|SPATIAL INFORMATION DESIGN LAB
Sarah Williams: Director Spatial Information Design Lab
Columbia University Graduate School of Architecture, Planning, and Preservation
1172 Amsterdam Ave \ NY, NY 10027 \ tel. 267.253.4124 \ email. sew@alum.mit.edu

APPLICATION FOR GIS SPECIALIST MEETING : DECEMEBER 13, 2007

Michael F Goodchild
National Center for Geographic Information and Analysis and Department
of Geography Ellison Hall 5707 University of California Santa Barbara,
CA 93106-4060, USA
Office: +1 805 893 8049
FAX: +1 805 893 3146
Mobile: +1 805 455 6529

Tuesday, September 25, 2007

Interaction with digital information has become part of daily life. We leave traces of data about ourselves everywhere we go. When we swipe our subway card - information is left about our travel patterns. When we use our cell phones, information is keep about where we used the phone, for how long, and at what time of day. When we connect to internet hubs, the volume of our use as well as what we visited is recorded. Mapping these data traces allows us to visualize the dynamic ways that people interact with the urban environment. The existence of “data traces” has become more pervasive as our lives have become more digitally oriented. Three of my recent research projects illustrate the possibilities for analyzing “data traces”. While “data traces” have inherent biases because they only record information about the people whom interact with digital technology, the data itself provides a resolution both in time and place, which is unprecedented. I am interested in testing the validity of using these “data traces” as a proxy for other survey methods.

Many of these “data traces” are documented by institutions, and therefore, have an accuracy level that reflects the interest of those who record the information. However, recently people have started geo-referencing photos and data they post to the web. Whether it is geo-tagging Flickr photos or RSS feed, these self-documented data traces provide a wealth of new spatially referenced information. At the same time the data itself has many validation issues. Those whom document information they post to the web do so for different reasons, therefore the accuracy of the information varies greatly depending on the context. Given these limitations I am interested in understanding how self-documented and geo-referenced data can be used for analyzing spatial patterns in the urban environment.

Looking at cell phone use in Milan provides and example of how curated “data traces” provide researchers with high resolution population information. Vodafone, the largest cell phone company in Europe, provided our research team with six months worth of spatially referenced cell phone activity data. Mapping cell phone activity levels across Milan, at every hour of the day uncovered Milan’s unique urban patterns. The data provided us with hourly population information, and allowed us to explore a different scale of the city from what census data can provide. Infrastructure planners could use this analysis to infer hourly urban densities, and therefore, create better plans for public transport or roadway restrictions. City Managers could use this real-time activity data to help create plans during emergency events. Cell phone activity maps explain one layer of Milan’s complex urban system, when complimented with additional mappings the data allows one to study the ecology of the urban environment. I am interested in understanding how well cell phone data can be used as a proxy for hourly populations and whether use of this data invades the privacy of people using cell phones.

Mapping New York City’s 311 call complaints provides another way to understand the ecology of the urban environment. New York City established the 311 call system for non-emergency and governmental calls. That means if New York citizens have a complaint - instead of calling 911 – they call 311. Calls requiring service are

Sarah Williams: Director Spatial Information Design Lab
Columbia University Graduate School of Architecture, Planning, and Preservation
1172 Amsterdam Ave \ NY, NY 10027 \ tel. 267.253.4124 \ email. sew@alum.mit.edu

logged by location and time. People call about everything from dead birds and potholes to juvenile loitering and noise control. Mapping these complaints tells us about current conditions in the city. For example high numbers of rat complaints are highly correlated with health code violations.

By studying noise, missed trash pick-ups, and homeless person 311 complaints I have found that the data tells more about the complainer than about the particular city condition. For example if you look at missed trash pick-up complaints. Callers from all over the city complain about missed trash pick-ups. However, it appears that fewer calls received in from low income /high minority areas. Does this mean that neighborhoods of high poverty have better service? It might, but is also might indicate that people from these neighborhoods are not aware that they can call 311 to complain. The City of New York is increasingly interested in using 311 data for management purposes. Therefore it is important that the caveats of this data are explored. I am interested in looking more closely at how complaint "data traces" can be used to manage city infrastructure, given that they may leave out large portions of the population.

Records of how people interact with the urban environment are not only self initiating, like making complaint calls to 311. Data is also recorded by people chosen to observe the urban environment. An example of this is illustrated by a recent study in which I analyzed the spatial patterns of photographs sold by Getty Images. Photos purchased from Getty are provided with information about where photos are taken and what type of event/topic they cover. Analyzing image locations in the Getty database showed clustering of arts and culture events in New York City and a dispersion of these events in Los Angeles.

While Getty images have verified locations because the photographer must confirm event addresses with the Getty organization, Flickr photos and RSS feeds are self-documented leaving validation up to the individual whom posts the information. The wealth of information provided by this self-documented data needs to be explored. I am becoming interested about the possibilities of using this data to understand spatial patterns in civic engagement. This past summer I taught an experimental course where I asked my students to mine the geo-tagged Flickr photos to determine, if like Getty images, the photos could tell us about hot spots for arts and culture. Mining the photos did provide some useful information, however we also found that many photos were tagged incorrectly. Given this limitation it appears that this self-documented and geo-located data does have some potential for analysis, but there are issues with documentation accuracy. These limitations may be related to the "newness" of the technology rather than the future possibilities of the data itself. As geo-tagging becomes more pervasive, mining images found on the web may allow self-documented photos to be used for new forms of analysis like the Getty images. Until that time it is essential that we determine how to work with the validation caveats created by this data.

"Data traces" can be curated or self-documented. As the existence of this data is becoming more prevalent it is essential that we understand the accuracy of both types of "data traces". Similar to most data "data traces" have inherent biases. For example cell phone data only tells us about those using their cell phones, and geo-referenced Flickr photos only provide information about those who know how to geo-reference their photos. Beyond these biases there are questions related to how each data set can be validated. How can others validate self-documented data sets? How can we use this data given these limitations? I am interested in attending the GIS Specialist meeting because of these questions. I believe the success of my future work is directly linked to the questions the group will be trying to address.