A Harvest of the Accessible: Potentials and Limitations of Utilizing Web 2.0 Data for Urban Analysis

User-generated digital information, or Web 2.0 data, facilitates a direct, digital connection of billions users. First defined in 1999, the term Web 2.0 describes a method of data transfer that is interactive and multi-directional, a departure from the static web pages of the initial internet. (Wikipedia, 2013A) Web 2.0 technology allows users to converse and exchange information through web apps that create user-generated content.

ALEXANDER WEBB
University of New Mexico

The content ranges from text to images, frequently tagged with geographical information.

In *The Image of the City*, Kevin Lynch wrote "Environmental images are the result of a two-way process between the observer and his environment. The environment suggests distinctions and relations, and the observer… selects, organizes, and endows with meaning what he sees" (Lynch, 1960). If designers could leverage Web 2.0 technology to retrieve user-generated content and filter it relative to geography and relevance, this relationship has the potential to be flipped; the environment, through the guidance of the designer, would select, organize, and endow meaning to the behavior of the inhabitants and the inhabitants would suggest distinctions and relations.

This inversion would allow urban interventions to be more intelligent and informed, gathering information from use patterns and experiential data, produced by users who are unaware their actions are being observed and collected for these purposes. This subversion of the Observer Effect, the principle that by observing an element inherently changes its behavior, promises a broader and more accurate form of inhabitant input than previously available.

While the potential of Web 2.0 data is seductive, much of the information available is far from comprehensive or relevant. This paper will clarify conditions surrounding existing Web 2.0 data, describe several approaches of leveraging existing data sources for design processes, and discuss potential implications of utilizing similar techniques with future information technology.

WEB 2.0

Web 2.0 technology facilitates a submitting data on their experiences, observations and locations; the geotagged locational information being particularly relevant for architects and urban designers. Geotagging allows for the user's locational data to be published as metadata alongside the information being shared (Hannay & Baatard, 2011). Positioning user-generated content relative to a geographic coordinate system enables a mapping and analysis of the behaviors, tendencies and experiences of a population relative to their urban condition. The relevance of this data is only limited by the Web 2.0 technologies currently available, the quantity of information and the user demographics.

With over one billion registered users, Facebook is the largest virtual community (Wikipedia, 2013B) and the most robust Web 2.0 technology in terms of user base. Facebook allows posting of text, images, video and file uploads- all with the ability to geotag. Though Facebook produces vast amounts of user-generated data, geotagged information by default is only available to direct connections of the user (Hannay & Baatard, 2011), making retrieving the data unfeasible. Google+, the second largest US-based virtual community (Wikipedia, 2013B), allows similar capabilities but with the same limitations relative to retrieving the geotagged information (Hannay & Baatard, 2011).

With over 500 million users, Twitter is the third largest US-based virtual community (Wikipedia, 2013B). In 2011, Twitter gained notoriety for its role as an organizational device in the Egyptian portion of the Arab Spring. "The speed of communication through digital channels gives activists unprecedented agility during street operations. Online, they can organize, debate, plan, and broadcast at a level of coordination that was unavailable, indeed unimaginable, in the past." (Wolman, 2013) Twitter has also been credited with facilitating the organization of the Occupy Wall Street Movement (Conover et al., 2013).

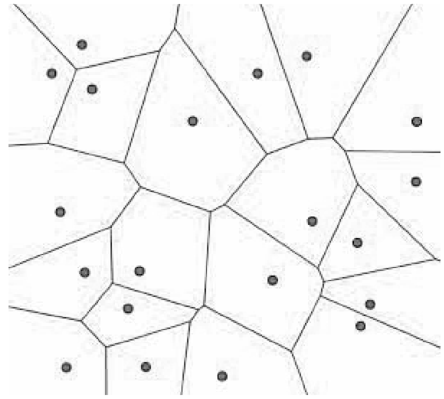
By comparison Twitter is a more streamlined platform than Facebook and Google+, allowing only text postings of a maximum of 140 characters. Users may choose to geotag their posts (referred to as "tweets"), which are then searchable by location. Though Twitter generates 400 million daily tweets (Moore, 2013), one 2010 study estimated that only .23% are geotagged (Bryant, 2010). The inability to search tweets more than ten days old (Twitter, 2013), combined with sparse geotagged data produces little information; at the time of this study a search within a ten mile area of Seattle, Washington produced only two geolocated tweets- not enough to provide meaningful insight into urban behavior.

STUDY OVERVIEW

Eliminating the sources already discussed in this paper, other relevant sources of user-generated data were determined. Sources were selected by relevance, accessibility of information, and amount of data available for the South Lake Union neighborhood of Seattle, Washington. This paper will describe investigations utilizing data from Flickr, MapMyRun, Google Maps, Yelp and Craigslist. This study does not mean to imply these sources are without biases and inconsistencies, but rather pose methods of leveraging the information that is available.

All of the studies involving these sources create a geographic datum for positioning their information. Leveraging the web app OpenStreetMap and it's ability to export .xml files of contextual information, base map data is brought into the Rhinoceros 3D plugin Grasshopper. Grasshopper is a graphical algorithm editor that enables computational design strategies within the Rhino environment (Grasshopper 3D, 2013). The OpenStreetMap .xml data is imported into Grasshopper through Elk, a Grasshopper component that utilizes OpenStreetMap and Shuttle Radar Topography Mission data to create intelligent vector maps. Objects within these maps are tagged with metadata which identifies the objects relative to OpenStreetMap categories- highways, property lines, bodies of water and other contextual elements within an urban environment. The metadata allows a variety of geometric operations on the disparate categories, which assists the development of a three dimensional map within Grasshopper.

Data harvested from the various sources of Web 2.0 data is positioned using the Grasshopper component gHowl. gHowl enables a translation of geo-coordinates into the Cartesian coordinate system of Grasshopper, as well as an export of Grasshopper-generated geometry into Google Earth. Google Earth enables a flexible, intuitive interface with the user-generated data that is dynamic, scripted, and embeddable, providing an ability to directly interact with the information within an articulate contextual environment.



1

2

Though many of these searches could have been automated to stream real-time data, these studies were conducted by architectural students with little to no programming experience. Accessibility and ease of data management were priorities in these inquiries.

GEOMETRIC STRATEGIES

Many of the inquiries were faced with a similar problem: how to assemble a set of points that describes relationships and tendencies. To solve this problem the metadata become a critical source of filtering, aggregating, and segregating different data relative to behavior and identity.

One strategy is utilizing the points to form metaballs. Metaballs, as defined by Greg Lynn, are geometric forms that "are surrounded by halos of influence. The inner volume defines a zone within which the meta-ball will connect with another meta-ball to form a single surface" (Lynn, 1995). As points aggregate, their metaballs aggregate as well to form larger metaballs. The metadata serves as a filter, separating different datasets by attributes, creating metaballs generated from points with shared parameters.

Figure 1: Voronoi Diagram. From ams.org.

Figure 2: Geo-Positioned Flickr Data.

3

Another strategy for formalizing points is to utilize a voronoi diagram. The segments of the Voronoi diagram are all the points in the plane that are equidistant to the two nearest sites (Wikipedia, 2013C), creating a cell around the point that is representative of all of the area that is less than half the distance to any other point. The voronoi diagram delineates "territories," regions based off of proximity.

FLICKR

The image and video hosting website Flickr has over 83 million users (Wikipedia, 2013D) and hosts over 3 billion images (Kaplan, A. M., & Haenlein, M., 2010). Though only 2 million images and videos are geotagged (Flickr, 2013), Flickr is one of the larger sources of accessible, geotagged user-generated content.

While some Flickr images are geo-tagged, the data is not necessarily applicable for making architectural or urban decisions. Images from Flickr have been leveraged for understanding existing architectural space (Christenson, 2011), but provide limited guidance when attempting to develop spatial constructs. The range of photographic subjects that are represented within geotagged images can be incredibly varied, sometimes not representing any information of the urban context at all.

Rather than attempt utilize the image themselves, the geotagged information was used to position image tags. A tag within Flickr is metadata attached to an image that allows other users to find an image through keyword searches. The Grasshopper plugin Mosquito has the ability to search Flickr for keywords and return all of the geo-locations of photos with their keyword tags. These locations are mapped to evaluate for trends in the disposition of the neighborhoods. Happy, joyful, fun, interesting, beautiful, elegant, and surprised are identified as positive keywords. Sad, cry, tear, depressed, angry, suck, and disgusting are identified as negative keywords. In both categories the collection of keywords aggregate into nodes, the nodes are connected to show connections between nodes. Metaball geometry is generated as an overlay to describe density of the different sets of keywords within nodes.

4

Figure 3: MapMyRun Vector Aggregation.

Figure 4: Keyword-Filtered Google Maps Data.

MAPMYRUN

MapMyRun is an application that allows users to track running routes via Global Positioning System-enabled (GPS-enabled) cellphones. Once uploaded, the routes are organized by location and shared with other athletes. While applications such as Strava and Ride With GPS provide similar services, MapMyRun allows the routes to be downloaded in both .gpx and .kml formats. The ability to download Google Earth .kml files of the routes facilitated a quick, efficient and accurate importation into Grasshopper.

Once the MapMyRun routes for South Lake Union are established, the .kml files are read into Grasshopper by gHowl. The vectors of the routes are divided into points, and the points are aggregated as metaballs to examine tendencies and use of routes. In this study, the metaballs describe a density of points by expanding in the Z axis and attaching to similarly sized point clusters. The height of the metaballs are further described by a color coding of the metaball rings for graphic clarity.

5

The geometry of the metaballs is then combined with additional geometric information to examine how the urban layout may influence running patterns further. Major roads and highways generate planes that interrupt and contain running behavior, parks are described as nodes of possible attraction and activity. The mapping of the routes shows the tendencies of local runners, and the mapping of parks, roads, and major roads begins to clarify the drivers of these tendencies.

GOOGLE MAPS

A web mapping service provided by Google, Google Maps is the most ubiquitous mobile app, used by 54% of all smartphone users (Smith 2013). Google Maps' robust platform is accentuated by route finding, satellite imagery, and business reviews. While many apps such as TripAdvisor and Urbanspoon provide similar reviews, Google Maps was selected for this study due to the app's facile interface that easily locates reviews of businesses based off of geographic location.

Google Maps' interface allows a synchronization and cross-referencing with the data available through the imported OpenStreetMap .xml information. OpenStreetMap

Figure 5: Food-Access Voronoi Diagram.

building footprints are tagged in Grasshopper with text information describing identity of the businesses, creating an intelligent vector map of business entities.

The Google Maps reviews were evaluated for predetermined keywords. The keywords are selected to support a conceit of behavioral regions and sub-cultural territories, but could have been selected to support any design or evaluative intent. Points derived from the vector footprint of the business are tagged with the keywords based off of use within the review and aggregated. These aggregations were used to drive metaballs which vary in height relative to the keyword's frequency of use. Graphical text tags of the keywords are positioned relative to the metaball aggregations in Grasshopper space to illustrate the terms that are aggregated. A second set of similar keywords were collected from the businesses' websites to drive cultural positioning further.

YELP & CRAIGLIST

Similar strategies were utilized with the information provided by the websites Yelp and Craigslist. Both sites provide locational information that can be mapped to understand behaviors and tendencies.

The local search website Yelp received average of approximately 108 million monthly unique visitors in the second quarter of 2013 (Yelp, 2013). Yelp provides business reviews similar to Google Maps, but allows a categorical search that enables greater specificity in returns. Users are able to search for specific business, such as "Taco Stands", "Dry Cleaners", or "Breweries." These searches then return businesses within that category for a given region, city or neighborhood.

In an attempt to better understand food availability of South Lake Union, grocery stores, convenience stores, bodegas and delicatessens in the neighborhood were located through Yelp's categorical search. These locations are plotted as points relative to the OpenStreetMap information, then the points are used as drivers of a Voronoi diagram. The Voronoi diagram demonstrates the proximity of one food vendor to the other, and divides South Lake Union into simple regions based off of which vendor is the closest to a given area within the neighborhood.

The classified advertisement website Craigslist provides location-based advertisements for over 700 cities in 70 countries, and receives over 50 billion page views per month from over 60 million users (Craigslist, 2013). Craigslist's interface allows a filtering by date and time, and enables search listings by keywords. These searches return content in text-based lists, image-based grids, and a location-based map. The information produced by location-based returns were mapped as an attempt to identify nodes of activity and behavioral tendencies within South Lake Union.

STUDY ASSESSMENT / FUTURE STEPS

The usefulness of location based information available is limited by precision of location, the selective nature of user-generated content, and the relevance of the content to the location. More comprehensive content aggregators such as Facebook and Twitter either do not provide locational information or allow users to censor it. More precise locational information providers such as Flickr or MapMyRun only provide a narrow range of content. Web 2.0 information provides limited implications to behavior within an urban condition at best, not nearly comprehensive or precise enough to begin to understand large-scale urban use patterns.

Future Web 3.0 technology could provide enough locationally-precise data to help these strategies become useful. Web 3.0 technology is difficult to define,

ENDNOTES

1. Briscoe, D. & Hadilou, A. (2013, May). Collective Intelligence: An Analytical Simulation of Social Interaction with Architectural System. In Open Systems: Proceedings of the 18th International Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA) 2013 Singapore (pp. 375-384).
2. Bryant, M. (2010, January). Twitter Geo-fail? Only 0.23% of Tweets Geotagged. In The Next Web. Retrieved from http://thenextweb.com/2010/01/15/twitter-geofail-023-tweets-geotagged/.
3. Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. ICWSM, 10, 10-17.
4. Christenson, M. (2011). On the architectural structure of photographic space. Architectural Science Review, 54(2), 93-100.
5. Conover, M. D., Ferrara, E., Menczer, F., & Flammini, A. (2013). The Digital Evolution of Occupy Wall Street. PloS one, 8(5), e64679.
6. Craigslist. (2013). Retrieved from http://www.craigslist.org/about/factsheet.
7. Flickr. (2013). Retrieved from http://www.flickr.com/map.
8. Grasshopper 3D. (2013). Retrieved from http://www.grasshopper3d.com/.
9. Hannay, P., & Baatard, G. (2011). GeoIntelligence: Data mining locational social media content for profiling and information gathering.
10. Hendler, Jim. "Web 3.0 Emerging." Computer 42.1 (2009): 111-113.
11.1 Jamison, J. (2012). Web 3.0: The Mobile Era. In Tech Crunch. Retrieved from http://techcrunch.com/2012/08/11/analysis-web-3-0-the-mobile-era/.
12. Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (pp. 56-65). ACM.
13. Kaplan, A. M., & Haenlein, M. (2010) Users of the world, unite! The challenges and opportunities of Social Media, Business horizons, 53 (1), 59-68.
14. Khondker, H. H. (2011). Role of the new media in the Arab Spring.In Globalizations, 8(5), 675-679.
15. Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (591-600). ACM.
16. Lynch, K. (1960). The Image of the City. Cambridge: The Technology Press & Harvard University Press (6).
17. Lynn, G. (1995). Blobs. In Journal of Philosophy and Visual Arts, (6).
18. Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011, July). Understanding the Demographics of Twitter Users. In ICWSM.
19. Moore, H. (2013, September). Twitter files for IPO in first stage of stock market launch. The Guardian. Retrieved from http://

as there is a wide range of potential definitions of Web 3.0 information ranging from unified information structures to behavioral adaptation and personalization (Wikipedia, 2013A). What could be useful regarding Web 3.0 data are the concepts of increased locality, ubiquitous connectivity, and linked data.

GPS enabled mobile devices enables a record and report of location-based data. This locational awareness allows proto-Web 3.0 interactions, some citing the location-based social networking app Foursquare as an early Web 3.0 app (Wikipedia, 2013A). Foursquare facilitates a level of connection based off of location, enabling users to communicate their commercial and experiential habits through location-based check-ins. As more apps integrate location-based capabilities, patterns and habits of users will be more readily accessible for analysis. This accessibility is facilitated by the ubiquitous connectivity of mobile devices, providing the opportunity to stream data in real-time to information archives.

The most advantageous element of Web 3.0 information for urban analysis will be the linking of data. A synonym for Web 3.0 is the Semantic Web, a linking of data through a common format (Wikipedia, 2013E). These technologies are largely based on mashups that occur at the data, rather than application, level (Hendler, 2009). If this component of Web 3.0 is realized, partitions between apps and websites will be removed, and apps will begin to speak with each other. This facile communication between apps is already evolving, a notable example being the connection between the GPS tracking app Strava and the photo-sharing app Instagram. Similar to MapMyRun, Strava tracks athletes as they bike or run on various routes. The connection to Instagram allows any image posted to Instagram during a ride to be posted on the Strava ride page, which can be publicly viewed.

This connection describes the intelligence gathered by enabling apps to exchange data between themselves. Determining what urban elements cyclists look at while on a ride would be difficult utilizing keyword searches through Flickr, but Strava's filtering of specific Instagram images enables a more intelligent selection based off of behavior.

Web 3.0 technology will not provide comprehensive urban behavioral data, but an increased location-based data stream through ubiquitous connectivity will address it. Web 3.0 linked data will not insure that all information is relevant, but it will help produce a more useful dataset by enabling a more robust behavioral and geolocated contextual analysis.

CONCLUSION

This paper identified the value and limitations of Web 2.0 information, discussed several attempts at harvesting existing data, and described the possibilities of Web 3.0 technologies. While available data sources provide limited insight into urban behavior, this paper posits that Web 2.0 information be treated as a testing ground for analytical methods as user-generated information becomes more pervasive and connected.

ACKNOWLEDGMENTS

The author would like to thank Amy Brasfield, Natalia Cundari, Catherine Harris, Stevie Hartmann, Dale Lusk, Michael Salinas, and Zhu Zhu. The author would also like to thank Timothy Logan for his work developing the Elk module, Luis Fraguada for his work developing the gHowl module, and Carson Smuts for his work developing Mosquito.

www.theguardian.com/technology/2013/sep/12/twitter-ipo-stock-market-launch?CMP=EMCNEWEML6619I2&et_cid=48826&et_rid=7107573&Linkid=http%3a%2f%2fwww.theguardian.com%2ftechnology%2f2013%2fsep%2f12%2ftwitter-ipo-stock-market-launch.

20. Barbara Ortutay "Live Action: Twitter Grabs Super Bowl Spotlight". Associated Press. Retrieved 5 February 2013.
21. Salim, F. D., Mulder, H., & Jaworski, P. (2010, October). Demonstration of an Open Platform for Tangible and Social Interactions with Responsive Models. In Association for Computer Aided Design in Architecture (ACADIA) 2010 Conference-Life in: formation, New York, US (pp. 21-24).
22. Smith, C. (2013, September). Google+ Is The Fourth Most-Used Smartphone App. In Business Insider. Retrieved from http://www.businessinsider.com/google-smartphone-app-popularity-2013-9#infographic.
23. Twitter. (2013). Retrieved from https://dev.twitter.com/docs/using-search.
24. Wikipedia. (2013A). Retrieved from http://en.wikipedia.org/wiki/Web_2.0.
25. Wikipedia. (2013B). Retrieved from http://en.wikipedia.org/wiki/List_of_virtual_communities_with_more_than_100_million_users.
26. Wikipedia. (2013C). Retrieved from http://en.wikipedia.org/wiki/Voronoi_diagram.
27. Wikipedia. (2013D). Retrieved from http://en.wikipedia.org/wiki/Flickr#cite_note-3.
28. Wikipedia. (2013E). Retrieved from http://en.wikipedia.org/wiki/Semantic_Web.
29. Wolman, D. (2013) Facebook, Twitter Help the Arab Spring Blossom. Wired. Retrieved from http://www.wired.com/magazine/2013/04/arabspring/.
30. Yelp. (2013). Retrieved from http://www.yelp.com/about.