services. From an academic perspective, COVID-19 also poses distinct ethical challenges to researchers, as ethnographic engagements, face-to-face interviews and focus groups pose clear risks of infection for researchers and participants. The ongoing pandemic has of course complicated our fieldwork plans. While the Netherlands did not experience the strictest forms of lockdown seen in neighbouring countries, public health policies heavily restricted non-essential travel within the country. Moreover, it would have been possible to comply with these guidelines and still put researchers and participants at risk of infection, for instance, if an asymptomatic carrier of COVID-19 was in prolonged contact with others during field work. In the interest of minimising any such risks, we have opted for remote work during this period, including further engagement with the above-mentioned data sources. Most ethical evaluations of projects do include accounting for environmental risks to researchers and participants, so these concerns are by no means new. Rather, the troubling novelty is the global scale upon which such concerns have become a part of ethics planning, coupled with the uncertainty about immediate as well as longer-term data collection. At the time of writing, it remains unclear when sustained in-person field work will revert to a low-risk or no-risk endeavour.

When engaging with many kinds of stakeholders, there is the risk of encountering and exploiting power relations between members of the broader community, as well as an overdependence on key informants. While first contacts, well-connected insiders and even gatekeepers may serve an important role in reaching individuals, researchers need to ensure that their reliance on these links does not distort their account of the neighbourhood (O'Reilly 2009). Rather than identifying a single contact person as an entry point to the community, we propose treating field engagement as a series of iterative first encounters with contacts who are unaffiliated with one another and diverse in terms of demographics and opinion (in addition to following up already existing contacts). This enables researchers to be attentive to marginal and under-represented understandings of technological norms, complicating what would otherwise be concise ethnographic accompaniments to urban data analytics.

When contacting property developers, entrepreneurs, landlords and employers, their relations of power via clients, employees, and other individuals could potentially provide researchers with privileged and even coercive access to them, or their digital presence. Scholars should be especially mindful of these dynamics and account for these power relations in their reporting. To mitigate

power imbalances between stakeholders, we recommend actively avoiding situations where a more powerful respondent or gatekeeper will directly benefit from recruiting others. This includes assessing one's own involvement with business improvement districts and comparable entities, to assess the possibility that the knowledge utilisation that universities and funding agencies enjoy could disproportionally benefit already capital-rich entities.

More recent studies propose "innovative methodological approaches" (Fransham 2020: 2), including machine learning (Reades et al. 2019), to understand and predict gentrification through large-scale data sets. Mobilising localised data can be both a means to understand, as well as a manifestation of surveillance capitalism. In considering Moerwijk and neighbourhoods sharing similar features, we advocate a sustained ethnographic presence, not only to supplement the data activities, but also to consider how residents and others make sense of these activities. Ethnographic research may address ethical issues stemming from quantitative analysis by mitigating biases and preconceptions. Yet ethnography also furthers concerns of exploitation, exposure and misrepresentation (Wacquant 2002; Bratich 2017), and should not simply be treated as a kind of 'ethical panacea' for big, open, or public data projects. A multifaceted approach to data collection offers a rounded account of contemporary urban data practices by drawing upon distinct and possibly conflicting accounts of social life, but it also augments the risks of bringing harm to the community under scrutiny. This includes unintended consequences of drifting away from the principle of data minimisation, through the exposure of participants. Hence, a delicate balance between the benefits of data heterogeneity and these risks must be reached. This research thus invokes the ethical implications of using data in studying people and neighbourhoods, notably when focusing on these subjects' own use of data.

Linking Data and People

Bringing together these two types of data facilitates an understanding of the relations between the data infrastructures of urban spaces, and the individuals who depend upon them. Digital data in the case of Moerwijk can provide guidance to ethnographic work, by helping identify trends and reputational concerns that may or may not be of relevance to stakeholders. Preliminary findings can also be critically interrogated through ethnographic work. Data yielded from platforms like Twitter and Instagram present a vision of neighbourhoods like Moerwijk that may not align with the residents' own experiences. When our methodology reflects facets of surveillance capitalism that we may otherwise critique, it is imperative to incorporate and centre perspectives that are underserved by digital platforms, in part by sharing preliminary findings with residents, and allocating space for their interpretation of these findings in subsequent work.

Gathering online data and qualitative approaches should be treated as complimentary yet separate activities. As stated above, a mixed methods approach is best suited for studying multi-contextual and potentially conflicting practices. In cases where qualitative fieldwork alerts researchers to local web resources that are relevant to data analysis (for example, a representative of a local initiative flags their organisation's website), scholars may seek permission to make use of any data on these sites, while ensuring that no individuals are identifiable in the process. While interviewees may be recruited during observations, these interviews should be treated as distinct research activities, and the relevant protocols (e.g., informing participants; finding an appropriate location) should be adhered to.

The collection and analysis of digital data sources will contribute to field work activities, including interviews and focus groups. This data serves as an account of Moerwijk and allows us to identify potential concerns for stakeholders, which they may or may not be aware of (for example, in the event that we identify a pattern of comparatively lower review ratings for local businesses and services). Yet we do not want data to speak on behalf of people, nor do we want it to guide our conversations with residents and other stakeholders exclusively. To this end we advocate that the substantive focus of our (and comparable) research is also guided by stakeholders. Practically speaking, this can be accomplished on site through semi-structured and open-ended interview guides. It can also occur in a more sustained manner, for example, by encouraging participants to raise local concerns that will in turn shape the types of questions we ask, and even our choice of data collection and knowledge utilisation activities.

Evaluating Risks of Identifying Neighbourhoods

Anonymisation and pseudonymisation are common practices for reporting research findings that focus on individuals, especially if dealing with sensitive or stigmatising topics. Our case study focuses on an aggregate – a neighbourhood. Identifying a neighbourhood could facilitate the identification of individuals when combined with other details about those people, such as their profession. While this is a risk, a separate and arguably more perplexing concern is the broader impact that research and reporting have on the reputation and public standing of Moerwijk itself.

The simplest approach would be to omit any direct mention of the neighbourhood in the published results. Yet this can do a potential disservice to the neighbourhood, which would be excluded from broader dialogue and engagement linked to the project. If Moerwijk were anonymised in a peer-reviewed publication, a press release on that publication would also need to exclude any mention of the neighbourhood, as would any knowledge utilisation activities that put these research findings to use. Although comparable research may concern neighbourhoods that experience stigmatisation, data collection and analysis should critically

assess the processes through which such stigmatisation is digitally enacted, to provide residents and other stakeholders with opportunities to correct or counter undesirable perceptions. It can therefore be justified to include the names of target neighbourhoods in publications and other knowledge utilisation activities. While talking about a neighbourhood that is coping with a troubled reputation could potentially lead to further stigmatisation, scholarly and societal framing of this topic may rest on the observation that this reputation is largely unwarranted (empirically, both in quantitative measures as well as qualitative accounts) and that residents and other stakeholder voices should feature more prominently in public discourse. Thus, by default we believe it will do a disservice to communities like Moerwijk if they are omitted from publications and knowledge utilisation initiatives.

Despite these recommendations, we may consider whether addressing categorical discrimination through dispersed urban surveillance contributes to a more prevailing bifurcation of 'good' and 'bad' neighbourhoods. Coupling residents' experiences with digital metrics may privilege the former over the latter but can also lead to additional ways to quantify and rate an urban area's market value. In situating our research in a broader context of coalescent surveillance practices, we may inadvertently contribute to the financial and cultural revaluation of neighbourhoods like Moerwijk. For this reason, researchers must consider whether the most ethical step we can take for a neighbourhood is simply not to study it?

Conclusions and Recommendations: Ethics Beyond Ethical Approval

In accounting for how urban data surveillance shapes the reputation of embattled and neglected neighbourhoods, we propose a datafied urban ethnography that situates local data practices in terms of local data handlers. We also acknowledge that such an approach must strike a balance between finding (and using) the best data available and any potential harm that reporting on such data might incur. In our case study, this compels our own research team to consider the ethical implications of our efforts. This leads us to generate data in ways that mean residents have early access to our findings, and a first word in interpreting them. This prioritisation is reflected in our ethnographic work, but also public outreach efforts that are centred on Moerwijk and its residents.

Mixed methods approaches to urban data collection uncover vulnerabilities caused by information technologies when they contribute to the symbolic re-shaping of neighbourhoods. Researchers are often outsiders to the communities under study. While this should not exclude the possibility of critical insights or positive interventions on their part, when it comes to drawing upon a normative understanding of what a neighbourhood like Moerwijk 'ought to be', researchers need to defer to those who reside in, depend upon, and invest in these spaces. This

can be accomplished throughout the course of a project, informing ethnographic encounters and interviews. Knowledge utilisation activities – whether they target a localised audience, an audience of professionals at a national scale, or an audience abroad – should also centre on and prioritise how local stakeholders understand public data and digital platforms. Of course, this is not to say that researchers should not contribute to these conversations. Inasmuch as they develop knowledge and expertise on these matters through field work and prior research, they are expected to offer external insights to conversations with stakeholders. But in doing so, they need to ensure that these encounters remain a dialogue.

Researchers must also recognise that the first account, the loudest voice, or the majority perspective they encounter does not necessarily represent the opinions and needs of the community under study. While dependent on those who set aside time to offer an account of their neighbourhood, we recognise that they may present locally contested accounts of its digital presence and may have their own interests in doing so. At its core, our case study is concerned with the development of a 'positive' digital presence for urban spaces. Such a presence is made up of an assemblage of automated and deliberate forms of dataveillance. We need to reflect further on what is meant by 'positive' in this context, and consider who benefits from normative understandings of a favourable symbolic trajectory.

Contemporary data practices force a reconsideration of the terms we use to describe social actors and their practices. Both researchers and residents appear to be engaged in (or at least beholden to) data handling practices. A normative and orthodox account of what 'proper and professional' data handling is like might mean that we delegitimize or overlook certain practices involving urban data. A qualitative, ethnographic and above all curious approach to data use in urban spaces should help us to recognise how those at the margins may live with data. Yet researchers must also question assumptions that our recognition of these practices is inherently good – for marginal communities more generally and for the subjects of our research. Researching and publishing about otherwise neglected data practices can easily impact these practices – for example, by making such communities even more self-conscious than they might be already, and consequently tentative of their practices, and in this way, risk furthering existing asymmetries of access, capital, and power.

We must therefore address normative assumptions about what a good online presence should be like (for individuals and organisations alike), as well as expectations of optimal digital media usage. Our own research is formulated in a way that suggests that Moerwijk's digital media configuration is sub-optimal, disadvantageous, or dysfunctional. This can be framed from the perspective of those who are most dependent on the neighbourhood, who are typically also those who are most disadvantaged more generally (e.g., low-income residents with limited opportunities for upward or even lateral mobility). Yet because of the multi-sector nature of this topic, coupled with normative assumptions that shape the use of technology (e.g., what elements should be included in a 'functional' Google Review for a small

business; what remedies should be in place if these are absent?), what begins as a concern for those who are marginalised or excluded may end up largely benefiting those with a comparative advantage such as entrepreneurs, property developers, and global digital media platforms. Ideally, scholarly concern with those on the margins should align with their interests and voices. Further, residents' having these interests met and having their voices heard should bring minimal risk or burden upon themselves. This paper aims to contribute to translating these lofty ideals into practical steps for urban data scholars and ensuring that these steps may be of value for comparable data practices.

References

Albrechtslund, Anders (2008): "Online Social Networking as Participatory Surveillance." In: First Monday 13/3 (https://firstmonday.org/ojs/index.php/fm/article/view/2142).

Association of Internet Researchers (2019). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf.

Bratich, Jack Z. (2017): "Observation in a surveilled world." The SAGE Handbook of Qualitative Research, edited by Norman K. Denzin and Yvonna S. Lincoln, 5th ed., Thousand Oaks, Calif.: SAGE, pp. 526–45.

Brighenti, Andrea (2007): "Visibility: A Category for the Social Sciences." In: Current Sociology 55/3, pp. 323-342.

Bucher, Taina (2012): Want to be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook. In: New Media & Society 14/7, pp. 1164-1180.

Burrows, Roger/Ellison, Nick (2004): "Sorting Places Out? Towards a Social Politics of Neighbourhood Informatization." In: Information Communication & Society 7/3, pp. 321-336.

Dalton, Craig M/Taylor, Linnet/Thatcher, Jim (2015): "Inflated Granularity: Spatial "Big Data" and Geodemographics." In: Big Data & Society 2/2, pp. 1-15.

De Vynck, Gerrit/Wong, Nathalie (2020): "Alphabet's Dream of a Smart City in Toronto Is Over" In: Bloomberg May 7 (https://www.bloomberg.com/news/articles/2020-05-07/alphabet-s-dream-of-a-smart-city-in-toronto-is-over).

Fiesler, Casey/Beard, Nathan/Keegan, Brian C. (2020): "No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service." In: Proceedings of the International AAAI Conference on Web and Social Media 14, pp. 187-196.

Fransham, Mark (2020): "Neighbourhood Gentrification, Displacement, and Poverty Dynamics in Post-Recession England." In: Population, Space and Place 26/5, pp. 1-13 (online first).

Gandy, Oscar (2016): Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage, London: Routledge.

Gutsche Jr. Robert E. (2015): "Boosterism as Banishment." In: Journalism Studies 16/4, pp. 497-512.

Haque, Akhlaque (2003): "Information Technology, GIS and Democratic Values: Ethical Implications for IT Professionals in Public Service." In: Ethics and Information Technology 5/1, pp. 39-48.

Kien, Nguyen Trung (2020): "Care of the Self in the Age of Algorithms: Early Thoughts from a Foucauldian Perspective." In: Ho Chi Minh City Open University Journal of Science - Social Sciences 10/1, pp. 79-90.

Landi, Annalisa/Thompson, Mark/Giannuzzi, Viviana/Bonifazi, Fedele/Labastida, Ignasi/da Silva Santos, Luiz Olavo Bonino da Silva Santos/Roos, Marco (2020): "The "A" of FAIR–as Open as Possible, as Closed as Necessary." In: Data Intelligence, 2/1-2, pp. 47-55.

Lee, Ju-Sung/Boy, John, D./Trottier, Daniel (2022): "Symbolic Trajectories in Action: Digital Technologies and Representations of a Stigmatized Neighborhood." Under review.

Mackenzie, Adrian/Munster, Anna (2019): "Platform Seeing: Image Ensembles and their Invisualities." In: Theory, Culture & Society 36/5, pp. 3–22.

Mandel, Jennifer L. (2003): "Negotiating Expectations in the Field: Gatekeepers, Research Fatigue and Cultural Biases." In: Singapore Journal of Tropical Geography 24/2, pp. 198-210.

Mann, Steve/Nolan, Jason/Wellman, Barry (2003): "Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments." In: Surveillance & Society 1/3, pp. 331-355.

Nissenbaum, Helen (2010): Privacy in Context: Technology, Policy, and the Integrity of Social Life, Stanford: Stanford University Press.

NWO (2021): "Open Science" (https://www.nwo.nl/en/open-science).

O'Neil, Cathy (2016): Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, New York: Broadway Books.

O'Reilly, Kare (2009): "Key informants and gatekeepers." Key concepts in ethnography. London: SAGE, pp. 132-137.

Reades, Jonathan/De Souza, Jordan/Hubbard, Phil (2019): "Understanding Urban Gentrification Through Machine Learning." In: Urban Studies, 56(5), pp. 922-942.

Reilly, Paul (2015): "Every Little Helps? YouTube, Sousveillance and the 'Anti-Tesco' Riot in Stokes Croft". In: New Media & Society 17/5, pp. 755-771.

Reyes, Victoria (2017): "Three Models of Transparency in Ethnographic Research: Naming Places, Naming People, and Shar-ing Data." Ethnography 19/2, pp. 204–26.

Thompson, John B. (2005): "The New Visibility." In: Theory, Culture & Society 22/6, pp. 31-51.

Van Dijck, José (2014): "Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology." In: Surveillance & Society 12/2, pp. 197-208.

Van Zoonen, Liesbet (2016): "Privacy Concerns in Smart Cities." In: Government Information Quarterly 33/3, pp. 472-480.

Wacquant, Loic (2002): "Scrutinizing the Street: Poverty, Morality, and the Pitfalls of Urban Ethnography." In: American Journal of Sociology 107/6, pp. 1468-1532.

Zimmer, Michael (2010): "But the Data is Already Public": On the Ethics of Research in Facebook." In: Ethics and Information Technology 12/4, pp. 313-325.

Zukin, Sharon/Lindeman, Scarlett/Hurson, Laurie (2017): "The omnivore's neighborhood? Online Restaurant Reviews, Race, and Gentrification." In: Journal of Consumer Culture 17/3, pp. 459-479.