May 8, 2013 Social Observatories Coordinating Network (SOCN) Workshop: Summary Report

Prepared by Sandra Hofferth and Emilio F. Moran

July 23, 2013

List of network members:
J. Lawrence Aber, Psychology, New York University
Henry E. Brady, Political Science, University of California at Berkeley
Susan Cutter, Geography, University of South Carolina
Dalton Conley, Sociology, New York University
Catherine Eckel, Economics, University of Texas at Dallas
Barbara Entwisle, Sociology, University of North Carolina at Chapel Hill
Darrick Hamilton, Economics, New School for Social Research
Klaus Hubacek, Geography, University of Maryland
John Scholz, Political Science, Florida State University

Co-chairs: Emilio F. Moran, Anthropology, Michigan State University, moranef@msu.edu
Sandra Hofferth, Family Science, University of Maryland, Hofferth@umd.edu

Funding was provided under Grant SES1237498 from the National Science Foundation.

I. Current Projects that are Developing Consortia.

Robert Goerge, from Chapin Hall, University of Chicago, spoke about "Bringing together data on people and places to support policymakers (and do research)."

For the past 30 years Goerge has been working on an integrated database of Child and Family programs in Illinois. The data span from 1977 to 2013, but most begin in 1990. The data, unless otherwise noted, are statewide – the focus is not just on Chicago. The data base includes information on children in child protective services, recipients of welfare, food stamps, TANF, child care subsidies, and early childhood programs. To data on individuals and their characteristics from administrative records, including arrest and juvenile justice records, they have linked wage and benefit records so they could follow children into adulthood.

This group has established an Urban Center for Computation and Data at the University of Chicago with multiple partners: 18 different organizations and multiple individuals. The city of Chicago decided they would build a "windy grid" that pulls together a significant amount of information every 30 seconds on some issues and daily on others. Data come from 311 service requests. The UCCD takes the data and makes them accessible to the community. At the moment there are existing software programs that are taking the data coming out of the continuously updated grid to create specific reports.

As part of their analysis they have examined growth in children in different Census tracts in Chicago. Growth rates vary across the city; one analysis shows that growth is fueled by increases in the Hispanic population. Another analysis shows considerable variation in the ratio of preschool programs for children across census tracts.

One project looked at the proportion of at-risk or multiproblem families across the city. The problems identified were foster care, substance abuse, mental health, and juvenile and adult incarceration. From administrative data bases, children in families with multiple risks were identified. The results showed that 23% of families experienced 63% of problems and 86% of dollars were spent on these families. In addition, 73% of multiproblem family children had experienced abuse or neglect. In some areas in Chicago, 61% of children 0-17 lived in multiproblem families. The purpose of examining the spatial concentration of such families was to better target services in needy areas. The governor asked to look at the population of children and families receiving services from the state and whether there were geographic concentrations of these families that could be targeted for programs to reduce their use of services.

This group also followed children in the Project on Human Development in Chicago Neighborhoods through high school, using public school and education records, and into young adulthood, linking them to employment and wage data from the Census Bureau Research Data Center.

Many elements in this study of Chicago/Illinois address the types of data linking and spatial analysis envisioned in the observatory framework.

Meg Merrick, from the Institute of Portland Metropolitan studies, Portland State University, presented "Data, Indicators, Visualizations, and the Challenge of Coordination."

The Greater Portland Pulse project is a collaborative data indicator project that makes indicators available for the entire 7-county region around greater Portland, Oregon. These indicators are disaggregated by Census tract and state so that maps can be produced showing indicators for the region. Portland is unique in having a regional government that is very interested in obtaining regular and up to date performance measures in the region.

There are nine domains in which indicators are constructed: economic opportunity, education and employment, safety, arts/culture, civic engagement, health, national environment, housing, and access and mobility. Within each area are constructed a set of 64 indicators requiring 111 variables.

The second project they are conducting is called the Regional Equity Atlas 2.0. In this effort, the data are broken out by neighborhood, not just by Census tract. This project, which has been used to put equity on the table for regional discussion, was funded primarily by health related foundations, including the Robert Woods Johnson foundation, Kaiser Permanente, and the NW Health foundation. The data for health indicators came from the state of Oregon health plan. The data were aggregated by census tract by a third party and provided only in aggregate form to the equity project. Another innovation – multiple layering – laid indicators such as the rate of diabetes over healthy life style composites such as walkability and access to farm markets. The data are solely place-based for now.

Tom Kingsley, from the Urban Institute described the "National Neighborhood Indicators Partnership (NNIP)."

NNIP started in 1995 with 6 institutions that had put together data sharing arrangements in their cities. Today this neighborhood indicators partnership has expanded to 37 cities and 12 more are talking about joining. These include community and social service nonprofits, university, metropolitan planning agencies, and foundations; all have a substantial community involvement component, i.e. they are close to the stakeholders. Data are used for performance measurement, management and policy analysis.

Activities consist of continuous indicator review and dissemination of information over the web. Indicators at the city or metro-wide level can be used to change laws and policies; for geographic targeting/coordination of resources for programs and investments; for individual neighborhood improvement initiatives; for performance management; and for program evaluation. The types of data include national data files –e.g., American Community Survey; open data – government administrative data; and integrated data systems.

One important substantive area studied was the effect of foreclosures on children. One program linked foreclosures in neighborhoods to the number of children in those areas. They did this by linking parcel level data on properties and neighborhood level data.

II. Data needs and current small scale observatories from computational social science

Alex Pentland, MIT, Living Labs

Living Labs are basically passively instrumented communities. Data collection is intensive and continuous. Pentland argues that more data are better than less. In addition, continuous is better than occasional collection. He has been building instrumentation to get large amounts of data for people over long periods of time. He uses smart phones/cell phones to collect mobility, call patterns, texting patterns, physical proximity, physical interaction, credit card interaction, health stats, and sleeping statistics. By collecting vast amounts of data, you can classify people according to their interaction patterns, for example. He calls this "behavioral demographics."

As one application, he is finding that some behavioral groups have a higher risk for diseases like diabetes. He is working with Mass General (Hospital) to examine risk behaviors of some of the groups. Another application is to determine from their behavior when someone is ill.

He believes that collecting data on context is the most important. This certainly fits with the views of the previous presentations that have focused on geographic/physical context and patterns of behavior across geographic regions. He is attempting to define groups on the basis of behavior, not geography.

One of the important possibilities is that of mapping social ties and networks. He is exploring how incentives could be provided to members of an individual's network to provide assistance when needed. He found that network incentives were 4 times as effective as individual incentives. He has deployed a set of software tools, Friends and Family, to track people's network activities in real time.

Finally, Pentland has worked to help develop a consumer privacy bill of rights. Data would be stored on a personal phone and the individual would have the right to decide what data leave that system. The key requirements for a secure system are real informed consent, metadata that describes permissions and provenance for all personal data, and the ability to automatically audit the flow of data.

Duncan Watts, from Microsoft, "Using the Web to do Social Science."

To Watts the question that has always been of greatest intellectual interest is what sociologists call the micro/macro problem - how do you put a bunch of people together and get something out of it that isn't just a bunch of people?

Sociologists have been aware that where this process comes from is the network of interactions between people that influence each other and somehow all of these nudges aggregate up to produce collective behavior. That may be true, but it's a difficult problem to study empirically. It's hard to do science when you can't observe most of what you care about and you can't do experiments.

Three classes of data are generated by web platforms: observational, survey, and on-line experiments.

1. Observational data. A lot of data are out there, but they can't be used for causal inference. Facebook represents online behavior. One category of data is search data; it can indicate popularity. Another example is influence. We can measure how far one tweet spreads. Can have retweets of retweets. Of tweets, 90% are not retweeted, 8% have 1 retweet, 1% have 2 retweets.
2. Survey data. Survey data have the advantage that you can tell what people care about and what they are thinking. Hypothesis: people are less similar to friends than they think. Put their idea on the X axis and their friend's on the Y axis, then examine the association.
3. Experimental data. Put up an experiment on line and have people come. One experiment found that the more you pay, the more work people do. But pay has no effect on accuracy. Also, the easier the work, the more people do.

Questions regarding the vision statement:

1. Why does the observatory have to be located so strongly in geography? Can see how helpful it is to obtain the data. Once you get the data, however, they can sit anywhere.
2. Why sampling frame? Representativeness is an issue, but not the only one. If you get everyone, then it is not an issue.
3. The biggest issue is how you join these data sets that combine individuals. e.g., Facebook, Google, twitter, etc. Will need to get the corporations together with others to be able to link people across data sets.

Julia Lane, from American Institutes for Research, "Privacy and Confidentiality in Social Observatories."

Lane is editing a new book: "Confidentiality in Data Access and the Use of big Data: Theory and Practical Approaches." The lead unit that is funding it is the Center for Urban Progress. It's very similar in concept to what they're doing in Chicago. The notion here is how to link together massive amounts of data from the NY administrative data systems and also transaction data, including taxicab data and subway ride data, and potentially video camera data. How do you pull all that together and help you manage that city in a number of ways? What they want to do is also make them available to researchers at NYU; that is the context that is informing this volume. Authors in this book are trying to think through the issues that need to be addressed.

The context is considerably different from 15 years ago. Needed is a new policy framework – what is the legal framework whereby you can bring all these data pieces together?

   a. What does Informed consent mean? No longer have a piece of paper individuals can sign.
   b. How will Institutional Review Boards review projects? They are going through revision of the Common Rule that governs these boards.
   c. What does Data stewardship mean? Before we began collecting data, we tell you what we're collecting, how we're going to collect it, and only collect the minimum of PII (personally identifiable information) necessary. How do we do this?

2) Measuring and optimizing utility, risk and harm. Anytime you're going to allow access to it, there's a risk of re-identification, what does the risk look like and what harm does it

do. Re-identification is the major problem with which agencies are concerned. It is likely that utility is positively associated with risk; the greater the utility, the greater the risk. The conceptual framework should include both potential harm and the risk of harm. Who determines what the risk is? Then there are three types of protections: control over inputs, control over outputs, and control over access.

Current approaches are to provide: 1) Public use data files; 2) Research data center – secure locations allow restricted access; 3) Restricted file for government use; 4) License; 5) Synthetic files – randomized data; 6) Query tools; 7) Job submission; and 8) Data enclave with remote access. The secure access data enclave appears to be the direction we are going.

A suggestion was to have risk x harm x utility tradeoffs.

Bill Rand, from the Center for Complexity in Business, University of Maryland, Comments.

There are two types of data: Type 1 data represent observational data on large numbers of people with thin information about each person. Type 2 data represent fewer people but very extensive data on each. Type 1 provides aggregate information. Type 2 provides data for building behavioral theories. He builds models on type 2 data which can be used to simulate population data and compare the results with type 1 data.

Complex systems is a view of the world that says the best way to look at the world is to look at the interconnectedness b/w individuals and the emergent properties that come about as a result of those interconnections. It is an inherently interdisciplinary approach.

Bill spends most of his time collecting open source data - not proprietary data that's sitting behind government or company walls. For instance he created a tool which is an open source application for collecting data from twitter. It is something like a social observatory but with a much more sophisticated warehouse set up to collect this data. We've recently began a way of extending it to collect mainstream news on an interface where we can collect all this stuff together on a database setup that we're calling Meter. Duncan mentioned that all of these data sources are essentially architected. Facebook is independent of twitter is independent of YouTube for that matter. They're sitting in very different styles. We need to collect this at a central location; what we're doing is building a tool that simultaneously collects data and dumps it into a central repository in some common format.

We've collected data such as the death of Bin Laden, Hurricane Irene, Hurricane Sandy, Election 2012. We're currently working with Red Cross, for instance, to see how we can take geolocated Hurricane Irene data to inform calls for donations, particularly, where to extend their efforts. The real thing I want to talk about with Observatories is that a lot of data are geotagged so that you can actually grab information about these individuals based on their location and based on conversations. One example of use is the Mineral VA earthquake near DC. People were reading about the earthquake before it hit them, because the earthquake moves at the speed of sound whereas twitter moves closer to the speed of light. You get people outside of 100km who can read about it before they actually feel the events happening. These uses of geolocated twitter are very interesting types of social media data.

I think the SOCN should think critically about how to make that available: How to build Observatories in such a way as to make modeling an easier task to handle within those setups. And to prevent some of the biases that we talked about this morning. How do you control for that? One way would be to have consultants or staff whose goal is to make sure the data are used in a responsible way both from a modeling perspective as well as from a privacy perspective.

Peter Muhlberger, from the National Science Foundation

There are two sets of central social science questions: 1) conflict, cooperation, and apathy and 2) rationality and irrationality.

1. There's an enormous of work in psychology, sociology, political science, whatever social science field you think of (almost) that reflects on what brings about conflict, cooperation, and apathy, what changes the outcomes and what interventions, social changes and so forth may affect the outcomes. What are some of the consequences involved? Issues of equity and economic well-being are central here as are democracy and openness.

2. The second issue is that of rationality and irrationality. It is ultimately what makes our choices and beliefs reasonable or unreasonable. What are some of the factors that lead to rationality or irrationality: the mechanisms behind this, and how can it be changed?

I believe that there are 3 concepts involved: identity, belief systems, and context. How do they interact in real time and as complex systems? We need to understand belief system and context as parts of complex systems and real time phenomena. What's missing from our understanding is the context in which A causes B. In the real world, in the field there are many factors that work at once. Knowing that one thing causes another in a lab isn't going to be very helpful.

In reference to the observatory project, there are substantive questions that can be generated. How one can connect this project to underlying questions in the social sciences.

III. Discussion of Linkages between Research Consortia and Big Data

Myron Gutmann – Comments from the National Science Foundation

The NSF is fully supportive of this project and is expected to remain so. The federal statistical agencies should be involved in this discussion so that there is no overlap or duplication of effort.

Robert Kaplan, from the National Institutes of Health, "The Importance of Behavioral and Social Science Research."

NIH's goal is "Science in pursuit of fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to extend healthy life and reduce the burdens of illness and disability."

There are a variety of different analyses - about 60 of them in the literature - that look at determinants of health; all of them show that behavioral patterns and social circumstances account for about 40-50% of variation in life expectancy. Medicine and health care account for 10% to 20%, but not much more than that. When you take that mission statement of the NIH about extending life and reducing the burdens of those with disability, behavioral patterns and social circumstances are hugely important, as are environmental factors. US male and female

newborns can expect to lose about 1.4 years and .8 years of life before age 50, respectively, about double those of Sweden.

The Importance of Representation

The first point is that bigger is not necessarily better. In the 1936 election between Landon and Roosevelt, the Literary Digest did a poll of 2.4 million people. The Gallup organization did a poll <2% as large. The Literary Digest incorrectly predicted that Landon would win by a landslide. The Gallup poll was within 1% of the correct numbers.

Multiple Comparisons Problem

Beginning in 1999, clinical trials were required to register their primary outcome. Prior to the online registry, clinicaltrials.gov, the common understanding from the literature was that there was a significant benefit of drugs in reducing mortality. For primary outcomes other than mortality, there was an even bigger effect of drugs. All combined there was a 24% benefit of intervention. After clinicaltrials.gov was started in 2000, the situation changed dramatically. There are no trials that show benefit of drugs on reducing mortality. The clinicians are still trying to figure out what happened.

Keys under the Lamp post - Selection problem

A high percentage of papers published in medical journals are based on people who are hospitalized at medical centers. Even when we include those studies based on electronic medical records, they are based on less than a third of all people. Those are the people who are getting services or even fewer than that because they're getting services in medical centers.

Neurologists believed that if your child had a seizure with a fever, they would have a high probability of having other seizures. If you just look at children who come to the neurologist, a high proportion had previous febrile seizures. But fever seizures are common; if you examine the entire population of those with fever-related seizures, the chance of progressing to full seizures is very small.

The point – it is important to inform researchers about selection bias and sampling.

General Discussion

The discussion centered on the importance of using place-based observatories for data collection versus simply collecting all of certain types of data for the entire US and then locating it in places through geocoding. The consensus was that place-based observatories were important for local involvement and social capital that would lead to better collaborative data-gathering relationships in communities. Plus this would be more attractive to policy makers and sustainable. However, not all data collection needed to be so located.

There was additional discussion about the role of the American Community Survey in replacing the census long form and the difficulties of obtaining trends over time with its fixed sample size and 5-year moving average design.