5

Integrating Knowledge from Individual Data to Population-Level Data

Claudia Bauzer Medeiros, Hye-Chung Kum, Sven Sandin, Cason D. Schmit, Kimberly M. Thompson, Henning Tiemeier, and Kimmo Kaski

Abstract

Knowledge integration permeates all scientific endeavors, which increasingly depend on interdisciplinary collaboration as well as on combining data from multiple sources and knowledge domains. Advances in digital ethology progressively rely on knowledge integration, which is enhanced, but also hampered, by the large volumes of heterogeneous data that need to be considered, the multiple aggregation levels to be considered, and the human expertise involved in answering research questions. Though considerable research efforts have focused on leveraging knowledge creation through data integration, many challenges remain. This chapter identifies and investigates some of these challenges, pointing out strategies toward the generation of knowledge while bearing incentives and barriers in mind. To investigate human behavior in the built, social, and/or natural environments, for example, what kinds of considerations exist when integrating individual and population data? Are big data an asset or a hindrance to such integration? Why should (or should not) researchers go through the effort of curating, documenting, and integrating multiscale data?

First and foremost, despite all the technological advances, human judgment remains a key factor in the selection of datasets to be integrated, in monitoring and validating the integration process, as well as in interpreting the results to extract knowledge. Moreover, quality factors, such as reproducibility or robustness, must be considered at all stages: data selection, design and implementation of the integration process, and result analysis. Appropriate documentation of data and processes must be ensured for fairness and reproducibility, and metadata quality is essential for sharing of data and processes. In conclusion, ethical and legal considerations interact in many complex ways, but there exist paths to move forward and overcome the barriers posed.

Introduction

Incentives behind Multiscale Knowledge Integration

Opportunities to integrate individual- and population-level data[1] to approach innovative research questions continue to expand as researchers recognize the benefits of interdisciplinary scholarship to better understand human behavior in the context of built, social, and natural environments. Although similar to the value of research partnerships and collaboration within domains of expertise, the need to combine individual data, often aggregated at multiple levels, to address some types of research questions, usually expands the number and types of disciplines and experts required to engage cooperatively in the process. Such efforts thus promote cross-fertilization of ideas and improve interdisciplinary understanding in the process of reaching shared insights.

Research projects that integrate data can also uncover new hypotheses as well as novel lines of inquiry, provide better insights about existing hypotheses and theories, and refine our understanding of observed phenomena, driving us to dig deeper to explain any differences in outcomes observed. By investigating some questions using integrated datasets, analysts can increase the ecological validity of findings and the generalizability of results. Recognizing the connectivity of different domains can provide further understanding of the structure and mechanisms operating in the complex human systems in which we observe patterns of behavior. Moreover, by linking individual and population data, the insights that exist primarily in one domain may take on broader relevance and importance.

The development of technological resources, the appearance of new platforms, and increased availability and access to digital data, including "data repositories," "data lakes," and federations thereof, contribute to this expansion of opportunities. Code repositories (e.g., GitHub), data repositories (e.g., NASA's satellite image repository), and registries of repositories (e.g., re3data.org) facilitate the identification of datasets and analysis code in different domains, which can then be reused or repurposed to answer new research questions. In addition, the development of ontologies[2]—such as LOINC for health-related measurements (McDonald et al. 2003), the human phenotype ontology (Robinson et al. 2008), and gene ontology for bioinformatics (Gene Ontology Consortium 2018)—help to support the translation and linking of data across datasets (Kamdar et al. 2019) as well as the characterization of geo-related scenarios (Huang et al. 2019).

Researchers who integrate individual and population data can benefit from using existing data (e.g., acquiring data much more quickly than collecting new data, saving time and money) and their reuse of the data may increase the value

1 Population-level data is the result of aggregating individual data into groups that abstract some of the individual-level properties to run an analysis.

2 We define an ontology as a data structure that organizes some field(s) of knowledge, by connecting terms to their meanings, usages, and relationships.

of existing data. In addition, reuse of previously collected data may be the only means of acquiring historical information. The recognition of researchers engaged in such efforts may lead to their identification as "connectors" and "translators" across disciplines who "think big," and increase the visibility of their domain-specific contributions in other domains. The development of broad expertise (i.e., across more than one domain) and increased professional visibility may further provide rewards in the form of increased funding opportunities, attraction of students and/or other collaborators, influence, and greater dissemination of results. The process of working on interesting and challenging research questions that require the integration of datasets and collaboration across domains can provide fun and intellectually challenging opportunities for learning with others about interdependent and multiple factors affecting outcomes that otherwise cannot be observed.

The Data Deluge and Research Questions

Integration of knowledge is always prompted by research questions, some of which can only now be answered thanks to the so-called data deluge (which, at the same time, poses new challenges to eliciting these answers). Technological advances in data collecting and processing devices have allowed massive availability of data on human behavior and activity at individual, group, community, and population levels, in different forms and storage organizations (e.g., databases, repositories, data lakes, and others). An estimate published by The Economist (2017) claimed that, by 2025, data generated per year will have reached an order of 170 zettabytes (zettabyte = 10^{21} bytes), and that it would take some 450 million years to transfer this amount of data from one place to another using the current data transfer technology. According to the same source, some 80% of these data are privately held or in hard-to-access forms; only 20% are found in various kinds of records (e.g., social or health data in registries) that are more accessible and regulated. Indeed, a wide variety of open big data sources provide select information on individuals and populations, summarized at different geographic or administrative levels (e.g., municipal, district, state, city, and country level) and by specific characteristics (e.g., age, education level, behaviors), as well as a myriad of files on conditions associated with the built, social, and natural environments (e.g., transportation, social networks, weather). The handling of such heterogeneous sources of information poses a number of conceptual and technical challenges.

Indeed, integration of data is arguably an important step in the attempt to develop new knowledge on human behavior and its constraints. While some platforms function primarily as repositories for data access, others support data collection, curation, security, anonymization/pseudonymization, as well as software tools, methodologies, and algorithms.

The key construct of science, namely the *formulation of a research question*, involves a long path to extract new knowledge through integration of

various sources of data. The nature of the question will determine the choice of research method. For example, is the research question guided by an existing theory or *hypothesis* to be tested, or is the research question related to empirical *exploration* without a hypothesis? In the exploratory study approach, the focus of the research question is on the properties of the system concerning its structure, how it functions and responds to different external conditions, in that order, using the methods of data analysis, computational modeling and simulation, respectively (Kumpula et al. 2007). This exploratory process to generate knowledge from data can be viewed as a continuum such that the data analysis primarily leads to insights about structural properties or correlations between entities. After this, additional studies may provide and would be required to obtain further insight into the functions or processes of the system. The progression from poorly structured mental models to mechanistic models that capture causal and dynamic relationships in physical and/or social systems may then support simulations to answer "what-if" questions and/or predictions of likely outcomes of future experiments or interventions (Saramäki and Kaski 2005). Learning and knowledge generation is not a linear process; rather, the knowledge obtained at each step may require going back to any of the previous steps, for example, to acquire more data or to change the modeling approach. In addition, individuals and their interactions with social, built, and natural environments (including the technology) continue to change over time, which means that our understanding of human behavior and our world also continues to evolve.

Research methods can take advantage of a number of well-established statistical analysis tools that are readily available for drawing inference from large amounts of data. Computational tools may use a phenomenological approach, a statistical approach, or a holistic approach that combines both. The phenomenological approach uses methods associated with, for example, network science and modeling to analyze links between entities, functions, or processes, in search of plausible mechanisms to understand the formation of human social networks and dynamics of human behavior in them. Statistical approaches are an integral part of data science, and cover statistical analysis or modeling, in which various machine-learning methods may play an increasing role for regression, clustering, and inference. Integration of data from various sources points, however, to the need to develop novel computational approaches, methods, and algorithms to get more detailed insight into human social behavior and population-level phenomena; regardless of the approach, human expertise is generally required (discussed further below).

As West (2017) pointed out in his data-driven studies of human social systems: "The underlying laws of complex social systems are not known, yet, but they show regularities so there must be governing principles." This, in turn, signals the relevance of integrating knowledge from data in multiple scales, collected at the individual, group, community, and/or population level. Our discussion begins with an overview of the main steps and approaches

for this kind of integration, and briefly delves into identifying questions while stressing the importance of human intervention. We then discuss some kinds of studies that lead to claims that may be made as a result of integration and analyze the soundness of data to support these claims. We address quality issues through the integration process and look at some of the factors that may hinder integration activities. Finally, we analyze ethical and legal questions that arise during integration and suggest future research directions for consideration.

Knowledge Acquisition through Data Integration: From the Individual to Populations

The integration of data to acquire knowledge can be seen as an iterative process that comprises four interrelated steps:

1. Defining and acquiring the data to be integrated.
2. Curating and preprocessing as needed.
3. Performing the integration through a number of strategies.
4. Performing computational analyses on the results of the integration.

This process may require backtracking to re-execute any activity, with potentially new data or strategies that may, for example, indicate the need for alternative or new data sources, or additional curation, or alternative integration methods, in which case one or more of the activities will be repeated until the researcher is satisfied with the result. In the context of place-based digital ethology, integration combines individual-level data (e.g., tabular records from administrative health databases; see Sandin, this volume) with area-level data about physical, built, and social environment (see also chapters by Smith, Lovasi et al., and Weigle et al., this volume).

Given a specific research question and datasets, results may be different and lead to distinct (and even contradictory) conclusions and claims depending on the choices made during steps 2, 3, and 4 and their interactions. This points to the need for separating the *concept* of integration from the *algorithmic strategies* used, as well as from the kind of *underlying physical storage* mechanisms (e.g., are the data in warehouses, repositories, or data lakes; are they provided through a single site or via a federation of sites or institutions). Here we will concentrate on concepts and high-level strategies and ignore computational implementation issues.

Many Names, One Goal: Acquiring Knowledge through Multiscale Data Integration

The integration of individual- and aggregate-level (in our context, most often area-level) data to derive new knowledge has been discussed in different

disciplines and research domains under a variety of names and contexts. It is sometimes called "multiscale data integration," in which the scale may be associated with the geographic space (Cui et al. 2022), but may also refer to different scales in human biology for health studies (Phan et al. 2012). Multiscale integration may interweave the data with the models that were used to produce data at different levels of complexity (Peng et al. 2021). Other names include "multilevel analysis" (Snijders 2011), "combination" of individual and aggregate data (Haneuse and Bartell 2011; Mezzetti et al. 2020; Raghunathan et al. 2003), "linking" (Paus et al. 2022), or "merging" (Gaubatz 2015; Hernández and Stolfo 1998) datasets.

Regardless of the name used, the ultimate goal is to acquire knowledge and get new insights about relationships among the real-world entities being represented by the data so that we can answer research questions. Through integration, new relationships emerge (Jo et al. 2014; Monsivais et al. 2017). Relationships may be explicit, such as those between "attributes"[3] (data properties) associated with a particular geolocation in multiple domains (e.g., rural, urban, demographics, records of social or medical services). Geolocation can be further enhanced by related information, such as temperature and length of daylight (Kovanen et al. 2013), obtained from open national meteorological and geophysical registries. Nonexplicit relationships (e.g., behavior patterns in a social network) can be obtained algorithmically by using, for instance, machine-learning techniques (see section on the Importance of Human Judgment in Data Integration).

Though ideally the research question at hand should decide which data source to use (step 1 of the iterative process), other considerations, like convenience and data availability, might also influence the selection of data. Regardless, the data sources chosen will have consequences on all analyses performed, statistical and scientific inferences, as well as which claims and conclusions we are able to draw. It is crucial for researchers to be explicit and clear about what they are proposing to measure and combine, and to ensure that the data they use are relevant to the task at hand. They must also understand and acknowledge the limitations in the data, analysis methods, and strategies (see Lovasi et al., and Kum et al., this volume).

Metadata

In parallel, researchers are often concerned about issues such as data access (how can I get the data I need; how do I know whether it exists, and where), data provenance (where did the data come from, how were the files produced,

[3] An attribute refers to a field in a file record and is sometimes called a property or feature, depending on the research domain. The term usually refers to textual or tabular files but may extend to nontextual files. Examples are the name of a person in a table, the coordinates of a region covered in a satellite image, or the amplitude of a sound wave.

and by whom), and responsible data management[4] as a whole. When looking for data that may be used in a research effort, metadata[5] are a valuable resource, since they describe a file and give information on authorship, provenance, quality, access rights, as well as other fields that may help in understanding the context in which sharing and reuse are allowed. (For a discussion on metadata and its value, see Lovasi et al., this volume.) Data registries, repositories, and federations thereof always contain catalogs of metadata—albeit of varying quality—that help to find the datasets of interest therein, in line with FAIR principles (Wilkinson et al. 2016). Given all these roles played by metadata records, metadata quality is a serious issue, often ignored by researchers. Issues related to quality are discussed below; see also Lovasi et al. and Weigle et al. (this volume) for further discussion on provenance and associated quality issues for data derived from interactions of humans with and within the built and social environments.

Integration Strategies

Integration strategies (step 3) are high-level procedures that can be applied to combine data from individual to multiple aggregate levels (e.g., area levels with different spatial granularities). Each strategy is refined depending on the data being integrated as well as quality and provenance issues. The actual computational implementation requires taking additional factors into account, such as performance, data volume, data placement, and even privacy concerns. The main groups of strategies relevant to the discussion in this chapter include:

- *Fusion* combines datasets into a single one by joining them along common attributes; this is often applied to tabular data (Bleiholder and Naumann 2009; Gagolewski 2015). Overlay is an example of a fusion technique in which the data to integrate are images whose contents, in digital ethology, are combined based on geolocation (Tsou 2004). In this case, the result is a compound image, in which each pixel corresponds to a value that represents a combination of the values of pixels of the overlaid images at that location. Individual- and population-level data can be fused, based on geolocation, when each individual is connected to a place; aggregate-level data refers to a polygon that contains

[4] For a comprehensive set of resources and standards on research data management and governance, see Research Data Alliance (https://www.rd-alliance.org/).

[5] Metadata are data that describe the contents of a file to help find and characterize it at a high level so as to preclude having to open the file to see what is inside. Metadata are always textual records. Metadata standards are domain- and research-group dependent and define which are the attributes of these records. A metadata record describing a satellite image includes attributes such as information on the sensors that captured the image, the date it was taken, and coordinates covered. A metadata record on a questionnaire applied in qualitative research may contain information on how interviewers were trained, or even a pointer to a particular term of consent.

the place, for example, as described by spatial join integration techniques (Brinkhoff et al. 1994).

- *Linkage* typically does not fuse datasets; rather, they may be kept apart but linked together (e.g., using tables) to form clusters of information about a given entity. An example is record linkage, also called entity resolution, which corresponds to recognizing different manifestations of the same entity in different files, and connecting their records based on an identifier, such as geolocation. Each integrated entity becomes a cluster of records, each of which addresses a specific kind of information, from individual to multilevel aggregates (e.g., income tax, criminal record, employment history, hospitalization history, census sectors). Linkage when the identifier is not unique or does not exist is a research problem. Herzog et al. (2007) treat a different aspect of this problem, and Kum et al. (this volume) discuss some approaches to dealing with privacy in record linkage when using individual-level data.
- *Semantic integration* connects separate files via ontology links (Noy 2004) by examining the semantics of their contents. Individual- and population-level data are connected together by the concepts they have in common and, in our case, considering geographic characteristics (Huang et al. 2019). Semantic integration often results in large graphs with millions of elements. Social networks are often processed using semantic integration mechanisms, in which clusters arise due to, for example, common behavior, expressed beliefs, or discussion topics (see Weigle et al., this volume). For a discussion of behavior patterns in digital ethology and associated data, see Dumas et al. (this volume).

Since integration starts by trying to identify commonalities across the files to be integrated, it is important to assess whether all files are minimally compatible. In particular, a combination of the above strategies may need to be applied, depending on the kinds of data types to be integrated (e.g., textual data, images, data streams, graphs of social networks, surveillance videos). The following is a succinct set of questions that need to be asked to identify commonalities among two or more datasets to facilitate integration:

- Is there any common set of features/fields/attributes/properties[6] that will allow, for instance, spatial or temporal units to be integrated, or the associated entity or characteristic to be represented, such as spatial extent, geographical characterization, measured variables, or category (e.g., land-use or socioeconomic factors)?
- At what granularity were attributes collected (e.g., meters, census units, years), and how were they expressed (e.g., frequency, intensity, time it takes to do something)? Are they qualitative or quantitative? Is there any kind of conversion between qualitative and quantitative that will

[6] Distinct research domains use these names to mean the same thing.

allow meaningful comparison? What does "near" mean in a location-based system, or "frequent" in medical reports? For a discussion on spatiotemporal granularity, see Lovasi et al. (this volume).

- Are these common sets of attributes compatible: Do they cover the same or overlapping spatial regions? Do they refer to the same or overlapping temporal windows?
- Did the datasets to be integrated already exist, or were they collected for the research effort? Are they raw, or derived, or synthetic? If derived or synthetic, what code was used to generate them? Note that synthetic data are common in situations where raw (real) data are hard to get, such as to protect individual privacy (Arora and Arora 2022).

The answers to these questions may indicate the need for data curation (a step toward increasing quality) or preprocessing (e.g., to fill in blanks or missing values, or to perform conversions). Examples of preprocessing involve converting temporal or measurement units, or aggregating/disaggregating records (e.g., transforming schools into school districts). Preprocessing may also involve additional methods, such as transforming images, sound, or videos into arrays that encode them in a more compact way (also called "descriptors" in image or sound processing).

An example of the need for such questions when integrating individual and population-level data is the so-called "modifiable areal unit problem" (MAUP) (Manley 2019). In a MAUP, the level of aggregation (e.g., administrative or census units) and the shape of the units will affect integration and subsequent analysis. Indeed, there is often an underlying assumption of population homogeneity within each aggregation unit, which is not always the case. Here, it is not enough to perform linking, or fusion, or semantic integration, without understanding the fitness for use of the individual and the population data.

Importance of Human Judgment

The consequences of different approaches to integrate individual and population data depend on how and when the integration occurs. The process affects the variability and clustering of the data ultimately used in the analysis (e.g., as in the MAUP situation just described) as well as the transparency in the judgment of the investigators involved in the process. Whatever approaches are used, substantial human judgment is involved, and domain expertise is essential (see the discussion on the importance of "human in the loop" by Kum et al., this volume).

Prior to the development of big data algorithms, analysts traditionally combined data through a process that involved the identification of data for potential aggregation and undertook a careful process of data curation with domain expertise to combine only what was needed. More is not better in these

situations; rather, integration typically required pulling together only the relevant parts of the data based on domain expertise.

In comparison, in the newer machine-learning approaches, the data selection process can include a wide range of data associated with the research question, but some of the integration relationships may not be known or established, or the association with the research question can be challenged. Here, more may be better. This is because the first step of data integration for these approaches is to get as much (potentially) relevant data together as possible, and then follow this step by data reduction and correlational analyses that identify relationships. In this process, the researchers face challenges to explain the data, and this process can lead to deeper investigation to identify causal relationships and sources of variability. Here, modern statistical and computational methods and techniques (e.g., machine learning) can elicit relationships that would not be identifiable in the more traditional knowledge integration processes.

In either scenario, the role of the domain experts is important for pulling together as much data as possible, for data reduction (Mattingly et al. 2019), or in understanding and validating the emerging relationships and results.

Some Typical Study Types and Associated Claims

Claims can be contextualized by the kind of studies with which they are associated, such as:

- *Descriptive studies.* These studies typically do not involve any elaborate claims and may be free from more formal statistical analyses and rely more on basic statistical methods (e.g., mean, median, distributions, confidence interval) but, ideally, include data representative of the target population. The goal is to describe what is being observed.
- *Estimating studies.* Can be seen as a deeper and more focused descriptive study, usually including formal statistical methods and inference quantifying an estimate of interest. The claim would relate to estimated effect size and magnitude. Their goal is to go beyond a simple description to look at relationships.
- *Hypothesis testing studies.* A study testing a prespecified scientific and statistical hypothesis, alternatively supporting equivalence of some kind. This study would include statistical methods; inference, estimation, and description would be included as supporting information. The claim would be very specific—for example, declaring presence of a difference.
- *Causality and mechanistic studies*. While hypothesis testing studies can be based on group-level data using population averages and correlations, this would be less likely for a study concluding causality where we would require a high degree of support from the data in order to claim that an association is a measure of a truly causal effect and not

driven by confounding factors or biases from mediating or unbalanced moderating factors. Patterns supporting a mechanism and causal effect include patterns across time or age, or dose response.

- *Normative studies.* These are studies that seek to make claims about whether observations are consistent with available prior observations. For example, normative claims using individual physiological data are familiar to most people (e.g., blood pressure is normal, or low or high relative to the reference range). Although there are no universally standardized reference ranges for human behavior, in human development some reference ranges exist (e.g., growth curves or developmental milestones, some of which vary by country). Similarly, psychologists and psychiatrists categorize some types of mental disorders and cognitive heuristics that impact behavior. In economics, there is a focus on observing human behavior by understanding choices/decisions, often in the context of preferences revealed by participation (or not) in markets.
- *Methodological studies.* These are studies that focus on demonstrating the functionality of algorithms and tools that make claims about the utility of the algorithm/process. These studies often start from defining an important computational problem that is useful to addressing human behavior if the problem can be solved with some algorithm. Here, results about human behavior may not be novel, but it is still important to demonstrate usefulness of the proposed methods through real case studies.

Evaluating the Use of Data to Support a Claim

There are at least two complementary approaches to evaluate how integrated data are used to support a claim. Both approaches address bias in research: one relies on statistical methods to check whether bias in integrated data produces biased claims; the other concentrates on methodological aspects in data collection and integration that may lead to misinterpretation of results.

In the first case, a major statistical approach is the analysis of confounding variables; namely, those in which external factors of no interest may influence integration outcomes, and thus the claims. Consider, for example, the use of environmental data to qualify the claim that "people work less than normal when it is hot." For this research question, and associated claim, consideration of the potential role of the omitted variables may explain the phenomenon (e.g., school vacations occur during the summer). Thus, temperature may not be the primary driver of this behavior, but rather the fact that children are out of school, which encourages families to take vacation and work less at the same time. In addition, high temperatures may spur government regulation when schools are open. In a more general sense, when integrating data to support claims, the analysis needs to include a process of not only validating the

accuracy and reliability of the data, but also the role that different variables may have on the research question at hand, and understanding the context in which the question is posed. A sequence of models and analyses may be required to test and evaluate the outcome under different assumptions related to the role of the variables as relevant with respect to discussion of a direct effect on the outcome or for the indirect effect on the outcome (i.e., as a valid indicator of something for which we do not have sufficient data). As always, analysts must remember "garbage in garbage out," and that quality issues must be considered at all research stages. The increasing use of machine learning as part of the analysis process has spurred the development of a wide range of statistical methods to check data and analysis bias on integration results (Ntoutsi et al. 2020).

The methodological approach (Brazhnik and Jones 2007), instead, is guided by questions on steps 1–4 of the integration process presented above. These questions can only be answered when the datasets and the steps were appropriately documented, in particular using metadata. The first set of questions concerns step 1: data selection. Was the choice of datasets to be integrated appropriately justified? Did these datasets already exist, or were they created for that research effort? If they existed, why were they chosen, and how were they found? Were they included just because they exist and are big (a self-justified choice)? Are they representative of the phenomena they purportedly describe?

Additional questions refer to how these datasets and their integration were documented. Are they appropriately described as to the spatiotemporal context in which they were created/collected? Are all units that characterize them stated? Are there standards against which we can analyze the suitability of the integration strategies adopted? What were the integration strategies performed, what kinds of preprocessing was conducted (e.g., curation, unit conversion)? Are they overly described (too many variables), requiring integration via multicriteria decision analysis? Or are they under-described, which would result in a poor analysis process and unsupported claim?

We now proceed to a discussion on quality, which is directly associated with all aspects previously discussed in this chapter.

Quality Considerations

Quality considerations permeate the knowledge acquisition process, from stating the research question to the final claims. During integration, quality checks apply to the four steps previously mentioned: data collection, curation and preprocessing, data integration, and computational analysis (and the selection of analysis methods and datasets). Such checks apply to the data (e.g., appropriateness of choice) as well as to the processes involved in integration and analysis. Which quality factors should be applied, and how should they be evaluated? Here, one must remember that data quality is also defined as

"fitness for use" (de Bruin et al. 2001) or "fitness for purpose," so that quality factors and their evaluation have to be specified relative to the research framework and acceptability of the results within that framework.

These factors, often called "quality dimensions" (Fox et al. 1994), include robustness, trustworthiness, generalizability, and reproducibility. When talking about big data, the term "veracity" is sometimes related to quality; namely, to which degree results or processes represent what they are supposed to. Weigle et al. (this volume) present many examples of quality dimensions associated with social media data, such as cohesion or coverage.

Here, we discuss how the integration of individual and population (area-level) data impacts the robustness, reproducibility, and generalizability of results. In particular, we offer recommendations on how to improve the trustworthiness and generalizability of results.

Robustness of associations and relationships found in integrated datasets will depend on modeling practice, measurement error, and sampling uncertainty. In all population studies, the robustness of associations is influenced by factors such as the basic model choice (e.g., structural equation vs. regression models), the degree to which model assumptions are met (such as the normal distribution of the outcome in linear models), and modeling choices, such as the number of knots in a spline regression (Klau et al. 2021).

In large-scale social media studies or large registries, some aspects of modeling are less impactful if the sample size increases. For example, a certain deviation from the normal distribution is more likely to influence results if less than a few hundred individuals are studied; very large datasets are often more robust to these assumptions (Schmidt and Finan 2018). The impact of many other model assumptions is independent of sample size. For example, any aggregated data will have to be analyzed accounting for the clustering of individuals in the study. Although standard practice, this is sometimes overlooked, in particular if the exposure of interest is based on individual-level data or if only confounders were obtained from aggregating data.

Measurement error is often only superficially discussed in datasets resulting from integration of population and individual-level data. It can occur in exposure, confounder, and outcome measures, but has been shown to impact results even if only occurring in one variable and even if datasets are large. Although measurement error is often nondifferential (i.e., associations would be weakened), it can also lead to overestimation of results. If adjustment variables are measured poorly, effect inflation is common. Aggregate-level data are often imprecise; for example, neighborhood data or measurements to model environmental data may have poor spatial resolutions. Hence, some scientists advocate careful analyses of measurement error, such that the possible degree of error is reviewed, modeled, and tested. Sensitivity analyses can be used to show the degree of measurement error that would make results disappear (Bennett et al. 2017). Good practice in the analyses of combined data with some reasonable doubt about measurement error should incorporate such

analyses to quantify robustness to measurement error. The practice is, however, uncommon.

Reproducibility: In this discussion, we will follow the report by the National Academy of Sciences (NASEM 2019) and distinguish reproducibility from replicability. Reproducibility is defined as obtaining the same results with the same protocol (measurements and model) in the same population. Open science advocates have called for codes or syntax and, ideally, the data to be made available publicly or at least on request. Likewise, analytical protocols and preregistration of analyses are suggested. These protocols should be specific and uploaded to registries and are ideally presented and discussed prior to any analyses. This increasingly common practice is important and useful even if no specific hypothesis is tested. That said, a formal evaluation of the progress in reproducibility achieved by the open science initiatives is lacking. Some guidelines have been suggested (e.g., transparency and openness guidelines; Nosek et al. 2015), but it is important that guidelines do not stifle innovation.

Replicability is the capacity to obtain consistent results across studies aimed at answering the same scientific question (NASEM 2019), "each of which has obtained its own data." The so-called replication crisis (Schooler 2014) has been discussed for over a decade. Several scientists have attempted to estimate the lack of replication in observational research and some state that many research findings are "false" (Ioannidis 2005). Although such claims cannot be quantified easily, combining data not initially collected for a certain research question or using large-scale social media data raise similar concerns. Replicability of results is important to guide policy and other implementation efforts. As Lash (2022) pointed out, however, replicability should not be judged by whether two results are both significant (or not). Rather, it is the slow accumulation of knowledge that mostly guides policy; and replication endeavors are an important part of this accumulation process. Limited sample size, chance findings, reliance on statistical testing, different forms of bias, selective reporting, and publication bias severely impact the ability of researchers to replicate results. Good practice in analyzing integrated data is not different from any other form of science. Some advocate analytical protocols and preregistration of analyses, but there are reasons to assume that this might improve reproducibility but not replicability (Hicks 2021). Others advocate for the use of careful multiple testing controls to reduce chance findings and data dredging. This, however, addresses only one problem of replicability and can increase the type II error (i.e., false negatives): the most significant associations are not necessarily the reproducible ones. Replication efforts using other samples to examine an association or other findings can be part of the original investigation. The current practice in some fields, like machine learning, is to reproduce a statistical model obtained in one sample by applying it to another independent sample, typically split off from the same dataset prior to analysis, and formally examine if the same result is obtained. In this framework, an algorithm is fitted on the training data and the model performance is tested on

such independent, unseen, test data. Well-powered studies could be redefined as allowing such replication. Yet, this is not a typical practice in population studies with aggregate and individual-level data, often because sample size does not permit such splitting of data and effects are commonly small.

A result is replicable if the design and findings of the original study and replication attempts are *qualitatively similar*. Because similarity of study includes the design, measurements, sampling frame, and analyses, and these assumptions are often implicit and involve judgment, replicability is almost always ambiguous if not put into context and can be highly controversial (Feest 2016). Another important facet of replicability endeavors is that they can unravel why associations differ, how variability in measurement or exposure distribution impact results. Large-scale social media or geocoded data may offer scientists the possibility to study the seeming lack of consistency and poor replicability of results, which point to cultural specificity, or the impact of measurement or design.

Generalizability of results is a major determinant of the usefulness of data. If findings cannot be generalized or extrapolated to specific, even if limited, populations or population subgroups, there is a limit to generalizable knowledge that can be obtained. Insights may still arise from studies where generalizable knowledge is not the goal (e.g., case studies), but care is needed not to over interpret the insights, especially when implementing interventions or policies. Generalizability is inherently subjective. It is conditional on a careful description of the research, not only the study population (characteristics, ascertainment, inclusion/exclusion criteria), but the exact study question, methodology adopted, and also outcome and exposure definition and assessment. Importantly, generalizability (external validity) is conditional on the (internal) validity of results. A biased finding may be reproducible even in different settings but generalizing it to the larger or any other population makes no sense. Hence a careful evaluation of possible measurement error, selection bias, and confounding is key.

Representativeness on key characteristics such as race/ethnicity or urbanicity is often used as an indicator of the generalizability of results to different populations. Such representativeness can also indicate lack of selection bias (internal validity) and that results apply widely to the general population. The degree to which a population is representative of a larger population is an indicator of sample generalizability. Without a clear sampling frame, however, representativeness may create the illusion of generalizability, for example, if the minorities included differ from minorities not sampled. Despite the appeal of representativeness, we encourage researchers to consider sampling nonrepresentative populations for certain questions. This may make sense for many reasons beyond practical ones. Opting for nonrepresentative populations can minimize bias (certain groups may be more reliable reporters), it can increase variability of the exposure, and it can help include or focus on subgroups (e.g.,

Indigenous or LGBTQ+ populations), which are often poorly represented in large population-based samples (Richiardi et al. 2013).

To increase generalizability, we recommend out-of-study reproducibility efforts as outlined above. Such efforts can truly help judge the extent of generalizability and further the process of evidence accumulation. Although replication efforts can best begin with populations and designs that are as similar as possible, often sample characteristics, settings, or measurements will differ to some degree. While some researchers recommend that analytical strategies and modeling practices should be kept the same in replication efforts, we argue research design can and, if possible, should be improved according to current insights. No single reproducibility study will show or refute generalizability, but out-of-study, rather than just an out-of-sample reproducibility, is needed to evaluate whether results hold in a different context and thus are generalizable. Even hypothesis-generating analyses of integrated data should aim to implement out-of-study reproducibility. In sum, replicability and generalizability are not tested, but depend on the quality of research that is carefully evaluated in a complex and often slow process.

Barriers to Multiscale Integration of Individual and Population Data

While most of this chapter focuses on the many benefits of multiscale integration, we must also consider some of the barriers. Table 5.1 summarizes some of the main incentives that influence multiscale/multilevel data integration. While the benefits that appear in the left column were emphasized in the introduction and assumed as given throughout the chapter, here we discuss potential disincentives listed in the right column.

While data reuse may come with savings in time and resources needed to collect new data, the process of understanding the study design and data selection processes that led to the reused datasets and obtaining these datasets may also require substantial investments of time. The failure to understand sufficiently the domain expertise that led to some of the data runs the risk of producing invalid results (see Lovasi et al., this volume). This implies further risk of the research becoming an example of "bad science" (Ritchie 2020) with potential criticism from domain experts and affected communities who may assert that the researchers did not sufficiently recognize the domain context and the need for relevant expertise, which can present a reputational risk to individuals, the group of collaborators, and any institutions with which they affiliate. Recognizing this possibility at the beginning of a project may lead to the need for an expanded research team with additional expertise, which implies the need for up-front resource investments to create new or negotiate expansion of research partnerships (e.g., to enable intentional and purposeful stakeholder involvement using value-sensitive designs; Friedman and Hendry

Table 5.1 Incentives that promote or hinder the integration of data.

Promote	Hinder
• Better insights, knowledge • Technological resources, tools, ontologies, statistical methods, GitHub, code sharing, open data, repositories, data lakes • Opportunities to innovate • Partnership, collaboration • Fun • Ecological validity, connectivity, generalizability, relevance • Quality, deeper, refined understanding • Collaboration • Refined understanding • Funding, reward, efficiency, value in influencing activities • Internal collaboration, visibility of research for expansion • Interest in interdisciplinary scholarship, publication, broader recognition • Domain expertise expansion	• Technological barriers • Legal agreements, divergence/complexity • Lack of expertise • Time pressure, time taken • Bureaucracy • Discrimination: communication, publication • Promotion of interdisciplinary work (independent vs. collaborative work) • Data quality • Collaboration • Risk of invalid results, domain context, expertise • Possession of data, fear of discovery • Stakeholders expanded • Fear of trying • Head in the sand, research suppression • Lack of recognition • Restrictions due to nature of data • Funding, variability across domains • People, training • Peer group • Demand for technical support, inertia associated with sharing data • Lack of training, opportunity costs • Misperception of costs, risks, benefits

2019; McIntyre 2008; Viswanathan et al. 2004). In addition, the nature of the datasets to be integrated may expand the size and number of stakeholders, specifically affected communities interested in engaging in the research process in some capacity, and may come with some restrictions that include ethical, legal, and institutional review. For example, if the research involves using data that are subject to a data use agreement, then the process of reusing the data may require negotiation with those involved in the specific data use agreement, and navigation of complicated and potentially divergent interests. In addition, if the data use agreement precludes sharing the data outside of the collaboration, then this may restrict options for publication of the results to journals that do not require deposition of the data into a repository.

Ethical and legal requirements can preclude the sharing of data at the same level of granularity (e.g., across country borders), leading to both bureaucratic and methodological challenges when collaborating researchers have access to different levels of detail. Interdisciplinary collaboration may introduce another complicating factor when distinct disciplines adopt noncompatible data sharing and reuse policies.

The identification of datasets for potential integration (step 1) does not mean that the researcher will gain access to the datasets in a usable format (or at all). Specifically, not all researchers (or, for that matter, institutions) share data. This may reflect their compliance with agreements they made to collect or assemble the data, interests in protecting data that they are actively analyzing or expect to analyze once the data collection ends (e.g., for a longitudinal study), or simply because not sharing data maintains control of further discovery, evaluation, and communications of the data and prevents misuse or uses that might harm the reputations of those who possess the data (e.g., discovery of errors in the data). Similarly, restricting access to data to prevent discoveries by others may reflect the preferences of some data owners to maintain the uncertainty and ambiguity that comes from lack of analysis, because providing data to independent researchers might lead to real or perceived risks of negative attention. For example, analyses of integrated datasets may result in identification of previously unidentified issues that some stakeholders may prefer not to become aware of (an attitude summarized as "head in the sand" in Table 5.1), lead to claims that require further resource investments, or create new risks for the data owners or stakeholders. In this regard, research that integrates data that may directly affect the activities of one stakeholder may encounter active research suppression efforts by others. Resistance for sharing data may also be due to fear of data misuse—the so-called dual use issues (Bezuidenhout 2013).

In spite of the development of tools, platforms, and advances in technology, research efforts that integrate individual and population data may encounter technological hurdles related to the nature of the datasets, issues with data quality, challenges with incompatibility between platforms and software used for processing data, inappropriate and/or insufficient ontologies required for coherent understanding of the concepts behind the data, insufficient data documentation, and computational demands that necessitate the engagement of technological or computational expertise in addition to any subject matter expertise. For example, while open data repositories mean that researchers may access datasets simply by downloading them, lack of documentation on these data, such as poor metadata, may render them unusable.

Along these lines, researchers who are willing to share data can upload the data with different levels of processing (e.g., raw, curated, derived) and their responsibility for data sharing ends with depositing the data into an adequate repository. Nevertheless, good data management practices, together with FAIR properties, require that datasets be documented by use of appropriate metadata records. Adequate documentation is itself a time-consuming activity that goes largely unrewarded, yet another barrier to good practices in data sharing.

There is, moreover, an expectation from researchers who want to reuse the data that the depositor of the data is responsible for answering questions, producing details about the data, essentially providing free technical support to potential data users. Since this kind of stewardship is seldom available, this

means that those seeking to use data from a repository may need to at least attempt to establish a collaboration with the data collector or generator and/or engage others to ensure appropriate interpretation of the data during integration. Alternatively, if data repositories come with expectations of perpetual stewardship of the data and responsibility for spending time helping any and all potential users, then this may create a disincentive for depositing data for reuse, or deposit only "self-explanatory" data, when research projects would potentially benefit from a more complete dataset.

At an individual level, engaging in research that integrates individual and population (e.g., area-level) data as part of a collaboration will likely mean sharing credit for the work. This may have substantial career implications for new and less-established researchers whose scholarship and promotion are judged by their independent contributions, and who may not receive sufficient recognition for their contributions as part of the team. In addition, the dissemination of the results may come with challenges related to communication of added complexities associated with the multiscale data integration, and difficulties finding an appropriate journal and/or opportunities to publish in high impact journals that may view the work as not a good disciplinary (or domain) fit. Similarly, the people who developed the original idea and intellectual property are rarely acknowledged, even though they managed to obtain funding, and performed data collection, cleaning, and storage to a level that would allow other researchers to use and access the data later are rarely acknowledged. This lack of acknowledgment may hamper data collection and sharing more broadly. The group dynamics of collaborative activities can provide a substantial disincentive and discourage even attempts to engage due to real or perceived pressures that researchers face to meet productivity targets ("publish or perish"), secure funding for research outside of established domain-specific funding streams or in domains with variable or little funding opportunities, and opportunity costs associated with investing in additional training and acquisition of staff with less-familiar skills and expertise.

The results that may come with the innovation of research that integrates multiscale data may also face challenges due to the absence of peer groups, or to experts in related domains who may perceive the research as a threat. All research projects come with some risk of failure (e.g., not resulting in outcomes worthy of publication or further pursuit), but some unique pathways of failure come from combining individual and population data. For example, the effort may fail after substantial investments in the up-front activities that lead to the integration process if the collaborators determine that the quality and fitness of the data when integrated do not support the analysis required to answer the research question. Those who perceive this and other risks as potentially very substantial may fear even trying to engage in this type of research. As with any research activity, individual researchers may misperceive the risks, costs, and benefits of participating in activities that integrate multiscale data, particularly

in the context of evaluating the opportunity costs. With time, as more efforts either succeed or fail, the risks may become more easily understood.

Ethical and Legal Considerations

Ethics and law are essential tools when making decisions about data use, but they are different constructs that provide different types of answers (Hulkower et al. 2020). If the question is, "Can I use these data?" ethics will help distinguish whether the answer is "right" or "wrong" or "should" or "should not." In contrast, the law helps distinguish between "yes," "no," or "maybe" and answers of "must" versus "may." It is also essential to recognize that ethical activities might not be legal (Hulkower et al. 2020) and that legal activities might not be ethical. Legal and ethical issues on data integration and use must be important considerations in determining whether a research project can or should proceed. Here, we focus on two critical considerations of integrating knowledge from individual and aggregate-level data: *group harms* and *legal uncertainty*.

Group Harms

In the ethical review process, the overwhelming focus is on the mitigation of individual-level risks. These risks are well documented, and research ethics committees are accustomed to weighing these risks against the perceived value of a proposed research project. To this end, the principal strategies include seeking an individual's consent, where practicable, and de-identification (for definitions on distinct forms of data privacy, see Table 1 in Kushida et al. 2012). Informed consent rests on the idea that the individual is best situated to evaluate the risks and benefits of participating in a research project. De-identification rests on the assumption that rendering individuals more difficult to identify will reduce the risks faced by those individuals ("data subjects"). Both strategies can, however, be legitimately criticized in big data contexts. When integrating large datasets involving individual- and aggregate-level data, the objective is often to gain insights about groups of people with similar characteristics (e.g., their geospatial location at a particular level of spatial granularity). These insights—well-meaning or not—can lead to substantial harm to these groups and the individuals within them. Thus, research that uses big data, especially when it involves integration, implies a different type of risk that is largely ignored by research ethics committees: group harms (Ienca et al. 2018).

Group harms are those harms that adversely affect the collective interests of individuals sharing common characteristics (Xafis et al. 2019). Some of these groups might have legal protections (Wachter 2022); for example, in the United States, various antidiscrimination laws protect racial groups legally. Other groups might have substantial predictive importance but lack any protections

under the law. For example, owning a dog is an important grouping characteristic used by many data brokers, yet "dog owners" is not a legally protected class under U.S. antidiscrimination laws (Federal Trade Commission 2014). Still other groupings, such as those derived through artificial intelligence, are entirely incomprehensible to humans (Wachter 2022). These incomprehensible groupings might include, for example, individuals with specific mouse movement patterns, or specific web-browsing behaviors (Wachter 2022).

Using de-identification or aggregation may protect the individual data subjects, but it shifts the focus of analysis, and the risks that come with it, to an identified and identifiable group. As a consequence, the grouping might aggravate risks for group members. For example, data aggregated using racial grouping criteria could facilitate erroneous stereotypes of that group and discrimination. Behavioral insights about a group like "dog owners," mentioned above, could enable harmful and potentially legal discrimination against individuals within the group. Also, de-identification may be meaningless as a privacy protection mechanism to individuals whose identity is strongly linked to the group they belong to, as is the case of many Indigenous groups, for which specific data governance principles exist (Carroll et al. 2020).

Similarly, consent is an imperfect tool to manage group harms. An individual who provides consent to research could face minimal individual risks, but the group the individual belongs to could face substantial group harms. For example, genetic data can be collected with minimal risk to an individual, but the use of the genetic data can have far-reaching impacts on the individual's family, community, and even culture, as was made painfully clear when genetic information from the Havasupai Native American tribe was used for research that caused significant cultural harm, stigma, and embarrassment. Genetic data collection is also a good example of another kind of group harm: by being "aggregated" into a group, the individual may not only incur harms–other members of that group, and sometimes even outside the group, may be harmed as well (e.g., allowing discovery of new knowledge through use of bioinformatics).

Moreover, most individuals cannot fully know or appreciate the implications of their "consent." For example, most Meta (Facebook) users might not appreciate that the broad consent they provided to Meta permitted widespread emotional experimentation on vulnerable social media users (Reilly 2017). An individual's ability to protect against group harms through withholding consent depends substantially on the individual's awareness of the group(s) they belong to.

Importantly, aggregation and grouping decisions during integration steps 1–4, described earlier in this chapter, can affect the distribution of group harms. Individuals and the communities they belong to have a right to be counted (Fairchild 2015). This right derives from the fact that information can empower individuals and communities to act. For example, the discovery that an industry is harming a community empowers the individuals within that community

to act to seek new regulations for the industry; that action would not, however, occur but for the knowledge of the harm. Similarly, the act of counting informs crucial resource allocation decisions. Consequently, inequitable counting begets inequitable resource distributions. In the extreme, inequitable counting can lead to so-called data genocide, whereby the undercounting of a particular group contributes to systemic exclusion of a group (and eventual extermination) (Urban Indian Health Institute 2021). For example, a 2021 report by the Urban Indian Health Institute alleged that inadequate reporting and sharing of COVID-19 surveillance data with tribal communities and governments contributed to ongoing data genocide of American Indian and Alaskan Native populations. For these and other reasons, great care should be taken to ensure that aggregation and grouping decisions do not contribute to systematic and inequitable disenfranchisement of vulnerable groups.

Critically, the group harms can extend beyond the specific subject matter of the data being aggregated or integrated. For example, consider a research project on school performance, where researchers report only aggregated student performance data at the school level to protect individual students. Although the reported data concern only specific schools, there might be group harms that extend beyond the study's focus. Neighborhoods surrounding poorly performing schools might see falling property values and increasing community stigma. Since the neighborhood residents were not the focus of the study, they might not have had an appropriate opportunity to raise their concerns with the researchers. In this way, researchers and research ethics committees should consider what groups, internal and external to the research focus, could face group harm from the research activity and weigh the risks and benefits to both individuals and groups accordingly.

Seeking a "social license" from relevant communities or groups is one approach to address potential group harms. Social license refers to the informal permission given by a community to a public or private entity to engage in a specific activity (Shaw et al. 2020; see also Weigle et al. this volume). In the context of digital ethology and other big data activities, a social license provides legitimacy to collect, use, or share data that is tied to the data subjects' communities. Additionally, the social license helps establish credibility and builds trust between the parties (Jijelava and Vanclay 2017). Careful and appropriate community consultation and engagement (Dickert and Sugarman 2005) can help develop a social license (Corscadden et al. 2012). For example, in the context of public health surveillance, the World Health Organization (WHO 2017) cites community consultation and involvement as one way to support ethical surveillance activities.

Legal Uncertainty

There are multiple dimensions of *legal uncertainty* in digital ethology and big data generally. First, the technology to easily share digital data across great

distances has existed for decades, but laws often make collecting, accessing, sharing, and using data exceptionally difficult in practice (Schmit et al. 2019). Laws vary across jurisdictional lines, and organizations interpret and operationalize laws into their internal policies in a variety of ways. Moreover, laws can regulate different types of data (e.g., health, census) or certain data activities (e.g., research, public health) differently (Schmit et al. 2022). These differences in laws must be carefully navigated when data that are regulated by different laws are integrated. This complexity creates both real and perceived legal barriers to data use. Consequently, the first and most challenging aspect of legal uncertainty in a data project is often understanding what legal rules apply (Public Health Informatics Institute 2021).

In addition to the legal complexity, technological innovation in big data analytics far outpaces the ability of regulators to manage new and emerging social risks. Bowman describes this problem using the parable of the race between the tortoise and the hare (Bowman 2013). In this analogy, technology is the hare—progressing at a rapid pace—and law is the tortoise—progressing at a much slower pace. When the gap between the two is too great, technological progress is impeded (i.e., the hare sleeps). This can happen when an out-of-date law is used to regulate a technological practice it was never intended to regulate, or when the uncertainty and legal risk of operating under out-of-date laws is too great. For example, the relative failure of the United States to keep pace with other countries' regulation of data protection led to the invalidation of the international data sharing agreement, the EU–U.S. Privacy Shield Framework by the Court of Justice of the European Union (Kerry 2021). This decision led to the cessation of many data sharing activities between European and U.S. researchers, and even questions concerning data transfer across the Atlantic (Hallinan et al. 2021). In this way, the failure of regulators to keep pace can interrupt scientific progress.

Rapid innovation also challenges regulators by making it difficult to define the subject of proposed regulation. Laws work by attaching legal prohibitions or permissions to words. Consequently, legal definitions of these words are incredibly important. Innovation-laden terms like "big data" or "artificial intelligence" have been difficult to define, and thus difficult to regulate. Rapid innovations in how the technologies are used make it difficult to balance precautionary risk-mitigation with appropriate room for technological progress.

Legal definitions can also lead to tremendous confusion because they can be counterfactual. A law might provide a definition for a de-identified dataset, but an individual can in fact be identified within a dataset by using an appropriate reidentification method. Here, the purpose of the legal definition is not to describe what is true, but rather to describe the thing that is subject to the law. Nevertheless, the discrepancy between legal definitions and technical terminology can lead to considerable confusion between parties—such as researchers, data custodians, research ethics committees, and data subjects. In these situations, it is important to clarify the intent of the terminology being

used when describing a data activity. For example, if data must be legally de-identified to comply with the law, then the legal definition is important. If, however, data identifiability is part of the data governance approach to manage an ethical concern like risk of harm, then the legal definition is less relevant and might either overmanage or undermanage the ethical issue.

Complexities and uncertainties in law and ethics can lead to both real and perceived barriers to data use. Well-intentioned individuals can reach reasonable (and seemingly intractable) disagreements on whether a data use is legal or ethical. Successfully navigating legal and ethical issues in digital ethology requires identifying these real and perceived barriers to data use. This will in turn require subsequent negotiation among all actors involved (legal, administrative, researchers) to achieve an agreement at some level ("getting to yes").

In these challenges, lawyers have a duty to advise their clients of the legal and ethical risks of a proposed activity. Ultimately, however, clients have the decision about whether to proceed with an activity in the face of the legal and ethical risks. For research institutions, there is unlikely to be a risk-free course of action in the face of these and other legal and ethical challenges. Unfortunately, often data sharing agreement negotiations can be bogged down by organizations (or their attorneys) aggressively pursuing a zero-risk agreement, resulting in protracted delays or restrictions that are neither legally nor ethically required. Some tolerance of risks—known and unknown—is necessary to ensure that socially beneficial research continues and the key to progress may require different ways for balanced risk management (e.g., Table 1 in Kum et al. 2014) rather than risk avoidance.

Conclusions and Additional Directions

This chapter analyzed some of the factors involved in generating knowledge from multiscale data integration, ranging from the individual to the population level. As seen throughout the text, in digital ethology such integration requires interdisciplinary cooperation. Indeed, one must never forget that data integration is not "just" integration of data, but also of the knowledge of domain experts.

The role of human expertise must be emphasized all through the integration process, since there will always be limits to what technology can provide. Humans intervene in selecting and curating the data, choosing the integration strategies, analyzing and interpreting results, documenting data and metadata, and checking quality at all integration stages. Quality assessment and monitoring throughout integration planning and execution are essential. Indeed, quality questions must be embedded into integration efforts. This might even be called a "quality by design" approach, in the sense that quality must be planned for, and designed into the integration of knowledge. The need for appropriate documentation, including metadata, is a requirement for checking quality and

also supporting FAIR principles. Multiscale data integration also requires navigation of ethical and legal paths and pitfalls to access, integrate, and analyze the integrated results. The associated risks must be acknowledged and considered by all actors involved in knowledge generation and governance, so that the barriers these risks pose can be overcome through cooperation.

While research collaborations are traditionally implemented through *direct* interactions among groups of researchers, the worldwide movement toward open science has introduced a new kind of interdisciplinarity in which groups collaborate through making the digital resources produced by their research (data, software, code) publicly available for reuse. This second type of collaboration, an *indirect* one, has been enabled thanks to progress in digital technologies. Here, the digital resources that are made available through, for example, repositories, data lakes, or federations, become de facto "collaboration mediators." Researchers who painstakingly prepare data to become available for sharing are assisting groups they may never meet; they are helping to solve yet-to-be-formulated research questions and, as such, are, indeed, collaborating with the future.

In this sense, open access to data and code are to be encouraged, and acknowledged, as a means of fostering scientific progress and new kinds of knowledge creation. Encouragement and acknowledgment also apply to institutions that provide resources to support appropriate data management and archival, thereby helping researchers to extend their cooperation networks. Digital ethologists whose research involves integration of multiscale data typically rely on datasets made available by others. Providing broad access and transparency can moreover foster reproducible research as well as scientific innovation in the methodologies developed, in the algorithms, in the code, and in the results themselves.

While the emphasis was on population-level data as a powerful kind of data aggregation that can help advance research in this field, other kinds of aggregation may also be considered, to which many of the issues raised in this chapter apply. This is the case, for instance, of satellite images, in which each pixel is a spatiotemporal aggregate of remotely sensed data that indicates human activity (or lack thereof). Spatialized pixels can be integrated with data on individuals and communities that inhabit that region or vicinity through use of coordinates and geo-statistics. Satellite images are aggregators of human activity, as in land-use maps, or as reflecting change in patterns of human behavior due to changes in the built or natural environment. For instance, forest fires or riverine pollution or erosion reflected in such images can be correlated with displacement of Indigenous populations, individual reports of respiratory diseases, or patterns in the spread of zoonotic diseases (Mishra et al. 2021). These are examples of aggregations that are not specifically computed as such; rather, they emerge from direct observation via the instruments used to collect such data.

Acknowledgments

The authors gratefully acknowledge the support of the Ernst Strüngmann Forum and the organizers of the Forum on Digital Ethology. Many thanks to all Forum colleagues for suggestions, comments, and stimulating discussions throughout and beyond the event.

Digital Ethology

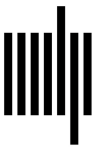
Human Behavior in Geospatial Context

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