Quilting a Time-Place Mosaic: Concluding Remarks

Social science is at a pivotal moment. The advent of "big data" from the Internet, social media, and smartphones as well as newly available administrative data from electronic sources has opened the door to new understandings of people and society. That said, realizing this promise requires a vision for the future and a practical plan for reaching it. The articles in this volume begin this work. Each addresses some aspect of data linkage. Each can be considered a patch in a time-place mosaic. This concluding article considers the articles as a collection and how they might be quilted together. It discusses the diversity of sources available, differences in time depth and sociospatial coverage, and the many challenges of using data not designed for research. It identifies the benefits that would cumulate as a set of regional data centers to assemble, link, curate, and share diverse data sources is established and coordinated. These data centers could provide the combination of national coverage, local depth, and temporal precision needed to advance our understanding of the American population.

Keywords: big data; social science; data linkage; data access; confidentiality; survey research; administrative data; social media

By
BARBARA ENTWISLE,
SANDRA L. HOFFERTH,
and
EMILIO F. MORAN

The social and behavioral sciences are at a remarkable juncture. Diverse data from transactional sources, the Internet, social media, and other sources not designed for

Barbara Entwisle is Kenan Distinguished Professor of Sociology and a fellow of the Carolina Population Center at the University of North Carolina at Chapel Hill. Her research focuses broadly on the study of social, natural, and built environments and their consequences for a range of demographic and health outcomes.

Sandra Hofferth is professor emerita, School of Public Health, and a research professor, Maryland Population Research Center, University of Maryland. Her research focuses on American's use of time; economic disadvantage, parental behavior, and child health and development; fathers and fathering; and immigrant youth's transition to adulthood.

Correspondence: entwisle@unc.edu

DOI: 10.1177/0002716216683698

research can now be used in research to varying degrees, making it possible to ask and answer questions about the dynamics of social interaction and behavior in new ways. This is particularly true when these data are combined with social surveys and other traditional social science data sources. Currently, however, the promise of these new opportunities is unfulfilled. One reason is that individual researchers and teams solve challenges associated with the use of new data alone, independently, for their own purposes. There is a potential for duplication of effort. Further, the linkages these researchers create, and the datasets they build, are not integrated into a larger, comprehensive resource. The vision of a social observatory or set of regional data centers laid out in the first article and that led to this volume is to build on and coordinate these efforts, strategically leveraging them to benefit social, behavioral, and economic science, as well to create a public good for the nation.

Our vision is to create a national resource that would assemble, document, curate, and disseminate the growing collection of data potentially relevant to social, behavioral, and economic research—a social observatory. Such a resource would make it possible to ask new questions and to answer others in new ways. It would promote data sharing, reduce duplication, and increase the cost-effectiveness of research. It would also enhance research quality and the reproducibility of results. There are many ways in which such a resource might be organized. Although it is possible to think in terms of a single national social observatory, it is more likely that there would be multiple entities, regional or distinguished according to foci that would cluster expertise (e.g., social programs, social mobility, migration, health, environment-related hazards). What is critical, however, is that these centers or observatories would play a key role in linking data and making them available to the rest of the science community and society. The challenges of linking data are significant and cannot be downplayed; it is for this reason that this volume—and the network of data centers it proposes—sees as a priority advancing the state of the science on this subject

The articles assembled for this special issue all address some aspect of data linkage. They cover a broad range of topics, including monitoring the health and well-being of the population (Dai et al.; Bader et al.); assessing the impact of social programs (O'Hara et al.; Digitale et al.; Leonard et al.); measuring concepts such as family and neighborhood in new ways (Bader et al.; O'Hara et al.; Browning et al.); planning for and responding to environmental hazards (Fussell et al.); and developing new approaches to movement, mobility, and migration (Browning et al.; Kislev; Fussell et al.). Linking diverse data sources is central to all of them.

In these concluding remarks, we consider the articles as a collection. We comment on the breadth of data sources used here, ways that data can be linked, and

---

Emilio F. Moran is Hannah Distinguished Professor at Michigan State University. He is the author of eleven books, fifteen edited volumes, and more than 200 journal articles and book chapters. His research addresses how humans interact with the environment. He was elected to the U.S. National Academy of Sciences in 2010.

NOTE: Funding for the Social Observatories Coordinating Network was provided by National Science Foundation grant SES1237498.

potential benefits to research quality that might flow from linking data. We also discuss challenges associated with linking diverse data in ways that are accurate and cumulative and with making the linked data accessible to a broad range of researchers. Collectively, the articles demonstrate the value of a social observatory that assembles sources of data on multiple topics, multiple types of information on a single topic, and even multiple measures of the same concept within a topic (e.g., measures of the family).

Sources, Links, and Coverage

The articles in this volume illustrate the breadth and diversity of data relevant to social, behavioral, and economic science, and in so doing reflect much of the opportunity as well as many of the challenges associated with their use. Included in these kinds of data are census, surveys, and traditional sources of administrative data. Also included are less traditional sources of administrative data such as from the Internal Revenue Service (IRS), the Department of Housing and Urban Development, and other federal agencies; state-level program data from the Temporary Assistance for Needy Families (TANF) program, Supplemental Nutrition Assistance Program (SNAP), and the Women, Infants, and Children (WIC) nutrition program; and data from nonprofits such as the Crossroads Program in Dallas, Texas. Data from "new" sources such as electronic medical records, Internet data (e.g., Google StreetView), social media (e.g., Twitter), and locational data from GPS-enabled mobile devices such as smartphones are also considered.

Linking these data is accomplished in a variety of ways. Sometimes, linking is at the person level, as discussed by O'Hara and her colleagues in their article. This requires careful attention to the privacy of the individuals involved and the confidentiality of their responses and, ultimately, may necessitate strict limits on access to the linked data. The datasets described by O'Hara et al. reside in highly secure data enclaves managed by the U.S. Census Bureau. Potential users of the data must go through a rigorous application and review process, and if approved, use the data onsite with strict oversight of any tables or graphs created. Privacy and confidentiality concerns can also arise when linking at the household level. Rather than limiting access, Leonard and her colleagues anonymize their data as a way to address these concerns while making the data more broadly available. Digitale and colleagues link individuals with the family-planning facilities they reported using; to maintain confidentiality, only broad characteristics of the facilities are analyzed. Bader and his team point out that identifying individual respondents' addresses in proprietary software may violate confidentiality and limit their use in conducting neighborhood audits.

Typically, linking is accomplished at a higher geographic level such as a block group (Browning et al.), census tract (Bader et al.), county (Fussell et al.), state (Dai et al.), or country (Kislev). It is worth noting that in many instances, less aggregated data are available. For example, Browning and his colleagues code

latitude-longitude coordinates collected by GPS-enabled smartphones to block groups rather than analyze them directly. In the most finely disaggregated analysis, Digitale and colleagues link participant residence with that of family-planning facilities within 10 kilometers using GPS coordinates. Dai and colleagues link Twitter data to state-level prevalence of asthma and other state characteristics from national surveys. It is theoretically possible for them to link data at lower levels of aggregation; however, at lower levels the precision of estimates of asthma prevalence is likely to be poor. Data at higher levels of aggregation are less likely to involve privacy and confidentiality concerns, although these cannot be entirely ruled out.

Linking data also involves temporal considerations. To be useful, linked data sources need to be temporally aligned, with putative causes measured at or before effects. Some of the new sources make data available instantaneously and continuously, in contrast to the annual, biannual, quinquennial, or decadal time steps of traditional sources. In the collection of articles that we have assembled, the finest temporal unit of analysis is the day, as presented by Dai and colleagues in their analysis of Twitter data linked to the Behavioral Risk Factor Surveillance System (BRFSS) and the American Community Survey (ACS). Temporally speaking, although fine resolution is available in the data sources, Dai and colleagues yield on geographic scale, grouping data into states or grouping states by quartiles or population characteristics. The same tendency toward geographic aggregation is found among others working at fine temporal scale. Indeed, there tends to be an inverse relationship between temporal and geographic detail. This relationship may reflect the capacity of individual researchers and the size of datasets that are easily managed. With proper support, regional data centers would not be limited in this way. These tradeoffs reflect the limitations of many individual research projects in the social and behavioral sciences, which are limited by small awards, small teams, and the need to reduce time and spatial coverage to achieve sufficient depth. Observatories, as long-term data centers, would go a long way toward helping to address these limitations.

Spatial and temporal coverage varies in this volume. Some of the authors assemble data that are national in scope. For example, Fussell and her colleagues integrate 1970 to 2010 data from the Spatial Hazard Events and Losses Database for the United States (SHELDUS) and U.S. census data using ArcGIS geo-referenced county-year FIPS codes and county-boundary files. Dai et al.'s and Kislev's studies are also national in scope but not as expansive in temporal coverage. Other authors focus on the local level. For example, Bader and his colleagues examine aging in place in four cities: New York, San Jose, Philadelphia, and Detroit. Browning and his colleagues examine adolescent exposure to violent locations in Columbus, Ohio. Leonard, Hughes, and Pruitt investigate responses of vulnerable households to unanticipated health events such as heart attacks, which they label "health shocks," in Dallas, Texas. It would be desirable to scale up these studies, or at least the data on which they are based, to a national level so that social and geographic inequalities and the typicality of these cases could be better understood. It would also be desirable to extend them temporally.

Finally, although we talk about social observatories in terms of research in the United States, it may make sense not to restrict the focus in this way. We live in a globalized world in which the flow of people does not stop at national boundaries, and our studies must follow people to understand their behavior. An example of this here is Kislev's article, which draws on survey data from Europe as well as census and survey data from the United States to examine immigrant outcomes in the United States. Although Dai et al. focus on the United States, they have access to Twitter data from countries around the world. Digitale and her colleagues link survey and administrative data in Malawi to look at the impact of family-planning facilities on contraceptive use among young women residing there. These articles remind us that social science is global; important questions are not confined to national borders. Indeed, some of the newer sources of data are not contained within nation-states. This is particularly true for social media data, where geographic boundaries are in a very real sense arbitrary.

Improved Measurement and Inference

Collectively, the articles in this volume demonstrate the value of a network of observatories or data centers for addressing topics critical to the advancement of social and behavioral science. In the process, they comment on measurement issues and potential problems of statistical inference that are worthy of consideration. We review the value of linked data for addressing these more methodological topics here.

One value of linked sources is that each can provide a check on the others. This is especially useful in leveraging new sources of data. The new "big data" sources have important advantages such as temporal and spatial detail more consistent with a dynamic and place-based understanding of social patterns and behavior than typically possible with traditional sources. However, because they were designed for purposes other than research, there are important questions to be asked about their provenance, availability, and quality, including coverage, representativeness, completeness, accuracy, and relevance of the information provided. For example, a moment-by-moment picture of some aspects of behavior can be obtained from the Internet, Twitter, and mobile devices, in ways that are not possible with traditional data sources. Whether this picture is representative is an open question. Not all Twitter data can be linked to a geographic unit, even a country—a problem noted by Dai and colleagues. Furthermore, data may not be population-representative; Twitter is used more by men than women, more by younger than older adults. Although traditional sources such as social surveys collect data only annually or at greater intervals, they are carefully designed and their coverage is well-documented. Hao, Lee, and Dai capitalize on contrasting strengths when they use BRFSS data to evaluate the potential use of Twitter data for monitoring asthma at the state level. They conclude that tweets can serve "as a rapid, cost-effective health detection system with real-time information to monitor chronic disease and track public sentiment," although

correlations of asthma-related tweets with BRFSS reports of asthma prevalence at the state level are modest (r = .32 to .38).

Another strength of nontraditional sources is potentially better coverage of hard-to-reach and frequently underreported groups. For example, Leonard and her colleagues argue that, in Dallas, vulnerable populations are better captured through visits to the Crossroads Food Pantry and electronic medical records from a local hospital than through traditional data sources. Further, while each of these sources may have limitations, together they may provide a more comprehensive picture than any of them could alone. Along similar lines, O'Hara and her colleagues report on studies examining administrative records on participants in state-supported social programs such as WIC, TANF, and SNAP and compare them to the self-reported participation from the ACS and the Survey of Income and Program Participation. These studies find underreporting of program participation ranging from 10 percent to 35 percent. Yet, turning it around, they report that supplementing census records with WIC and SNAP records substantially improves the coverage of young children, for example. The complementary perspectives of multiple sources may provide a comprehensive picture of the phenomenon in question. Centralizing the assembly of these data makes it possible to play to the strengths of one against the strengths of the other—making the challenge of linking data all the more necessary and valuable.

Information available in one dataset may be used to supplement another, as a way to address unit and item nonresponse. Plans for the 2020 U.S. Census call for the use of administrative data as a supplement to the main data collection (O'Hara et al.). In addition to reducing missing data, using administrative datasets to "pre-fill" already known information into survey instruments reduces the burden in the interview.

Bader and his colleagues comment on how valuable multiple data sources are for addressing potential "same source" bias. When measures for both hypothesized causes and effects come from the same source, distinctive features of that source may lead to correlated error, which in turn may lead to biased estimates of causal effects in statistical analysis. When multiple sources are available, this problem can be avoided by substituting one source for the other. Multiple sources also make it possible to use sophisticated approaches to measurement such as multitrait multimethod and other types of structural equation models. A multitrait multimethod model is one type of model.

Finally, linked data make possible entirely new approaches to examining key social science concepts. O'Hara and her colleagues capitalize on linked data to explore family units. In their definition, families are sets of relationships that shift over time as unions are formed and dissolved, children are born, and family members move and die. Families are related to, but not synonymous with, households. It takes multiple datasets to identify family relationships and track them over time. In another example, Browning and his colleagues show that neighborhoods need not be identical for everyone who lives in a particular location. Other geographies may come into play. With GPS-enabled smartphones, it is possible to take an individualized approach to daily activity patterns and, in the process, redefine what is meant by "neighborhood."

Challenges

As the articles demonstrate, linking data from multiple sources provides many benefits. There are also challenges—How do we link diverse forms of data in a way that is accurate, cumulative, and accessible to a broad range of researchers?

As the articles show, the quality of the link between data sources is critical, as is clear documentation of any problems encountered, particularly when newer, nontraditional sources are used. For example, location information needed to link data may be missing or of questionable quality. Dai and colleagues impute geographic location of tweeters based on their self-reported city, state, and country location in the Twitter metadata. Of all the tweets in the 1 percent public access sample they were using, approximately 25 percent had country information, and of these, 16 percent were from the United States. In interpreting this statistic, it would be helpful to know whether tweeters in the United States were more or less likely to reveal their country of residence than tweeters from other countries. Whatever the percentage of U.S. tweeters who report that they are based in the United States, it will fall short of 100 percent, perhaps substantially so, and this will degrade links to other data sources. Fortunately, of those identifying as a U.S. tweeter, almost all (91 percent) supplied a state of residence, a much more reassuring statistic. Even when links exist, there may be questions about their quality. First, the choice of keywords can have a huge impact on the correlation and prediction results. How should researchers choose keywords? Is it based on review of the literature, directly extracted from other sources, or from domain experts? It is important to select and justify these keywords carefully. Second, the data are noisy. In the Dai et al. study, for example, nearly 6 percent of asthma-related tweets were from the top ten users and 19 percent had URLs, suggesting that some of these tweets might have been sent through twitter bots or spammers. Third, assessing sentiment from the language in tweets is not an exact science. Given the complexity of language, true emotions or feelings might not be captured.

An observatory or set of regional data centers would be an ideal place to develop such information and make it broadly available to users. Variables would be created and links performed and documented cumulatively. Indeed, doing so creates economies of scale. Many researchers are creating the same measures—tract-level measures of poverty, race-ethnic composition, and immigrant status—from the same sources. Making these measures available centrally would save time and money. It could also enhance the reproducibility of results, as common variables would be carefully constructed and assessed by experts.

Data could be assembled once, rather than separately by each project. For instance, Bader and his colleagues use Google Street View as a less costly alternative to neighborhood audits, but even this approach can be expensive. To keep costs down, street segments were sampled, only one side of the street was coded, interpolation techniques were used to generate a spatially continuous surface, and only four cities were analyzed. With a coordinated and centralized approach, it might be possible to include all of the street segments, not only a sample, for all of the country, not only four cities. That way, observationally based measures such as neighborhood disorder could be developed for variably defined spatial units (including custom definitions) and then used by many researchers, for many purposes. For example, one

can imagine that neighborhood disorder might be useful in Browning et al.'s assessment of adolescent exposure to violence. With automation, it is becoming increasingly feasible to think of national data coded at the street segment level. Whether variables developed from Street View could be made available in this way would depend on arrangements with Google and their willingness to allow it.

Special arrangements would be needed for other sources of data as well. For example, Fussell and her colleagues purchased the Spatial Hazard Events and Losses for the United States database from the University of South Carolina for their analysis of environmental hazards and migration. Costs are modest for academic users, but subscriptions for governmental, nonprofit, and especially corporate use are more expensive. The license agreement makes it clear that the data may not be shared further.[1] In cases such as this one, the observatory might provide easily applied protocols for linking to these data rather than the data themselves.

Access is central to our concept of a social observatory. The U.S. Census Bureau houses data not only from its data collections but also from federal agencies including the IRS, the Department of Housing and Urban Development, the Social Security Administration, and the Centers for Medicare and Medicaid Services, as well as state-level data on TANF, SNAP, and WIC. As O'Hara and her colleagues show, when combined, these data sources open up new approaches to the study of families and their complexities. However, access to these resources is limited to projects that benefit the Census Bureau and are conducted within their highly secure Remote Data Centers. We imagine a complementary resource in which data are more readily available. Well-designed and thoroughly tested approaches to sharing even highly confidential data are available that allow for a broader array of potential uses (e.g., see those developed by the National Longitudinal Study of Adolescent to Adult Health).[2]

Wrapping Up

Each article in this special issue is a valuable contribution in its own right. Each discusses the data sources and linking challenges associated with a particular social or behavioral science question of interest to the authors. As we discussed here, the articles are also valuable as a collection. They demonstrate a variety of approaches to linking diverse datasets, address the tradeoffs of temporal and spatial detail, raise important questions about access, illustrate a variety of challenges associated with the use of data not designed for research, and set an agenda for the future. Additionally, as a collection, they show many different ways in which linking data from diverse sources can improve analyses beyond the research possible with only one of those sources. They show how diverse sources with varying strengths might be leveraged and how linking them can improve research that is done with only one of those sources. New questions can be addressed. We know that social mobility and life chances vary from place to place in the United States; but why? What can be done to reverse the erosion of the middle class? Improved identification of vulnerable populations is critical if we

are to develop more compassionate public and private social welfare programs to serve these populations. Concepts such as family and neighborhood can be defined in innovative ways with new data sources. Indeed, family arrangements have changed dramatically over the past half century, and yet our ways of measuring and describing them have not.

Collectively, the articles offer broad commentary on the potential value of and challenges associated with the creation of a social observatory, especially the linked data that would be central to such a resource, and point to next steps. We can imagine using them and especially the data on which they are based, to quilt a time-place mosaic and have an efficient and national resource for improving science, policymaking, and services for the benefit of the American people. In broad strokes, we can paint a picture of what this resource might look like in the longer term. Consistent with federal policy, it will promote data sharing and re-use. In this way, it follows in the footsteps of the infrastructure social surveys that have served the needs of the social sciences so admirably for the past half century, including the National Longitudinal Surveys, the Panel Study of Income Dynamics, the American National Election Studies, and the General Social Survey (GSS). As one example, take the GSS, which is funded by the National Science Foundation. Its director estimates that there have been more than 27,000 articles, dissertations, books, and conference papers based on the GSS, and that each year, 400,000 students use the GSS in a class they take. In 2016, GSS data informed news coverage of such diverse topics as racial attitudes, child care, corporal punishment, and marital happiness. This model contrasts with the situation as it exists now with the new "big data." Many research projects using these data develop their linkages and measures independently, for their own purposes. In some cases, the same variables based on the same source are created again and again, a duplication of effort that we can ill afford. One such example is neighborhood poverty rates based on the ACS, a product of the U.S. Census Bureau that has been used by hundreds of studies. A centralized source for such measures would create efficiencies, and also, because of improved oversight, contribute to robust and rigorous social science. Each study represents one patch in the time-place mosaic referenced above. If researchers contributed the measures they create—i.e., other patches—based on some of the newer data sources such as Twitter, Google Street View, and administrative records, this could be of broad benefit to social science and society. The local variability in social mobility and life chances documented in the introduction to this volume, and on full display in the recent election, would be fully captured. When fully assembled, the set of observatories or data centers will provide the combination of national coverage, local depth, and temporal precision needed to advance our understanding of the American population.

Notes

1. http://hvri.geog.sc.edu/SHELDUS/docs/END_USER_LICENSE_AGREEMENT.pdf
2. http://www.cpc.unc.edu/projects/addhealth/contracts/add-health-contracts-homepage.