Christopher Round
Urban Ecology
Final Research Paper

Harnessing the Power of Citizen Science and Open Data for Urban Ecology

Abstract: Citizen science is both a research methodology and a social scientific movement that is growing rapidly. Citizen science differentiates itself from other research methodologies in that it involves the general public in data collection and interpretation. This movement has the potential to be intertwined with the open data movement, and thus in a sense "democratizing" science. Citizen science methodologies could be particularly potent for the field of urban ecology. "Urban" environments contain greater concentrations of citizens and thus a greater potential number of citizen scientists. Fortunately, the tools such as cloud computing and ecoinformatics have matured in time to take advantage of the growth of citizen science. The purpose of this research paper is to summarize how these tools, when combined with the open data movement, can benefit the study of urban ecology. As a test case, utilizing the citizen generated data set developed at the Cornell Orthinology laboratory (known as eBird); I investigated the number of species found per observer. The results indicated that Suffolk County may be effectively saturated with birders, leading to a thorough census of the biodiversity of the area. When it is considered that this field work came at little cost to researchers, state regulatory bodies, and other traditional research institutions, it reveals how powerful citizen science can be.

Introduction:

Urban ecology and citizen science are both rapidly growing aspects of ecological science. Urban environments with greater concentrations of citizens provide a greater number of potential citizen scientists. Two-thirds of the world human population is predicted to be living in urban areas by 2050 (Norbert & Werner, 2010). As the human population has journeyed into urban environments, native and nonnative biota has followed them. Thus biodiversity within towns and cities will play an

important role if biodiversity is to be protected (Norbert & Werner, 2010). This is especially pertinent as most of the predicted urban population growth is expected in fast-developing countries in South America, Africa, and Asia in areas near global biodiversity hotspots.

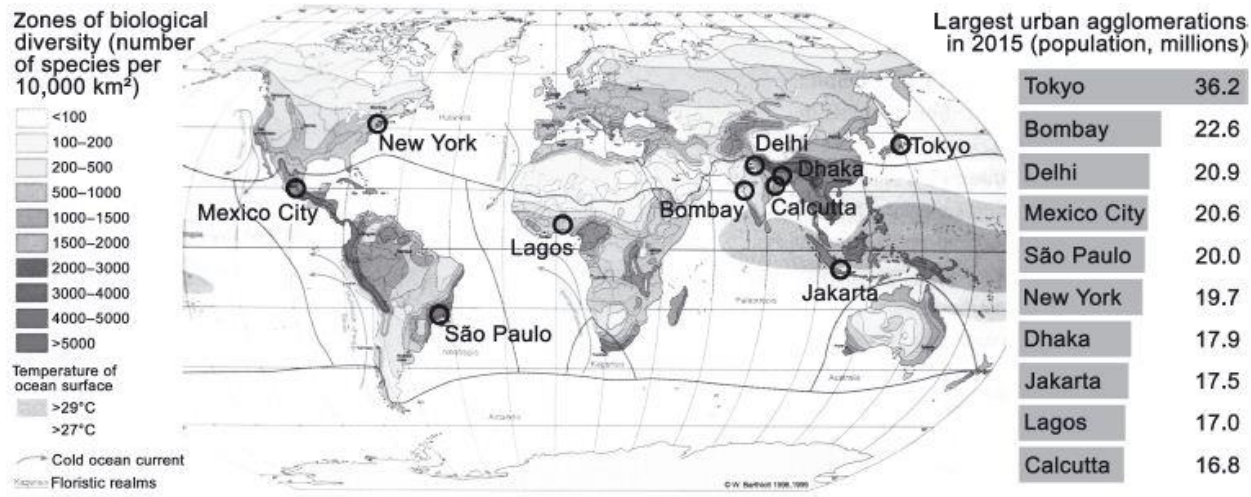


Figure 1: Hot spots of global biodiversity and the 10 largest urban agglomerations in 2015. Map amended from (Barthlott et al, 1999) for (Norbert & Werner, 2010).

Citizen science

Citizen science is not a new aspect of research, as demonstrated by the National Audubon Society's annual Christmas Bird Count (dating back to 1900) (Bonney et al., 2009). Citizen science is the conducting of scientific research or collection of data by amateurs, in collaboration with professional researchers (Francis & Chadwick, 2013). This could be particularly important for conservation efforts. The conservation of a species starts with understanding its distribution, abundance, habitat preferences, and movements across geographic areas and long periods of time (Hochachka et al., 2012). Typically only human observers can identify organisms to the species level (Hochachka et al., 2007). Intensive surveys by expert observers while accurate, are expensive and can be impractical due to limited availability (Hochachka et al., 2012). Citizen science differentiates itself from other methodologies in its ability to generate and/or analyze massive volumes of data over a short period of time. Unlike expert surveys, citizen science projects can cover large spatial extents, across many

years (Hochachka et al., 2012). One of the inherent problems of citizen-science projects for other environmental fields is a bias towards regions with high human populations (Kelling et al., n.d.). This is not a problem but inherently a boon for urban ecology researchers.

Open Data: The open data movement runs on the idea that certain data should be available for use by everyone and republished as they wish without control mechanisms. This approach has been argued to be particularly useful for science, as it hypothetically could enable a greater number of individuals to analyze data. The idea began to gain popularity in the scientific community in the 1950s with the development of the World Data Center system. The advent of the internet has reduced the cost and increased the efficiency for open data systems. This also provides an additional layer of accountability to those actively publishing using open data sets, as it is easier to demonstrate fault in the event of faulty data analysis. Open data could be especially useful for younger scientists, who often do not yet have access to the resources needed to generate their own data sets in their respective areas of interest. An example of an open data project that was successful was the galaxy zoo project.

Cloud computing: Cloud computing could be a major lynchpin in marrying the citizen science movement and open data efforts. Cloud computing is when software, platforms, or infrastructure are provided "as a service" over a network (commonly the internet). New services such as github make it easier for collaborations to occur on large data sets. An important example for this paper was Indiana University providing infrastructure as a service in the form of Quarry. Quarry is a Linux based super-computing cluster capable of handling extremely large data sets. It was utilized for the experimental portion of this paper. The University provides Quarry as a computing environment that can be accessed remotely.

Ecoinformatics: What ultimately makes all of this useful is the rise of ecoinformatics. Ecoinformatics offers tools for managing ecological data and transforming it into information (Michener & Jones, 2012). Conventional ecological analyses of species occurrence based on statistical techniques are capable of identifying changes in the occurrence of important subsets of variables (Species et al., 2006). Traditional statistical techniques can be overwhelmed by complications such as missing data, the need to fit non-linear effects, and potentially many interactions (Species et al., 2006). Data mining tools offer better flexibility and near automatic application when compared to traditional statistics techniques (Species et al., 2006).

Examples of Citizen Science, Open Data, and the Cloud:

Galaxy zoo is a citizen science project that combined open data and citizen scientists. The Sloan Digital Sky Survey compiled a list of more than 1 million galaxies. The only way to classify each galaxy was to look at them individually. Hoping to eventually investigate the formation and subsequent evolution of galaxies, Lintott et al, invited the general public to visually inspect and classify the galaxies in the survey (Zooniverse Development Team, 2014). Over 100,000 volunteers were engaged in the project (Zooniverse Development Team, 2014). Over 24 published articles have been produced from the Galaxy Zoo project between 2008 and 2012 (Zooniverse Development Team, 2014).

An example that would be more pertinent to ecology is the app "Instant Wild". Instant Wild is an example of citizen science and cloud computing coming together. Instant wild is an iphone application that acts as a cloud client, enabling users to browse a hosted data set made up of pictures gathered from camera traps. Users then browse the pictures, identifying wildlife for the data set owners. The project is part of the EDGE of Existence program, which is overseen by the Zoological Society of London. EDGE stands for Evolutionarily Distinct and Globally Endangered Species. The Edge of Existence program hopes to identify at risk species and map their locations in

order to help develop better conservation programs. It should be noted that unlike Galaxy Zoo or eBird, EDGE does not have an open data aspect.

What does this all mean for Urban Ecology?

Urban ecology has the potential to take advantage of the benefits of citizen science. Urban ecology as a field has the inherent advantage of being a far more convenient field for potential volunteers. Simply put, what is being studied is near people. This has already been recognized in the well publicized "Float" project, which hopes to take advantage of the large number of kite enthusiasts in China in order to better monitor air pollution (Solon, 2012). Urban ecology research into the impacts of cat predation on urban birds, butterfly biodiversity, micro site use by urban bees, human interactions with coyotes, and garden use by mammals and birds have all been investigated (Chadwick, 2013). Open data sets provide greater opportunities for researchers to investigate urban biodiversity, and allows trained interested citizens to investigate on their own. Ecoinformatics provides the technology and methods to data mine these massive data sets. This should accelerate the generation of knowledge in the field.

The Test Case: Birding is a popular hobby in the United States. Birders over time have the opportunity to become experienced naturalists who are capable of providing quality sightings data for data analysis. Recognizing this, the Cornell Laboratory of Orthinology developed one of the most well-known citizen science projects: eBird. It allows birders to submit their observations via the internet through handheld devices. All data is free and accessible (with permission) from the Avian Knowledge Networks. By building tools that appeal to and provide a service for birders, researchers were able to stimulate rapid and sustained growth of the data set. It was estimated that by 2011, the number of checklist submissions had exceeded 1.8 million from more than 210 countries (Wood, Sullivan, Iliff, Fink, & Kelling, 2011).

Such a large data set puts substantial knowledge into the hands of researchers. Occurrence maps such as figure 2 would be extremely expensive to generate utilizing expert driven analysis. Analyzing urban avian biodiversity is useful for more than just demonstrating the power of citizen science. The increasing interest in reconciliation ecology for the city requires an understanding of what biota is in the city. Urban environments have been demonstrated to hold novel assemblages. We must understand the ecology of the city in order to prescribe ecological solutions for the city. To demonstrate the potential of citizen science, open data, and informatics for urban ecology, I opted to examine a species survey in Suffolk County, Massachusetts.

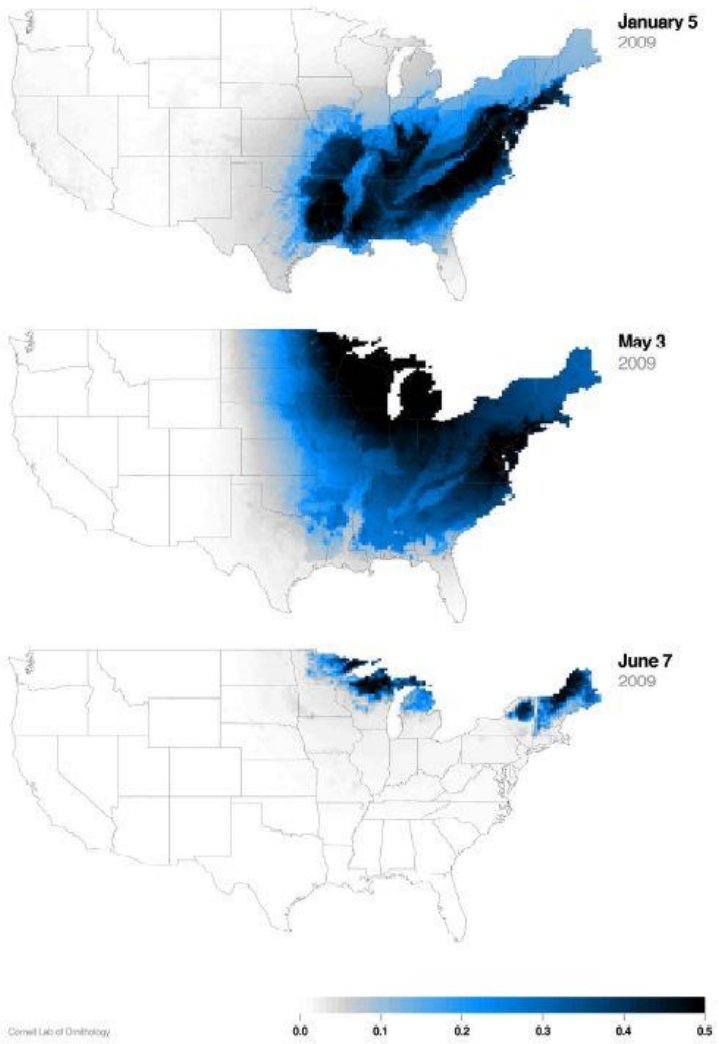


Figure 2: White-throated Sparrow occurrence maps generated by the eBird data set. (Wood et al., 2011)

Methodology:

The dataset contains count data for bird species observed by novice and experienced bird observers (a.k.a.birders). The data was submitted by volunteers to the eBird Citizen Science Project, run by the Cornell Lab of Ornithology and the National Audubon Society (Cornell Lab of Ornithology, 2014). A record in this dataset corresponds to a checklist that a birder uses to mark the number of birds of each species detected; one checklist is submitted per sampling event (i.e. birding session). Each checklist submitted from the 48 states in the contiguous United States is additionally annotated with

hundreds of predictor variables (called covariates below) that are derived from the location of the sampling event (Munson et al., 2012).

Data analysis was performed utilizing the Quarry supercomputing structure at Indiana University. Computation was done utilizing a Bash Linux shell. Information for Suffolk County in Massachusetts was extracted for analysis. Massachusetts had a large number of participating birders, and Suffolk County contains the city of Boston, Chelsea, Revere, and Winthrop. Suffolk County has a human population density of 12,721 people per square mile, qualifying it under most definitions as "urban". Data for a twenty year period (1993-2013) was loaded into excel, and was analyzed using pivot tables. Figures were created investigating the number of species found over time, the number of participants over time, and the ratio of species to participants.

Results:

As time went on, more species were found (figure 3) and more participants were engaged. The fewest species and observers occurred in 1993. After 1994, the number of species found hovered between 150 and 200 until 2005 when it steadily climbed up to 300. Participation grew much slower, but began to steadily increase in 2001 from less than 50 to over 350 observers. The unique species to observer ratio (figure 5) steadily decreased. The final values for 2013 were less than one species per participant, suggesting observer saturation. This also suggests that the number of avian species to be found in the county has been nearly exhausted.

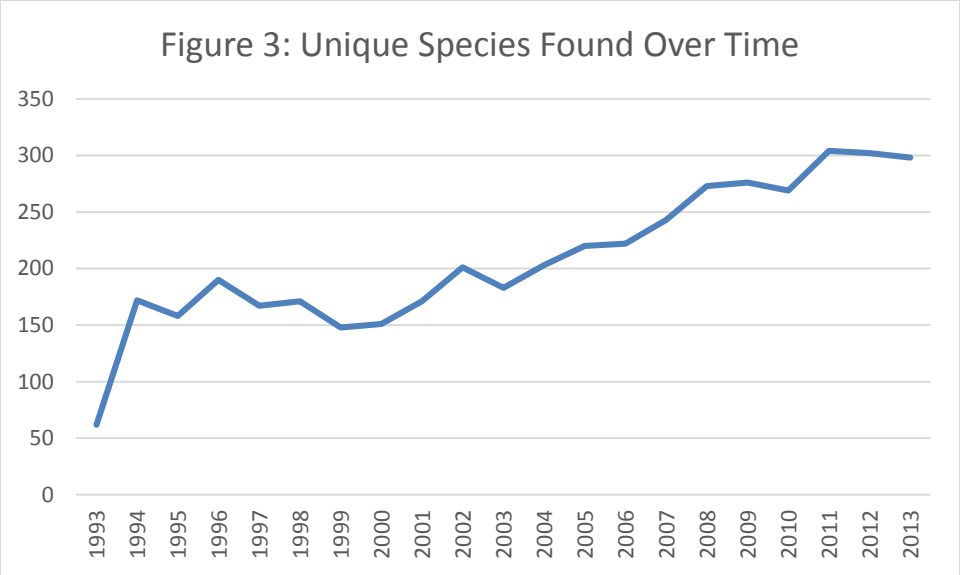


Figure 3: Unique Species Found Over Time
350
300
250
200
150
100
50
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

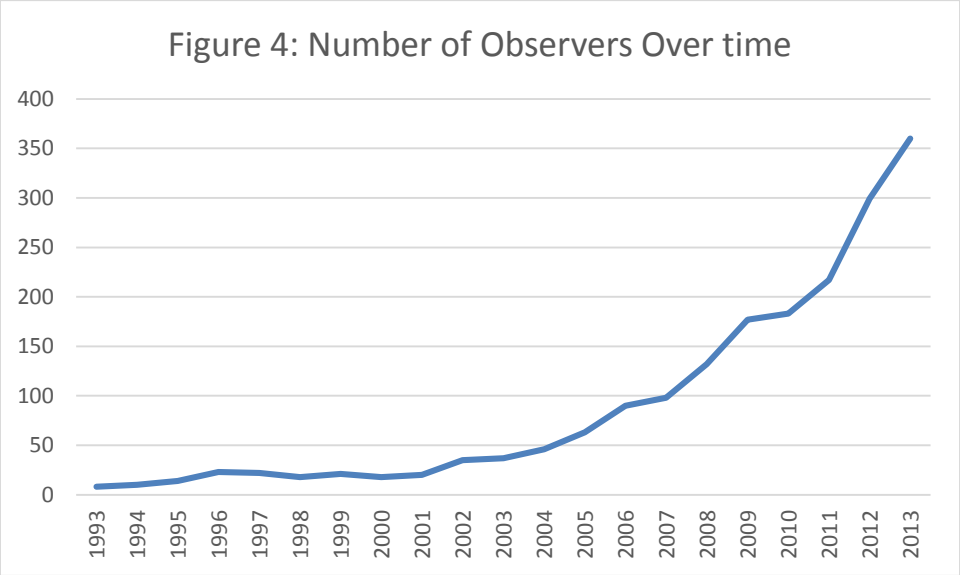
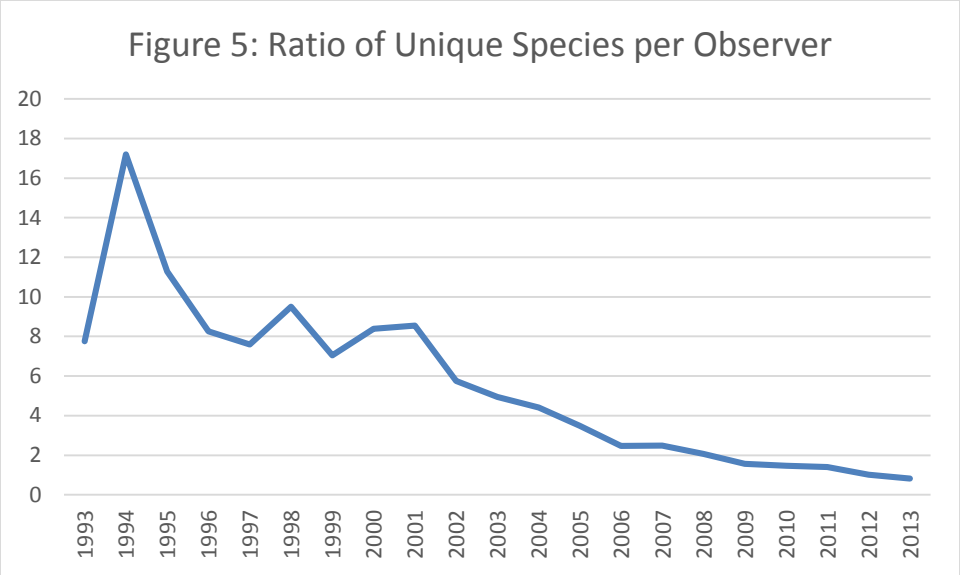


Figure 4: Number of Observers Over time
400
350
300
250
200
150
100
50
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013



Discussion:

The decreasing number of unique species found per observer suggests that the survey has been highly comprehensive and therefore gives us a good look at biodiversity in Suffolk County. Using this data we can now begin to analyze trends in the biodiversity of Suffolk County. Due to the geographic and temporal scales the larger data set encompasses, analysis of changing biodiversity on a continental scale is possible, and will be the next step for this research.

-Potential political advantages of Citizen Science

"If we can teach people about wildlife, they will be touched. Share my wildlife with me. Because humans want to save things that they love." – Steve Irwin

Conservation projects have two inherent problems that create difficulty with the general populace: they carry an economic opportunity costs and may lack support due to apathy. Opposition amongst policymakers from the first is likely made worse by the later. If voters do not care about conservation, policymakers most likely will not either. Citizen science projects enable science to step out of the ivory tower and onto the ground level. Incorporating the public in urban biodiversity research creates opportunities for people to develop an appreciation for it (Dinetti, 2010). This

could increase the likelihood that citizens would support conservation and reconciliation efforts. Urban scientists could potentially work with grassroots organizations to their mutual benefit.

-Issues with data mining and the current state of taxonomy

One of the core problems that data mining in ecology must deal with is the current state of taxonomy. The rise of informatics, open data, and citizen science could be particularly valuable to taxonomists. The current taxonomy system is regrettably not optimized for intensive data mining (Deans, Yoder, & Balhoff, 2012). Phenome annotations (which are used by taxonomists to describe species), tend to be composed in natural language and thus are difficult to data mine effectively due to a lack of standardization (Deans et al., 2012). They also do not usually reference explicit, logical definitions of concepts (Deans et al., 2012). While natural language is fine for humans, text data mining programs struggle with it. I had previously had difficulty with this problem while working on my undergraduate thesis attempting to correct for duplication in a marine mammal data set. This creates a limitation for urban ecologists seeking to integrate taxonomic data mining with large open data sets like eBird. Therefore, the questions urban ecologists could ask are limited, or at least more computationally complicated to answer. For example: if an urban ecologist was interested in asking questions regarding whether or not a species is more likely to take hold if it utilizes a specific color palate, currently it would require the ecologist to go through each species that has been spotted and identify their color palates. A more standardized taxonomic database could enable an urban ecologist to engage this question easier. Please see the figure below for an example, highlighting a proposed methodology made by (Deans et al., 2012).

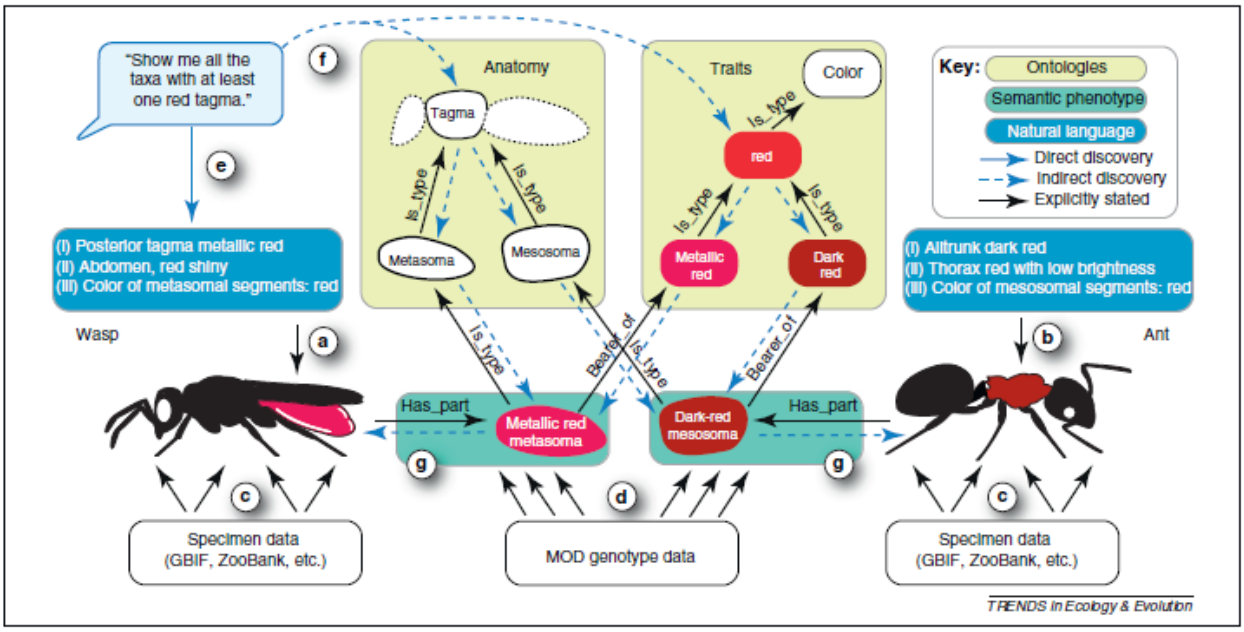
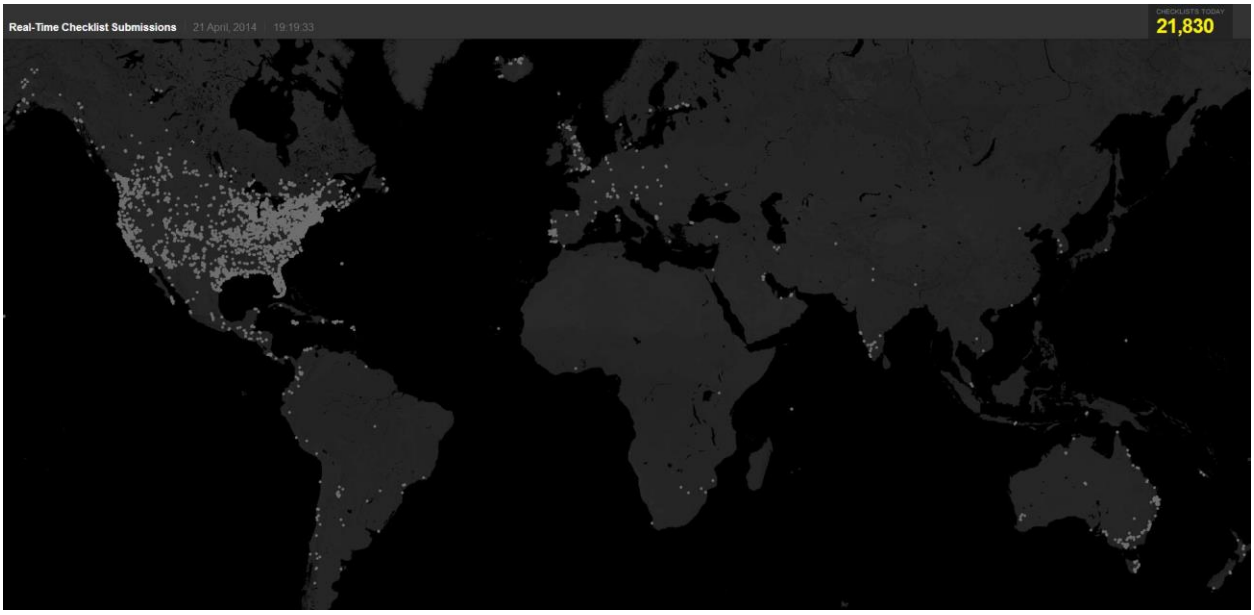


Figure 6: Benefits of semantic phenotypes. Taken from (Deans et al., 2012). The wasp and the ant are described in natural language (a,b) and connected to data in existing databases (c); this is the extent of current taxonomic practice. A biologist (upper left) queries against the corpus of biodiversity knowledge and fails to discover the ant when manually reading natural language (e), but finds the wasp and the ant when using the logic inherent in ontologies to query across semantic phenotypes (f). Model organism curators already collect genotype data and apply them to phenotypes using semantic methods (d). In the proposed method from Dean et al, taxonomists or other annotators produce phenome annotations (g) built following logical rules and referencing anatomy and trait ontologies (tan boxes). Their example here is simplified to illustrate the major connections. The graph representing an actual semantic phenotype within an explicit logical model would be more complex: briefly, an individual specimen might be asserted to be an instance of an OWL class expression such as 'has part some (metasoma and bearer of some metallic red)'. Abbreviations: GBIF, Global Biodiversity Information Facility; MOD, model organism database; OWL, Web ontology language (Deans et al., 2012).

Conclusion:

The field of urban ecology stands to benefit greatly from the rise of citizen science, open data, and advances in computer science such as ecoinformatics and cloud computing. These tools reduce the operating costs of large-scale environmental studies, and provide the tools to analyze massive data sets. Advances in technology have enabled citizen science to go from a linear structure in which the fate of data analysis lies purely on the shoulders of the data owner, to being available to be queried by more individuals. The greater opportunity analysis will no doubt increase the knowledge generated data points. At the same time efforts must be made to ensure the integrity of data submitted, and that data points be designed for ease of analysis.

Figure 7: Checklist submissions to eBird on the day of April 21st 2014. Retrieved from eBird website.



Works Cited

Barthlott, W., Biedinger, N., Braun, G., Feig, F., Kier, G. & Mutke, J. (1999). Terminological and methodological aspects of the mapping and analysis of global biodiversity. *Acta Botanica Fennica*, *162*, 103–110.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, *59*(11), 977–984. doi:10.1525/bio.2009.59.11.9

Chadwick, R. A. F. and M. A. (2013). *Urban Ecosystems: Understanding the Human Environment*.

Cornell Lab of Ornithology. (2014). eBird Basic Dataset.

Deans, A. R., Yoder, M. J., & Balhoff, J. P. (2012). Time to change how we describe biodiversity. *Trends in Ecology & Evolution*, *27*(2), 78–84. doi:10.1016/j.tree.2011.11.007

Dinetti, M. (2010). Attracting Interest in Urban Biodiversity with Bird Studies in Italy. *Urban Biodiversity and Design*, 453–462.

Francis, R., & Chadwick, M. (2013). *Urban Ecosystems: Understanding the Human Environment*. New York: Routledge.

Hochachka, W. M., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., & Kelling, S. (2007). Data-Mining Discovery of Pattern and Process in Ecological Systems. *Journal of Wildlife Management*, *71*(7), 2427. doi:10.2193/2006-503

Hochachka, W. M., Fink, D., Hutchinson, R. a, Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, *27*(2), 130–7. doi:10.1016/j.tree.2011.11.006

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Damoulas, T., & Gomes, C. (n.d.). eBird : A Human / Computer Learning Network for Biodiversity Conservation and Research, 2229–2236.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85–93. doi:10.1016/j.tree.2011.11.016

Munson, M. A., Webb, K., Sheldon, D., Fink, D., Hochachka, W. M., Iliff, M., … Kelling, S. (2012). The eBird Reference Dataset , Version 4 . 0, 1–11.

Norbert, M., & Werner, P. (2010). *Urban Biodiversity and the Case for Implementing the Convention on Biological Diversity in Towns and Cities*.

Solon, O. (2012). Pollution-detecting kites to monitor Beijing ' s air quality (Wired UK) Pollution-detecting kites to monitor Beijing ' s air quality (Wired UK). *Wired UK*, *4*.

Species, B., Caruana, R., Elhawary, M., Munson, A., Riedewald, M., Fink, D., & Hochachka, W. M. (2006). *Mining Citizen Science Data to Predict Prevalence of Wild*. Retrieved from file:///C:/Users/cround/Desktop/Spring 2014 Documents/Urban Ecology/Research Paper/Methodology.pdf

Wood, C., Sullivan, B., Iliff, M., Fink, D., & Kelling, S. (2011). eBird: engaging birders in science and conservation. *PLoS Biology*, *9*(12), e1001220. doi:10.1371/journal.pbio.1001220

Zooniverse Development Team. (2014). Galaxy Zoo. Retrieved January 04, 2014, from http://www.galaxyzoo.org/#/story