3

Paths to Public Benefit

Constructing Meaning from Our Physical and Built Environments through Digital Observation

Gina S. Lovasi, Steven Bedrick, Michael Brauer, Megan Doerr, Fabio Kon, Lindsey Smith, and Beate Ritz

Abstract

Digital data can be used to observe human behavior as well as aspects of the physical, built, and natural environment that provide context for such behaviors. Data extracted from communities through surveillance have rightfully been the subject of concern, yet such data hold great potential for benefits, including knowledge generation and dissemination to advance human health and equity. Benefits will depend on what is measured and who sets the agenda. Here, ways to organize available and future physical, built, and natural environment measures are discussed, and approaches are proposed to guide the use of such data to generate knowledge while keeping in mind varied value judgments and goals. Metadata are identified as a key tool to deter misrepresentation and misuse of data. To serve this purpose, metadata could be expanded in several ways, including historical context and intent of data collection as well as limitations and permissions to be aware of while planning use and interpreting findings. As data are used, subsequent versions of metadata could record information to inform future use, including a statement of social license updated as the individuals and communities affected by use of the data reflect on harms and benefits. The process of seeking social license for use of geographically referenced data itself has potential to add to our understanding of human agency and to inform ethical inquiry about the structural determinants and individual choices that play out in communities. Opportunities to fill gaps and meet future challenges are identified. Further, attention must be given to incentives across the funding, publishing, and institutional landscape so that envisioned change can be realized and sustained.

Introduction: Physical, Natural, and Built Environment Measurement for Digital Ethology

We are living amidst a revolution of geospatial data generation and use. Such data have the power to be transformative by improving our understanding of the physical, natural, and built environment and benefiting the public and individuals through valued outcomes such as health. Here, we consider the potential of place-based data within the emerging interdisciplinary field of digital ethology, which brings a multimodal perspective to the potential for accumulating data to describe and explain the bidirectional relationship between human behavior and its geospatial context (see Paus, this volume).

Implications of Accumulating Place-Based Digital Data

As we live our lives, data accumulate in records that are increasingly in a digital format. When we access our phones, we generate vast amounts of behavioral data, much of which can be anchored to our location at the moment and our recurring travel patterns. Further, we may add sensors to our homes to detect water leaks or other disturbances, and municipalities and governments digitally monitor and report on air quality and temperature. Individually, we benefit when we use location, satellite imagery, and real-time traffic congestion data to navigate to a restaurant or clinic. To attain these benefits efficiently, we may agree to monitoring of our mobility as a part of traffic density surveillance, which is then made available for broader use that extends far beyond our own planning. Stored imagery or video footage of public spaces, such as that recorded for security-related purposes, could additionally be used by researchers to study human behavior in daily life, as highlighted by Pallante et al. (this volume). Thus, knowledge generation[1] goals may be among uses that extend beyond those originally envisioned in planning or permitting digital data collection. Such research applications could use geospatial data resulting from digital surveillance for the common good; there is also a need, however, to manage and mitigate potential harm.

Potential Harms

While digital surveillance is an increasing and nearly ubiquitous reality,[2] digital surveillance has negative connotations due to known, suspected, and

1 In line with Kum et al. (this volume), we view knowledge as being created from data and usable to inform action. Many steps and much potential for missteps lie along the path from data to knowledge to action.

2 Some applications of geolocated data are designed for public health surveillance purposes, such as to monitor infectious disease outbreaks, as highlighted by Sarker (this volume). Here, we include not only passive methods that capture information about people, but also those that capture spatiotemporal variation in the spaces inhabited, traversed, or otherwise used by people.

feared uses and abuses. Many possible uses are not anticipated or may not be welcomed by the individuals whose data are assembled. For example, commercially marketed cell phone–location data have been used by police during criminal investigations (Burke and Dearen 2022) and can reveal presence at sites where sensitive medical services are provided (e.g., reproductive, mental, or behavioral health clinics) (Fair 2022). These present uses add to historic precedent to substantiate concerns that place-based digital data can reify and exacerbate systemic inequities and power imbalances. Harmful use of digital place-based data can include restriction of individual and collective rights.[3] Other costs to individuals and communities may result from commercialization of these data in ways that are misaligned with or undermine advances toward equity. Documented harms from use and abuse of digital data, even if well-intended, provoke questions about the legitimacy and acceptance of digital observation and resultant data use.

To mitigate harms that can arise from digital surveillance, strategies that increase transparency and limit abuse potential are required. In some instances, the risk of harm or lack of consent may be most appropriately addressed by not accumulating data. Where digital data can be ethically collected, however, their use should benefit the individuals and communities whose surroundings and activities are represented, such as through remediation of environmental harms that undermine health. Here, we frame this as *using geospatial data for good*, while recognizing that notions of "good" are highly subjective. Proactively thinking in these terms frames our obligation to produce public benefits while averting harm. Further, it highlights the need to include and amplify the voices of the communities who contribute to the data from the outset. Finally, investment in dissemination and translation is needed so that observation and knowledge generation can contribute to communities' data-informed advocacy and action.

This chapter distills our multifaceted discussions from the Ernst Strüngmann Forum in July 2022. During this week-long immersive event, we put forward a vision to advance scientific and societal benefits made possible by assembling digital data on the physical, natural, and built environment. To do so we identified types of data to be included, implications of sharing access to and power over such data, and strategies for creating and disseminating knowledge with attention to challenges specific to spatial data and to the values and needs of communities represented in this data. Before concluding the chapter, we highlight opportunities for team formation, cross-disciplinary training, and ways to shape our funding allocation, publication, and institutional incentives to support sustained progress toward our vision.

---

3 Here we are referring to rights, such as the right to life and liberty, but also note that the right to privacy is closely connected to concerns raised about digital surveillance. For further reading, see Chapters 10, 11, and 12 (this volume); for an overview of how human rights could inform ethical work with big data, see Mantelero (2018).

Types of Digital Data on Physical, Built, and Natural Environments

As noted by Smith (this volume), multiple existing data sources capture aspects of what is present in the environment (e.g., land cover such as pavement), how it is used (e.g., parking, playground), and quantitative characteristics that vary spatially (e.g., surface temperature, air pollutant concentration, annual precipitation, daily average sound levels). The lens of digital ethology suggests making human habitual behavior central to our typology of environmental measurement.

Human-Centric Quest for Measurement

For the purposes of this chapter, we chose a human-centric approach to classifying measures of the environment. Our emphasis is on public benefits and harms, where the humans who make up this public have lived experience expertise and value perspectives that need to be considered. This should not be interpreted as the only lens through which one can view potential global benefits of digital geospatial data; alternatives may emphasize aspects of the biosphere affecting multiple species. Here, we identify that a human-centric approach can bring attention to the following questions:

- Fitness for fulfilling human needs: How fit is the environment for fulfillment of human needs? In what ways does the environment create opportunities from the most fundamental (e.g., breathing clean air) to the most aspirational (e.g., artistic expression, co-creation of knowledge)?
- Suitability to how a place is actually being used: How fit is the environment for the currently enacted or desired use[4] by the community? What features may enhance uses for which a place was designed? What features contribute to unintended side effects, including those that arise from the mismatch between originally intended and current de facto uses? What constructs relate to fitness for the current de facto or emergent proposed use, such as livability, walkability, or accessibility?
- Design and redesign to encourage intended uses: What immediate- and long-term uses were deliberately accommodated or discouraged as the environment was built and rebuilt over time? Do we have direct accounts of the intentions[5] (e.g., oral history, transcribed discussion at

4 Uses of the surrounding built environment range widely, including the acquisition of food and other goods and services, mobility and physical activity, and social interactions from casual greetings to building collective identity and action (see Weigle et al., this volume, on social environment).

5 Intentions might, for example, be revealed by noting exclusive attention to private vehicle use in a planning document for gridded streets.

planning meetings or public hearings, archived documents, legislation) or can intentions be inferred based on the specific features present (e.g., hostile architecture to deter homeless encampments, loitering, or skateboarding (Petty 2016)?

These questions help to organize existing measures and can also lead us toward what additional data are required as new uses of environments are initiated or proposed, or as needs are newly articulated by communities.

Notably, the intentions of those who design and the needs of communities who use the environment are brought into the foreground. These intentions can be mutually informed, as emphasized by architects such as the Brazilian landscape designer Roberto Burle Marx who revisited design decisions after actual use has been observed (Montero and Marx 2001).[6] For example, Burle Marx maintained that the paths in newly opened public gardens should be formalized only a year after the space becomes available, reflecting the footpaths created by frequent community use (i.e., those routes through the space that have been demonstrated to be convenient and useful). This attention to emerging use can be applicable even in cities with a long history of human habitation and built environment change. There is also the possibility for observation across domains, measurement scales and time periods, and across emerging frontiers of measurement to inspire entirely new questions as we wonder about ways in which humans respond to the built environment "in the wild."[7]

Domains: What to Measure That Is Relevant to Human Needs and Uses of the Environment

As we explore digital data related to the lived environment, we find ourselves encountering a wide variety of domains of data, situated at varied levels of resolution and abstraction.

Beginning with impediments to foundational needs such as breathing clean air and sustaining thermal comfort, we may first consider data describing atmospheric properties of the physical environment (e.g., particulate concentration, humidity). Topological characteristics and type of land cover may affect these properties, along with how suited the landscape is for providing nourishment and shelter, what resources can be accessed, and what uses the spaces may support. Beyond describing the places used for housing, work, and leisure,

6 Other practices applied by Burle Marx in his work in Brazil may have relevance to natural features integrated into the built environment (e.g., specifying that gardens should prioritize native species and taking into consideration the preexisting natural and physical landscape). While on the surface these may not seem crucial to a human-centric approach, the perspectives of Indigenous peoples may bring further attention to these and other aspects of how we build.

7 The phrase "in the wild" is used here to convey that these are not settings artificially contrived to manipulate human behavior for research purposes, as might be seen in a laboratory setting. Humans in their current habitat largely means humans surrounded by structures and urban spaces built by and for humans.

geographic data can incorporate notions of secure tenure (ownership), safety, and private or restricted use spaces.

Going beyond physical attributes of terrain (i.e., topology, geology) to consider fitness for intended uses requires distinct measurement approaches even when situated at a similar geographic scale. Remote sensing and stationary sensors are especially valuable for visible environment measurement, including the presence of buildings and transportation-related structures. In contrast, administrative and participatory digital data collection approaches are often needed to capture aspects of the built environment[8] that relate to intended and actual use over time (e.g., availability and accessibility of health-care delivery or food establishments). Notably, quantities such as auditory noise may be operationalized via relatively objective measurements of physical properties at a particular point in space and time (i.e., ambient decibel level), yet whether a given decibel level is perceived as *unwelcome* noise can depend on the source, the listener, and the surrounding context. Many measures that relate to fitness for use are quite complex and inherently subjective, such as a "walkability score" (Wang and Yang 2019), which may be computed in any number of different ways and often relies on combining multiple sources of data. Developing and agreeing on methods of measurement is critical for deriving value from geospatial data in terms of how these data relate to human–environment interactions. Clarity about what to measure is a prerequisite to selection of relevant data sources,[9] and also to noting limitations specific to the task at hand. Going beyond methodological limitations, it is also important to explicitly examine and document sources of bias within one's data and choice of measures.

Further enhancing our understanding of the environment, we may consider data representing (and possibly directly generated by) discrete and ongoing human activities, including "sensor data."[10] This could include readings from traffic counters with relevance to mobility and vehicle emissions, as well as data sources providing insight into how people feel or act in a given space, such as geotagged social media posts. These data types may illuminate barriers to realizing benefits of intended land use. For example, two otherwise similar parks may be quite different with respect to physical and mental health benefits due to differences in surrounding vehicle traffic and associated noise, air pollution, and injury hazards.

---

8 We define the built environment as including human-built or modified structures, transportation systems, and features such as buildings, roads, plazas, and parks as well as fixed features such as fire hydrants and light posts.

9 The proposed use will determine whether available data are sufficiently relevant, and correspondence between what we aspire to measure and what we have represented in our data is never perfect.

10 This is intended broadly to include not only stationary sensors deployed for purposes of measurement but also device-based data such as accelerometry and geolocation data generated as people carry cell phones throughout their activity space.

There is a feedback relationship between the design and use of the environment; enough people participating in an individual activity can result in structural change, and vice versa. Further, observational data generation and subsequent knowledge generation can make ongoing use of a space more evident, and awareness of how a space is used can itself change use (e.g., people changing their behavior or chosen route in response to the presence of a cycle-counting monitor) or can bolster the case for sustained investment to facilitate use (e.g., monitoring the number of cyclists following the development of protected bicycle lanes can be used to make the case to maintain and scale up such protections).

The same physical feature may simultaneously span multiple domains categorized based on type of human use. For example, a bus stop could be both relevant to current community use for mobility as well as providing for rest or shelter because of the presence of a bench. Likewise, a mixed-use building including ground-floor retail and apartments may play a role in both the food environment and walkability at the neighborhood scale.[11]

Variable Scale and Timing Require Attention to Human-Drawn (and Redrawn) Boundaries

Data representing features of the physical and built environment range in scale in terms of spatial and temporal resolution, density, and precision.

Advantages of digital data include its volume and frequency—for example, measures that capture seasonal and even hourly fluctuations in air quality. Some digital data can be archived and later processed and transformed to limit the uncertainty due to temporal gaps in an observation series.

At a given time point, geographic space is divided into units of observation in ways that may align with how they are designed or used, ranging from a simple grid to human-drawn administrative units, including parcels, zoning areas, and plots of variable shape and size.

A challenge in digitally derived environment data is posed by human-drawn boundaries and features (rather than those that are naturally occurring and enduring). Human-drawn boundaries have different social, economic, and political functions, and are commonly used in research relying on geographic information systems (GIS), as described by Smith (this volume). Our organization and depiction of such boundaries benefit from a notion of spatial hierarchy, yet the spatial nesting of smaller areas within larger ones may be imperfect. In some scenarios, these hierarchies may be complex, and involve plural,

[11] Neighborhoods have been variously defined to include activity spaces frequently visited, areas important to resident identity, or the postal and other administrative units that provide a convenient but imperfect operationalization of neighborhoods (Lovasi et al. 2012). Together, neighborhoods contribute to larger geographic contexts such as the city-level patterns that connect physical and social environments; for further discussion, see Balsa-Barreiro and Menendez (this volume).

non-overlapping, and highly irregular attributes. Boundaries may be closely tied to elements of physical geography or existing infrastructure, such as bodies of water or utility and sewage networks. Historic processes shaping delineation may themselves be harmful, as in the case of gerrymandering (Sánchez 2018) or municipal fragmentation (André Hutson et al. 2012). Understanding the origin of these boundaries may have implications for contemporary use, including efforts to explain how various land uses arose and changed over time. For example, analyses concerned with equity and resource distribution benefit from use of historical information about boundaries such as those associated with *redlining* in the United States, which determined unequal access to loans and housing by race (Rothstein 2017).

Human-drawn boundaries may be driven by power or bias and are subject to challenge or overthrow. The built spaces marked by these changing boundaries may respond incrementally or suddenly.[12] An example of the latter can be noted in the city in which our Forum discussions to conceptualize this chapter took place: Frankfurt am Main, Germany. Frankfurt's old town underwent substantial physical rebuilding and administrative changes after bombing during World War II had destroyed most of its physical infrastructure (Lehné et al. 2013). Changing boundaries may occur in response to population growth or migration, which requires attention when working with longitudinal population characteristics based on census boundaries (Logan et al. 2014). These shifts can pose a challenge to systems of data management, knowledge representation, and statistical analysis.

Contextualizing Imposed Labels and Current Practices: A Case for Increasingly Inclusive Teams

Digital data that we have access to or envision to create arise from a legacy of geospatial work. Further, those engaged in generating and using environmental data to explain human behavior and health are influenced by our training to think of the world as compartmental and to formulate questions according to our specific professional lens as well as other aspects of identity. This can impede the match between community needs and what is measured about the environment. For example, within mobility research, amenities and services may not be equally matched to the needs of all demographic groups, and in particular, accessible toilets and benches that are critical mobility determinants for seniors have not been routinely captured in walkability measures. Thus, collaborations inclusive of perspectives across demographic categories such as age may yield new insights even for commonly addressed topics.

Provenance and identification of those who should set the agenda for measurement of and changing use of a space (e.g., public vs. private control) may

---

[12] In addition, the location and nature of boundaries can, of course, be disputed between groups of people or organizations, adding an additional layer of complexity.

not be easily established. The concepts and definitions discussed here and by Smith (this volume) are illustrative of measures commonly encountered in urban spaces in western, educated, industrialized, rich, democratic (WEIRD) countries where the systematic collection of data from the physical and built environment started several decades ago. Yet, the vast majority of humanity does not live in areas that have data availability typical of WEIRD countries. As we consider a truly global research agenda for digital data about the environment, collaboration with a broader cross section of researchers, communities, and policy makers from around the world will be essential.

Future research teams may include perspectives we are missing and may accordingly judge some ways of categorizing or labeling domains of environment measurement to be inappropriate. It is particularly important to recognize the need for bridging to work on topics of importance to equity, such as housing instability, with research in understudied parts of our human habitat. For example, Weinstein pointed out the narrow view of North American scholars on the topic of evictions and wrote an article on reconceptualizing housing insecurity by looking at the work carried out by scholars in India and South Africa on urban "slum" evictions (Weinstein 2021). The field will benefit from critically appraising current practice, assessing ways in which our categorization of data is or is not appropriate to other research scenarios, and articulating additional concepts that need to be developed.

Expanding Frontiers of Environment Measurement

Digital observation may open the door to research efforts, collaborations, and exchanges that cross national boundaries. Some types of data are already collected globally (e.g., Landsat, which collects satellite imagery from the entire Earth). Such data can now also be used with tools such as artificial intelligence algorithms in innovative ways (e.g., remote sensing images from the Amazon Rainforest to detect deforestation areas with the help of machine learning and citizen science) (Dallaqua et al. 2021).

Some types of data offer flexibility in generating data categories and constructs, and we note that imagery is one such type of data. Imagery can be used to capture pre-determined features and to enable future uses not envisioned at the start of data collection. For example, at present digital data to characterize quality and use of indoor environments is limited, even though these are the environments where most people around the world spend most of their time. Potential indoor environment data sources include indirect information derived from exterior imagery (e.g., building structure and details visible through remote sensing or façade features from street-level imagery) as well as imagery that more directly shows the indoor environment, but which may not represent the typical condition of that environment over time (e.g., from online real estate resources which include indoor images). Importantly, despite the flexibility of working with imagery, challenges arise due to measurement that is inferred, available only for a biased sample of places or times, or unreliable such as due to variation in weather, lighting conditions, obstructions, or other temporal aspects that may affect observation (e.g., image capture of a street before or after trash collection). Further, when using human raters of imagery to capture information such as perceived safety, there is a risk of embedding into resultant metrics any salient human biases, such as an implicit association of racial composition of a neighborhood with perceived safety. Preferences and perceptions differ, which makes an inclusive research team composition and practices like community consultation valuable in understanding what is being observed in digital data. Relevant to safety perceptions and equity, for example, experiences of over-policing may result in divergent responses by race to police presence.

Alongside digital datasets about the built and physical environment, spatially referenced human reactions to events can be captured, particularly through data sources like social media, as discussed by Sarker (this volume).[13] Social media can capture conscious reactions to physical features or associated construction efforts, possibly leading to behavior change or public demands. In contrast, users of a space may not be able to sense air pollution or notice resultant cumulative health effects, and therefore reactions to unseen or gradually harmful exposures are unlikely to be captured in social media posts. Novel insights and innovations may be facilitated by the increasing use of social media as a source of data, including insights into the perspectives of geographically delimited communities and other social or professional groups. Representativeness of such data must be considered, however, as different social media platforms may have greater affinity from particular user groups while other parts of society may be entirely excluded.

Beyond what data are presently recorded or monitored describing our built and physical environment, it is important to be aware of what *is not* being measured. Even where certain aspects of the physical or built environment are currently challenging to measure, determining that something is worth measuring or sensing digitally has the potential to drive down costs of data acquisition, as has been the case with the cost of remote sensing imagery.

As a metaphor, we find it helpful to think of the digital measures that are currently in common use for understanding the environment as those found "under the lamp post." As data needs are articulated and the range of domains covered by available geospatial data broadens, we will expand and spread the light of the lamp post and increasingly be able to see what has until now been hidden. In full awareness of our current imperfect vision of what is possible, we endeavor to provide ideas and questions about how emerging frontiers of

13 It must be remembered that social media data introduces issues of sampling bias; for example, a dataset comprised of geolocated Twitter/X posts will underrepresent voices of older users or of users without smartphones. We discuss this issue in detail later in this chapter.

data generation could fit with previously used data sources. In doing so we aspire to catalyze continued conversation and elaboration by others.

Who Has Input and Access to Environmental Data and Metadata from Digital Surveillance?

In this section, we consider how we can improve access to digital data toward an overall goal of "data for public good." In doing so, we consider that with improved technology there may be opportunities to measure previously understudied aspects of the environment.

In considering who has access to data, there are a number of existing constraints. Not all data can be shared without relevant security or legal clearance. For example, some imagery is classified (collected for military purposes) or when released obscures specific features. Data may come at a financial cost or require payment for transformations needed to make it ready for use. Storage systems could create barriers to access or pose additional costs. Some data may only be available for a limited period of time, either after an embargo period or before it must be deleted. Of course, beyond data access, appropriate and informed use of data requires understanding the underlying methodology and purpose, making metadata invaluable.

Metadata Wishlist

As noted by Miller (2022), metadata is data about data, taking the form of structured statements that inform efforts to organize, describe, locate, index, structure, navigate, and manage data resources. Metadata creation and contribution of metadata to repositories are important ways to increase responsible use of digital data (Leipzig et al. 2021). Both those sharing and accessing data and data repositories (e.g., Dataverse; King 2007) will benefit from the skills of data governance experts and data librarians (Lagoze et al. 2006).

Some novel aspects of metadata that we propose below go beyond fixed technical specifications and may need updates subsequent to initial data dissemination. This means that a system that handles versioning is needed, perhaps building on practices developed for GitHub (Crystal-Ornelas et al. 2021).

We note that the data versus metadata distinction can seem arbitrary and, in fact, the same observations may both be represented as data in one database and be summarized in the metadata for a different but spatially overlapping database. The use as data or metadata will depend on the specifics of any given analytical or data management scenario.

Metadata about Original Purpose for Data Collection

Some recontextualization of data can be achieved when including information about the original data collection purpose within metadata. For example, Google imagery and maps have become useful tools for the characterization of the built environment for health research (Rzotkiewicz et al. 2018). Google Street View (Gallo and Kettani 2020) had a primary purpose of improving the spatial and temporal accuracy of Google Maps, for purposes which included identifying commercial locations and increasing advertising revenue. As a consequence, derived data based on these private sector efforts are expected to represent retail settings more accurately than other aspects of the environment such as bike routes. Commensurate with its primary purpose, the image availability and recency vary systematically with socioeconomic conditions (Fry et al. 2020). Researchers may, however, use Google Maps/Google Street View for efficient characterization of the environment at a scale that would not be feasible using field audits.[14]

Other examples in environment measurement likewise benefit from understanding the original purpose and potential for blind spots and bias in the data. Differing susceptibility to bias based on origin can be articulated even among data sources in a similar domain, such as traffic counts computed by a city's bureau of transportation as compared with user-contributed data for smartphone-derived traffic apps such as TomTom or Waze. Whereas a transportation bureau may collect data for meeting reporting requirements or informing intersection changes to improve safety, traffic apps are likely seeking to increase user engagement and associated revenue. Data users could be more cognizant of the data origins and differences in sampling density if these are routinely contained in metadata.

Privatization of data generation intensifies the need for metadata to highlight the reasons for data collection and the related implications for their secondary use. Potential biases, blind spots, or inconsistency may arise related to the original commercial purpose motivating data generation.

Public Open Data projects (e.g., Open Street Map) where "the community" can upload and update data are an alternative that is commonly used in research, especially in locations where government or private sector data may not exist, are not trusted, or lack granularity. In working with such community-generated data, users should be aware of ongoing updates and gaps based on data provider capacity or interest in specific locations (e.g., locations with higher proportions of populations with technical GIS proficiency may have more detailed information; points of interest to specific groups, such as caregivers, may be underrepresented).

Metadata Relevant to Generalizability: Incorporating Structured Information about Communities

Metadata illuminate how data are viewed from multiple perspectives (Lagoze 2001), including attention to the communities represented or omitted.

[14] For example, de Macedo Oliveira and Hirata Jr. developed a system that analyzes thousands of Google Street View images with machine learning to investigate the greenery in a megalopolis like São Paulo (de Macedo Oliveira and Hirata Jr. 2021).

Omission can be the result of structural racism, marginalization, and related social processes that the data creator may not acknowledge or endorse. Thus, a structured requirement for attention to representativeness within the metadata itself is useful, especially if accompanied by an inclusive process. Multiple perspectives can allow a team to draft more robustly and update metadata, documenting a range of cautions to consider when generalizing to a larger set of individuals or geographic areas.

When a community is not represented in data on the environment, there will be missed opportunities to inform decision making (e.g., due to insufficient attention to hazards in the environment or incorrect attribution of harmful effects to the wrong cause), potentially resulting in a community missing out on beneficial place-based or policy changes as a result. As an example where errors in attribution could result in missed benefits, consider how a focus on physical signs of disinvestment could be interpreted as supporting different action strategies. One response might involve attending to the visible signs (e.g., by fixing broken windows or planting trees in deprived neighborhoods); however, even if appreciated by residents, this may fall short of enduring change if the underlying cause is not identified. Alternatively, the underlying disinvestment could be addressed more directly, such as through investment in education or job creation in the same neighborhoods to foster social mobility. Thus, attributing any observed harm to what is proximal and visible risks superficial action; that is not only ineffective, it also diverts attention from alternative actions responsive to the underlying cause and with greater potential for enduring benefits.

Metadata that incorporate a structured ontology (Norris et al. 2019) for social context could help researchers delimit their findings by identifying populations that were entirely or disproportionately excluded. This would facilitate systematic efforts to describe and fill gaps in the availability of actionable knowledge that result from historical and present inequity.

Metadata about Data Sharing and Social License

Data access and sharing practices can also be highlighted in metadata, for example, as described by the Data Use Ontology standard (Lawson et al. 2021). Crucially, this can include who can access data and potentially also how the data have been used over time. Through data reuse ("secondary use") of large-scale data, community data may become divorced from their source context. Structured approaches are needed to reconnect datasets to their originating and dynamic social context. We propose that this can be achieved by bringing community voice alongside application of established guidelines, such as Maelstrom for data harmonization (Eva et al. 2022).

How a dataset is used is subject to change over time, requiring updates to information about how it has been or could be used. Such information includes who has used the data and how data transformations and linkages have been made or could be made. Such metadata would bring users' attention to any distinction between the data in circulation and aggregated or enriched data available upon request. For example, food environment data from the Canadian Urban Environmental Health Research Consortium (CANUE) repository is being released to general users at a higher level of aggregation than the version released to approved research teams (Doiron et al. 2018).

A consideration relevant to stewardship of data is social license, defined as the acceptance granted by a given community or public to a company or organization for a particular activity. Social license could both be described in metadata and seen as a prerequisite to using data about the physical, natural, and built environment. One example of a deliberative process leading to documented social license is the vast network of CCTV cameras in the United Kingdom. These cameras generate data about the environment and human interactions with and within those environments. The use of CCTV for surveillance is considered an extension of the principles of "policing by consent" established in 1829 (GOV.UK 2012a). Permitted use of the resulting data is formalized through the Protection of Freedoms Act (GOV.UK 2012b) which includes the Surveillance Camera Code of Practice. This legislation established the Surveillance Camera Commissioner (updated in 2022 to the Biometrics and Surveillance Camera Commissioner) as an authority to guide the use of this technology for one of the most visually surveilled countries in the world. Systematic attention to social license as metadata is created and shared could promote communication among users and with the original community or its descendants, and also capture efforts over time to reconfirm or revise the agreed terms.

Metadata about Other Data Limitations

Metadata should be designed to include aspects relevant to understanding and communicating data limitations, such as coarseness of the data that potentially masks important variation. Thus, metadata should note quantifiable sources of error and uncertainty. A critical component of data is the characterization of measurement uncertainty (as distinct from true observable variability). Uncertainty may be due to the quality of the measurement itself and may also arise due to sampling error (e.g., gaps in spatial and temporal sampling). No measurement is exact, but measurements may be compared against some practical benchmark or reference value.

Attention to error and uncertainty can aid in not only articulation of limitations but also harmonization and triangulation with other sources. For example, a current measurement with an improved spatial resolution (measurement A) could be combined with a historical measurement with more coarse resolution (measurement B) using comparative analyses (e.g., by linear regression) which itself has some uncertainty (MacEachren et al. 2012). This uncertainty should be propagated together with the uncertainty of the initial measurements. This permits all available data to be used in a way that reflects the reduced certainty of estimated values as compared with measured values. Likewise, interpolation (e.g., filling in missing data that is within the spatial or temporal bounds of the measured data) is another process where uncertainty should be propagated based on both the original measurements and their modeled relationships.

Using Data

Once data are assembled and access is being provided alongside metadata, further steps allow data to be used to generate knowledge and benefits. Importantly, even before turning to strategies for dissemination, we consider how to make GIS-informed knowledge replicable and reproducible (Peng and Hicks 2021). Key steps include integration, analysis, and interpretation refined through multiple perspectives.

Integration

Across disciplines, good practices are needed for data stewardship (Wilson et al. 2017a), including planning for data storage (Hart et al. 2016). Errors are caught and transparency of algorithms improved through practices such as code review (Vable et al. 2021) and code sharing (Peng and Hicks 2021). We note, however, that code that only works on a transformed dataset is not sufficient to allow for external verification, even if the raw data are publicly available; sharing the code (or at least a narrative) that details steps involved in the data transformation and integration can limit redundant work or use of unnecessarily flawed data.

Two major approaches can be adopted when dealing with large amounts of data from different sources, differentiated by whether transformations are done up front or later as needed.

First, a *data warehouse* (Vaisman and Zimányi 2014) is a very large, highly structured database built by extracting, transforming, and loading data from its original sources. The warehouse can be updated periodically through an automated process. Metadata are included. Transformation may include spatial, temporal, and semantic alignment. Data warehouses are designed to enable their users to perform analytical queries (e.g., summarizing data, computing aggregate measures). As such, designers consider the specific analytical needs that the warehouse will support, embedding aspects of their knowledge and intention into the resulting data warehouse design.

In contrast, a *data lake* (Gorelik 2019) is a repository of heterogeneous data collected from multiple sources and stored in its original, raw format. A data lake typically holds a huge amount of data, in a similarly huge variety of different formats. A key advantage of a data lake is that new types of data can be added quickly and with minimal effort. The trade-off, of course, is that the end user of a data lake will need more work to harmonize, format, or link data before starting analyses than is typical for users of a data warehouse. The greater control that the user has over decisions on how to transform or link data may have advantages, however, especially if the analytical needs differ greatly between data users.

Incentive structures that encourage or impede data integration are themselves considered by Balsa-Barreiro and Menendez (this volume), including factors that influence the perceived opportunity costs and benefits (both intrinsic and extrinsic).

Analysis

Emulating study designs that can provide a strong basis for causal inference and pre-specification of analysis plans are among practices whose benefits have been articulated elsewhere (see Dumas et al. and Medeiros et al., this volume). As such, we acknowledge these but focus mainly on challenges related to interdisciplinary collaborations and place-based analyses typically encountered in work with geographically referenced data relevant to environmental constraints on human well-being and behavior.

Expectations for rigor and transparency (such as use of code sharing platforms like GitHub) vary across fields, and collaboration with computer science researchers from an early planning phase can help to ensure adequate resources and capacities. Care is needed for analyses of geospatial data. Current practice ranges from regression approaches to neural network techniques. Widely used programming environments bundled as libraries (such as in R, Python; ESRI, QGIS) may reduce user error and encourage code checking. Investigators focused on causal hypothesis testing may benefit from applying approaches such as directed acyclic graphs to identify confounders or colliders (Pearl and Mackenzie 2018); in other phases of research, undirected exploratory analyses may be more useful.

Even when considering analyses of a single environmental measure, independence assumptions may be violated because spatially closer or neighboring units are similar. Data reduction and modeling techniques can help to quantify or account for this, such as through hotspot analyses or geo-aware clustering algorithms. Highly correlated spatial characteristics are also commonly encountered in datasets derived from geospatial sources, requiring analytical methods to take this correlated nature of spatial measures into consideration.[15]

[15] For example, Dias et al. (2023) were able to find a causal connection between the use of glyphosate in genetically modified soybean crops and infant mortality by taking into consideration the geographical dispersion of the pesticide via Brazilian rivers. Aleixo et al. (2022) were able to develop a machine learning model capable of predicting dengue fever outbreaks in individual neighborhoods of Rio de Janeiro by carefully analyzing the geographical distribution of tens of variables.

Geographic location or other characteristics derived from geographic data may have a relationship to the dependent variable outcome that is nonlinear and, sometimes, completely unpredictable. For example, the distribution of bicycle-based mobility flows within a city is influenced by geography but also by city points of interest, residential, work, and leisure areas, as well as the existing transit infrastructure (Kon et al. 2022). A robust understanding of how multiple characteristics of the environment contributed to observed spatial patterns can be promoted by considering multiple measures, study designs, and statistical analysis methodologies.

Interpretation

Initial interpretation of analysis output by researchers should be informed by known or likely data limitations, including those documented in the metadata, as well as other questions and considerations shown in Figure 3.1. Importantly, this should be a starting point to participatory input from others, allowing potential harms or overlooked aspects to be considered. Involvement of broader communities to inform interpretation as conclusions are reached will be more effective if it is based on a prior foundation of working with communities as true partners across the entire research life cycle. Models for such engagement include citizen or community-based science practices and participatory research methods, including community-based participatory research; place-based work on human use of environments may be an especially good fit for such approaches.

Interactive visualizations are a promising way to allow audiences to select options aligned with their interests and needs, increasing the opportunities for engagement in ways that inform interpretation. There exist specialized information visualization platforms for communicating narratives with a geospatial component, such as ESRI's Story Map platform (Alemy et al. 2017).[16] Interactive geographic dashboards (e.g., InterSCity; see Batista et al. 2016) provide another powerful visual tool capable of giving insights and evidence for stakeholders including health professionals and urban planners.

Audience Engagement to Refine and Disseminate Knowledge

For the knowledge generated from geospatial studies to result in public good, efforts to disseminate knowledge must be tailored to multiple audiences. This requires methods and skills for effective dissemination as well as an investment

16 For example, see the story map produced by the Confederated Tribes of the Grande Ronde (available at https://arcg.is/0v1TO0), which illustrates the geographic history of the various original treaties with the United States and includes a series of interactive maps, narrative text, and other multimedia elements. This story map was one of the winning entries in the 2019 ESRI "Tribal Story Map Challenge."

Enhancing the Use of Data for Public Good: Key Considerations What do data users need to be aware of? What is the optimal description of the data? Who has access and under what conditions?	
Collection What data we collect ("under the lamp post") • Spatial, temporal coverage and resolution • May be agnostic to (e.g., Landsat) or aligned to human-drawn boundaries (e.g., census tracts) • Resolution/quality varies over space and time What data we don't collect • Deep historical context (e.g., changes in use and boundaries) • Relationships between layers • Data settings that have presented logistical challenges (e.g., informal communities, indoor environments)	**Access / Who owns the data** Open / Government / Research / Private sector • Made publically accessible • Licensing/cost • Maintenance / storage control (e.g., servers) Context of data collection • Intended uses • What was not collected/pro-cessed • What was collected but not accessible
Description (what "rides along") Resolution, coverage, quality (and how this varies over time) Collection instruments, methods, and context Post-processing (availability of raw/more granular versions) Access and use restrictions Social license: how data were collected, are being used	**Limitations** What do data not depict and what is incomplete/missing What may be lost in raw to processed conversion What are known problems with representativeness based on incomplete coverage of target areas/population Errors and uncertainty in measured or estimated values

Figure 3.1 To enhance the use of data for public good, the following must be taken into consideration: What do data users need to be aware of? How can data be optimally described? Who has access, and under which conditions, to the data?

of time and money. Below, we summarize the prosocial motivations by group and offer guidance to aid optimal dissemination to:

1. The *general public*: to encourage critical appraisal, build acceptance and support of data-driven activities, and increase the potential for collective action potential, and enable citizens to hold policy makers accountable.
2. *Target populations*: to benefit and empower specific populations or communities and to support relationships between researchers and these communities.

3. *Study populations*: to benefit directly/indirectly and empower those who have contributed to the greater understanding of their environment, to invest in reciprocity, transparency, and accountability of the research process, and to return value through capacity building.

To reach these first three groups, direct outreach efforts will be invaluable, as will working with journalists, creating community-driven data browsers and visualization tools, utilizing popular platforms (e.g., social media, television, podcasts, online courses, museum talks), and leveraging features of social media platforms that promote dissemination (e.g., consider "bots for scientific good"). Additional groups are important to reach:

4. *Practitioners*: to align practice with evidence, to drive practice change and innovation, and to inform interventions and planning. This can best be achieved through existing structures for training/skill development (e.g., continuing education, professional associations).
5. *Policy makers*: to promote evidence-driven policy making, to support critical appraisal and adjustment of existing policies, and to prevent distortion of scientific results. Here, the preparation of contextually framed policy briefs is imperative as well as tapping into existing mechanisms for public comment as well as funder and advocacy organizations' lobbying networks.
6. S*cientific community*: to ensure accountability, to foster collaboration, to generate new ideas, and to receive feedback. Efforts should focus on refereed journals, scientific meetings, etc., and through cooperation with the various professional societies.
7. *Private sector*: to increase knowledge that supports harm avoidance, thereby increasing public good; to increase legal culpability; and to encourage doing good while doing well. Efforts will need to communicate contextually framed information on the potential for harm and interventions for future harm avoidance.

Across all audiences, the following pitfalls need to be considered as strategies are developed:

- Reification of bias
- Limited expertise, experience, and resources
- Competing priorities and time commitments of research team
- Competing demands for audience attention
- Difficulty of contextualization
- Temptation to oversimplify or overhype
- Insufficient accessibility of language or terminology
- The elongated timescale of science generally or of a given research project specifically as compared with the "media cycle"

To engage people who could be affected by physical and built environment characteristics,[17] it is important to distinguish between the individuals directly engaged in a study (participants), other residents/users of the studied settings, and broader populations spanning other settings to whom the research conclusions may be generalized.[18] Investigators need to adopt specific strategies to reach groups affected by their research, address any risks for group harm, and ensure transparency through communication. Working with journalists, and science or data journalists in particular, can be a strategy to reach multiple audiences and foster trust and recognition for scientific endeavors and advances in our fields.

Yet not all dissemination efforts will be equally effective. Brevity is valued. Timing of dissemination can determine receptivity to research findings. The optimal time to speak to the media or to publish op-eds based upon research findings may not be when the research has been completed/published, but rather when a relevant issue arises within the public discourse or policy agenda. For example, emerging attention to wildfires may present an unanticipated window of opportunity to disseminate research on health effects from air pollution or land management decisions. Complementary engagement may be considered across multiple formats (e.g., reaching policy makers through both producing a two-page policy brief and contributing comments on a public notice with obligatory response to relevant comments). Beyond the efforts of individual investigators, there may be a role for professional societies to scan for relevant actions (such as public comment periods for certain subjects) so that membership can be made aware of relevant rulemaking.

Nuances and caveats typically reserved for communication to a specialist audience may be important to translate to a broader audience in a concise and accessible way that conveys which findings are fragile versus robust; oversimplifying the message may be expedient but could later backfire as subsequent and seemingly conflicting findings impede understanding or undermine trust. Without appropriate training, dissemination opportunities can be mishandled, communities offended or harmed, misinformation reinforced, and disinformation propagated. Inaccessible language can impede effective communication that meets audience needs, including due to unintended connotations of some commonly used scientific terms (Somerville 2012).

How research should be disseminated will also depend on characteristics of the researcher and audience. In some instances, rules established by funder

---

17 Often, those affected by the environment are discussed as "stakeholders." We note, however, that there was not agreement within our group about the utility and appropriateness of that term.

18 The known limitations of such generalization are important to articulate, though generalization beyond the observations included requires strong assumptions. A key assumption for generalizing claims about an environmental effect on human health, for example, would be that there is no effect modification (even if unmodeled) that results in a different strength or direction of association, or that any effect modifier is not differently distributed between the measured and target population or settings.

organizations or governing the organization where the research is conducted may shape what is allowed to be communicated to elected officials. Other potential obstacles may be institutional fear of upsetting funders or board members. In order to reach the private sector where the aim includes establishments of culpability, contextualization of the information may be warranted such that not only possibility for harm is demonstrated but interventions for future harm avoidance are proposed.

Ultimately, a dissemination process that reaches a broader audience is critical to professional advancement of researchers individually and stands to benefit the field through influence on funders and policy maker priorities (Dudo and Besley 2016). Dissemination should be wide ranging, encourage multidirectional discourse, and may frequently extend beyond the conclusions of a particular project (e.g., massively open online courses [MOOCs] or other formats for continuing education, TED talks, science museum events, podcasts, engagement on social media platforms, and other ongoing public outreach roles). Specific cultural contexts may also create other avenues to move science messaging to be more resonant with popular culture forms. For example, a Canadian public service announcement designed to reach a Punjabi-speaking population used video featuring bhangra dance and a well-known Indo-Canadian actress to enliven delivery of the message about pesticide safety and laundry instructions (Murphy and Nicol 2010).

Critical consumption of knowledge warrants attention across all sectors of society and all career stages as the methods used continue to advance (Few 2019). Especially with complex topics, a craving for simplistic solutions and a rush to attribution may lead to trendy, overhyped, and misleading science. To avoid dissemination of simplistic conclusions which can undermine public trust, attention is needed to who is engaged in research, how we train investigators, and system-wide incentive structures.

Future Directions

For digital environmental measurement and research to advance in ways envisioned as beneficial to humanity, attention will be needed on inclusive team formation, cross-disciplinary training, and leverage points for sustaining change.

Team Formation: Include Diverse Perspectives Early

The need to include diverse perspectives is a unifying theme throughout our vision for determining what to measure about the environment, documenting how and why, and bringing intentionality to the generation and dissemination of knowledge. A team science approach offers the potential to combine the strengths of multiple fields. For place-based digital ethology, teams should consider including computer research scientists, geographers, and others skilled in working with geographically referenced data (as highlighted by Brinkhoff, this volume); those with expertise in the building of the environment such as urban planners and architects; and domain scientists with expertise in how environments affect the people who live or spend time in them (e.g., environmental health, environmental psychology, urban health). Bringing together these multiple perspectives early recognizes the depth of expertise and limits the risk of redundant effort rediscovering what is already established in another field. Further, teams and partnerships that bring together a diversity of disciplinary, identity-based, and lived experience may be especially crucial to scrutinizing unsupported assumptions and addressing shortcomings of conventional practices.

Interdisciplinary researcher teams are poised for effectiveness by encompassing knowledge about standard and emerging practices for measuring and investigating the environment, including promising practices from across disciplines in data stewardship, responsible use of data, and supporting audiences in the critical consumption of knowledge produced with data. Skills and roles that allow at least some team members to work with nonacademic partners (e.g., inviting input and exploring social license for use of data directly with communities, cultivating audience ties including through interactions with media) are most beneficial if incorporated from the earliest stages of collaboration planning.

Training Needs to Leverage Place-Based Digital Ethology for Good

Cross-disciplinary training was identified as a priority to support working across silos, as well as building networks that span disciplines, allowing for rapid dissemination of promising approaches and ideas. Cross-disciplinary training relevant to digital ethology should provide a foundation for future collaboration among those with knowledge of geographic settings (e.g., urban planners, architects), physical and mental health in humans (e.g., medicine, public health, psychiatry, neuroscience), social processes (e.g., sociology, anthropology), technology to collect and use digital data (e.g., computer science research), and outreach to partners and audiences (e.g., community-engaged research, communication and implementation science).

As no one person can reasonably cover this full range of skills and expertise, disciplinary silos impede collaboration. To foster awareness and appreciation of other disciplines among disciplinary specialists, it will be necessary to develop models for cross-training at different career stages. This may take the form of courses offering basic skill building or a primer for topics outside of one's own discipline. There can also be value in dual degree or exchange programs, for instance embedding journalism trainees within science teams, which promotes two-way learning.

Even among fully trained professionals who are active in the field, ongoing learning opportunities are needed. Workshops and discussions across disciplines (such as the Forum which resulted in this volume) may create the opportunity to discuss and mitigate potential harms of siloed research (e.g., awareness of legal, ethical considerations that are relevant to avoiding harmful consequences of technically possible research).

Setting the Stage for Sustainable Change

Given the challenge of competing priorities, attention is needed on upstream structures and incentives that can sustain change in project planning and implementation, dissemination, and other scientific professional practice. Here we present some preliminary ideas concerning strategies for organizations that provide research funding, journal editors, and academic institutions.

Funders

Organizations that fund research teams and projects have leverage to incentivize sustained change. Some funders already incorporate planning grants, dissemination requirements, or other approaches designed to foster inclusive team formation and knowledge generation that reaches those who can take action to reduce harms or create benefits. Since harvesting metadata on a large scale can be challenging (Lagoze et al. 2006), standards and citation rules for meta-analyses should be supported and incentivized by funders. Funders can require and financially support the generation of user-ready GIS data repositories, such as efforts supported currently by the Lacuna Fund. This may require special initiatives to generate such repositories similar to one designed to address environmental influences on child health outcomes (Smith et al. 2018), or special funding that would be provided for the work necessary to prepare and add data to repositories at the end of studies. Complementary to traditional seed funding to incubate a nascent project, funders may consider "harvest funding" to amplify the ability of teams to articulate, document, and disseminate both insights and data from a project as it concludes.

Journal Editors

As peer-reviewed articles continue to be an important signal of research reputation used in making funding and promotion decisions, journal editors can set the stage for improved practice through required reporting and flexible formats.

Journals increasingly ask authors to include statements about data sharing or the involvements of affected populations (e.g., the British Medical Journal required reporting on patient and public involvement) (Boivin et al. 2018). The acknowledgment of acceptable social license could, in the future, be considered a requirement for publication with digital data of certain types, in a way similar to the ethics assessment carried out by Institutional Review Boards.

Beyond current reporting requirements for articles, journal editors can create spaces for published datasets and metadata with shareable digital object identifiers and promote citing these as a way to connect related work and ensure credit. This could take the form of an article type or follow the model of a dedicated journal as illustrated by the *Nature* journal *Scientific Data*, which includes examples such as a description of global emissions mapping data (Weng et al. 2020).

Academic Institutions

Academic institutions can update promotion and tenure guidelines and processes such that they reward the investment of time in activities described above, from building of database structures and repositories to developing trusting partnerships with communities and other knowledge dissemination audiences. Institutions may decide to set up or further invest in structures to facilitate cross-disciplinary collaboration in the form of Institutes or Centers. These can foster the necessary policy-oriented communication capacities and practical collaborative infrastructure that allows for feasible incorporation of skills that any given research project may not have sufficient funding to sustain (e.g., web designers to provide audiences with access to knowledge synthesis and interactive mapping).

Conclusion

With the considerations described in this chapter, we envision a world in which digital data on the physical, natural, and built environment are useful and used for public good. Data volume, scope, depth, and quality are likely to increase in the future. We are already seeing multiple benefits, although missteps may be an inevitable part of our path forward as the field evolves. Boundless potential gives us optimism for appropriate use, while recognizing that attention is needed to amplify responsible use of digital data for knowledge generation, equity, and other public benefits.

Acknowledgments

The authors would like to acknowledge funding from Dr. Ernst Strüngmann Foundation; inspiration and feedback from those who participated in the Digital Ethology Forum, and assistance from Julia Lupp, Marta Gruszka, and all at the Ernst Strüngmann Forum who set the stage for and supported our work.

This is a section of doi:10.7551/mitpress/15532.001.0001

Digital Ethology

Human Behavior in Geospatial Context

Edited by: Tomáš Paus, Hye-Chung Kum

Citation:

Digital Ethology: Human Behavior in Geospatial Context
Edited by: Tomáš Paus, Hye-Chung Kum
DOI: 10.7551/mitpress/15532.001.0001
ISBN (electronic): 9780262378840
Publisher: The MIT Press
Published: 2024

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open

The MIT Press

Series Editor: J. R. Lupp
Editorial Assistance: A. Gessner, C. Stephen
Lektorat: BerlinScienceWorks

The book was set in TimesNewRoman and Arial.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-0-262-54813-7

10 9 8 7 6 5 4 3 2 1