IJDC | *Peer-Reviewed Paper*

eBird: Curating Citizen Science Data for Use by Diverse Communities

Carl Lagoze
School of Information,
University of Michigan

Abstract

In this paper we describe eBird, a highly successful citizen science project. With over 150,000 participants worldwide and an accumulation of over 140,000,000 bird observations globally in the last decade, eBird has evolved into a major tool for scientific investigations in diverse fields such as ornithology, computer science, statistics, ecology and climate change. eBird's impact in scientific research is grounded in careful data curation practices that pay attention to all stages of the data lifecycle, and attend to the needs of stakeholders engaged in that data lifecycle. We describe the important aspects of eBird, paying particular attention to the mechanisms to improve data quality; describe the data products that are available to the global community; investigate some aspects of the downloading community; and demonstrate significant results that derive from the use of openly-available eBird data.

Received 22 October 2013 | *Accepted* 26 February 2014

Correspondence should be addressed to Carl Lagoze, 4444 North Quad, 105 S. State Street, Ann Arbor, MI 48104. Email: clagoze@umich.edu

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v9i1.302
DOI: 10.2218/ijdc.v9i1.302

Introduction

The explosion of the Internet as an everywhere-accessible technology has embedded crowdsourcing into all aspects of our lives. On a daily basis, we use information collected by and contributed by masses of volunteer participants when we seek facts about the world around us (e.g., Wikipedia), search for entertainment options (e.g., recommendations on Amazon or Netflix), and try to figure out the best way to get to work[1]. Even if the "wisdom of crowds"(Surowiecki, 2004) is not always wise, certainly the *influence* of crowds has reached new levels.

One particularly noteworthy application of crowdsourcing, and the subject of attention in this paper, is citizen science, which engages numerous volunteers as participants in large-scale scientific endeavours. These volunteer participants may play the role of either a processor or a sensor. A well known example of the processor role is the Zooniverse family of projects (Savage, 2012), in which volunteers classify or extract information from images. In this paper, we examine the eBird[2] project (Sullivan et al., 2009), an exemplar of the sensor model in which human volunteers independently collect data from the field and submit it through intuitive user interfaces on mobile devices or desktop computers. This submitted data is then made available in a variety of forms for a variety of use purposes.

From the perspective of data curation, in which ensuring data quality is a fundamental part of the data lifecycle, the distinction between the processor and sensor role is important. Processing tasks are usually repeatable; a given task can be undertaken a number of times by different volunteers, thus providing a cross check on the accuracy of the data. Furthermore, if necessary, the task can be repeated by an expert to provide a gold standard for quality. In contrast, the majority of sensor tasks cannot be repeated; there is no ground truth against which a volunteer observation can be validated. Because of this, the effect on data quality of widely varying skill levels among volunteer observers, and the inability to validate their submissions, is a long-standing point of concern and contention in the scientific community (Sauer, Peterjohn, & Link, 1994).

Despite these data quality concerns, for a variety of ecological phenomena, large-scale human sensing is the only viable means to collect sufficient quantities of data for analysis. Humans are extraordinarily capable of making observations of events in their surroundings and providing detailed descriptions of these events in ways that mechanical sensors cannot. The complexity of bird observation, which can involve a mixture of subtle visual features, nuanced audible signals in noisy environments, and contextual knowledge about habitat and environmental conditions, demands this unique capability of human intelligence. Furthermore, the global scale of ecological sensing and the variety of habitats and locations for which data are needed prohibits the exclusive use of highly trained and paid "professional" observers or scientists.

In this paper we examine the factors that contribute to eBird's success. One facet of its success is the number of volunteer participants and the quantity of observations submitted at the producer end of the data lifecycle. But an equally important measure of success is the utility and quality of data products that are made available at the consuming end of that lifecycle. As we describe in the remainder of this paper, this

[1] See: http://mashable.com/2013/08/20/google-maps-adds-waze-data/

[2] eBird: http://ebird.org

utility and quality is an outcome of careful eBird data curation practices that span the scope of the entire data lifecycle and are attentive to the requirements of the multiple stakeholders in the eBird community. This data lifecycle is illustrated in Figure 1. A particularly interesting aspect of the consuming end of the lifecycle is that eBird collects metadata about the consumers of its data products and their intended use of these products, currently an uncommon practice among citizen science projects. This metadata is the basis of the analyses we report later in this paper.

We begin with a description of eBird and the factors that contribute to both its success and the quality of the data that it collects. We then describe the data products that are made available to the global scientific communities, both professional and avocational. We follow that with an examination of the nature of the data use community based on the metadata collected in the download logs. Finally, we describe some of the notable usage domains of eBird data. We close with future directions of the project.

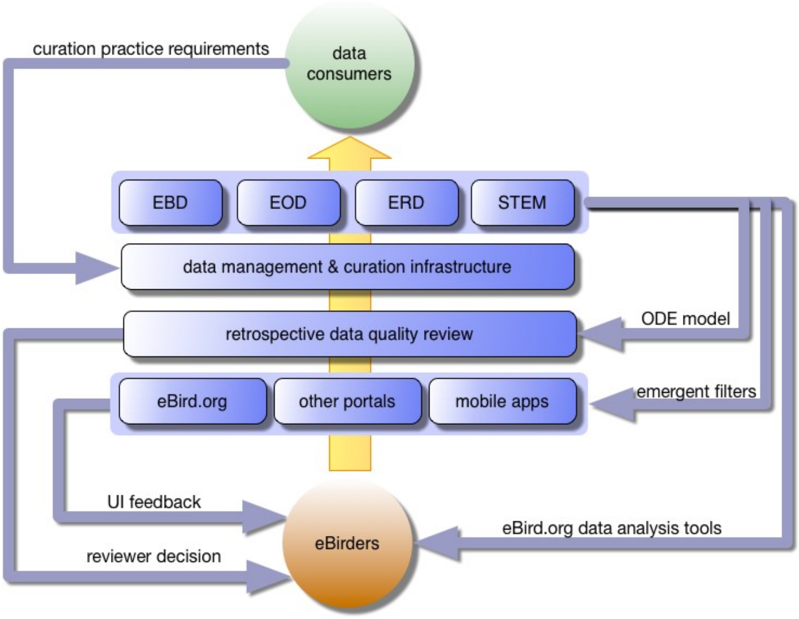
curation practice requirements
data consumers
EBD
EOD
ERD
STEM
data management & curation infrastructure
ODE model
retrospective data quality review
eBird.org
other portals
mobile apps
emergent filters
UI feedback
eBirders
eBird.org data analysis tools
reviewer decision

Figure 1. eBird data flow. Primary flow (yellow arrow) is up from observers to consumers. Note the feedback loops (blue arrows) at various levels. The observer and consumer communities overlap.

eBird: Features and Practices

eBird is a successful citizen science project for a variety of reasons. First among them is the manner in which it easily facilitates the tasks that birders most care about: entering, storing and accumulating their field observations. This practice makes use of the notion of checklists that are aggregations of species observations, a metaphor that predates eBird. In addition, it adds to this core historically-based functionality a number of new and innovative benefits enabled by its crowdsourced foundations. These include providing the tools with which birders can compare their birding accomplishments to those of fellow birders, thus appealing to the benevolent competitiveness of the birding community, and providing a variety of data exploration facilities that allow birders to explore the presence of species locally and throughout the world. Many of these features are based upon and facilitate long-standing community practices, substantially motivating adoption and participation. These tools make use of the data submitted by eBirders and help members of the eBird community to improve and advance their birding skills. These features, appealing to competitiveness and providing for self-improvement, are ideal examples of how a citizen science project can attend to self-interest while achieving the larger goal of contributing to scientific knowledge.

The growth rate of eBird contributions is illustrated in Figure 2. By the end of 2013, over 150 million observations will have been submitted by 150,000 unique observers, who spent 10.5 million hours in the field collecting data. This has generated an extraordinary biodiversity dataset that includes data from all countries in the world, representing more than 95% of known bird species.

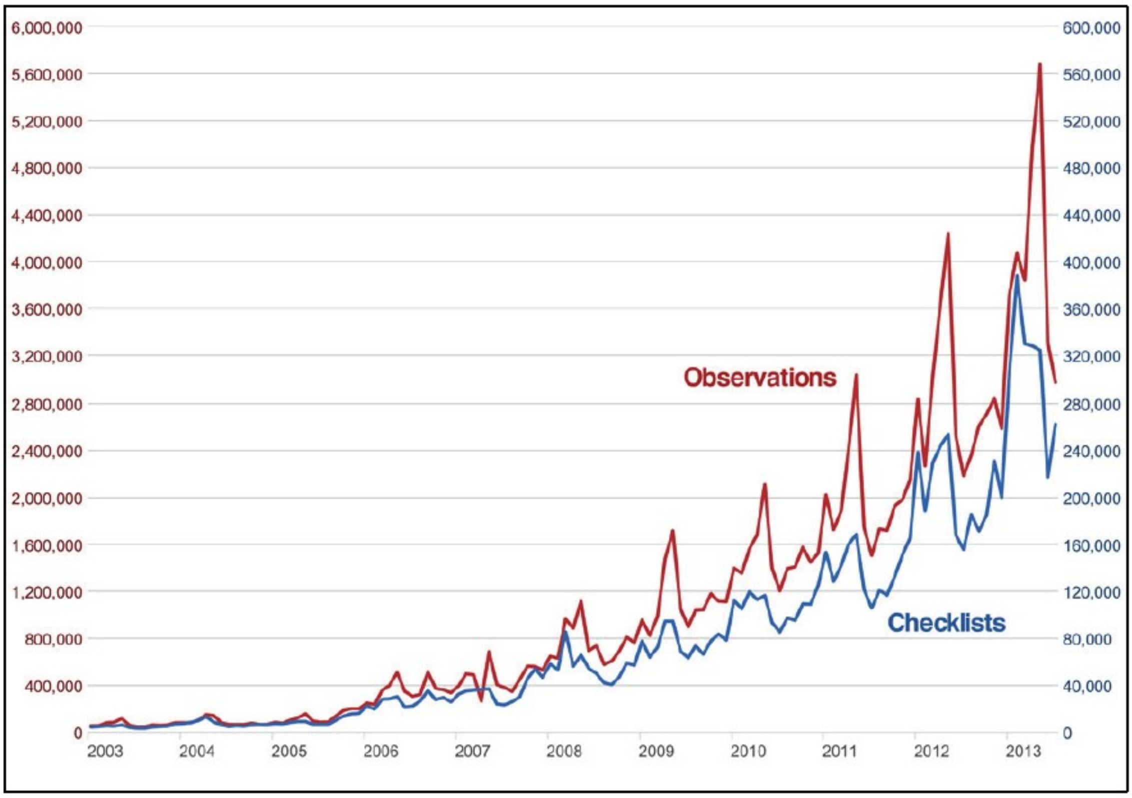

Figure 2. Growth in eBird contributions since its introduction in 2002. Note the seasonal fluctuations with most contributions occurring in May of each year, corresponding to cycles of high species activity and correspondingly high birder activity.

The primary interface to eBird since its introduction has been the website: eBird.org. Leveraging the rapid proliferation of mobile devices, this interface has been supplemented by a smartphone app through which, at a recent count, 20% of entries are now submitted. Both interfaces allow a volunteer participant to submit a set of observations as a checklist and provide basic metadata about them. Checklist metadata includes the time and day of the set of observations on the checklist, the location of the observations (specified at a variety of granularities from latitude-longitude to city and state name), whether the observations were recorded from a stationary position or moving (in which case distance covered is recorded), and how many observers were present. One additional and notable piece of metadata for the checklist is the “all species reported” field, by which the observer reports whether they have entered a complete list of all species identified at the site. This information is scientifically relevant because it provides both species presence and allows for the inference of species absence. Metadata for each species observation on the checklist includes a number of individuals observed and optional evidence, such as photographs.

Other than the requirement that participants acquire a username and password, and the minimal effort to enter observations, there are few barriers to participation[3]. For example, there are no tests for expertise that interested individuals must complete before participation. While such tests might contribute to some overall data quality, they might ultimately decrease participation. In the end, this would have a negative impact on the scientific utility of eBird; as shown by Hochachka et al. (2012), for citizen science projects patterns and signals are more effectively detected when data quantity is high, rather than from smaller amounts of high quality data.

Addressing Data Quality and Scientific Utility

An inevitable consequence of reduced barriers to participation is high variability in the expertise of observers and the quality of contributions. This variability, as noted earlier, raises a number of questions about the effectiveness of the data as usable scientific evidence. In response to this problem, eBird has developed and is developing three main quality control strategies to improve the integrity of the data that it collects, curates, and makes available to the community for research, without compromising the low-barriers-to-entry principle.

The first strategy is prospective: that is, implemented at the point of data entry. This proactive approach employs data-driven user-interface biases that favor the submission of plausible data. In effect, the user interface makes it easy for the observer to record species that are plausible at the spatiotemporal coordinates of the observation, and in quantities that conform to historical precedent. The automated filters that enable this selective presentation to emerge from the huge amount of validated historical eBird data (thus their characterization as emergent filters (Kelling, Yu, Gerbracht, & Wong, 2011)).

With a little extra effort, the interface does allow the observer to record a sighting of a highly unusual species or unusually large number of individuals, as such outliers are sometimes genuine and of particular interest to both birders and data users. In this case, the participant is prompted to confirm that the entry was not accidental and to provide additional details. A retrospective data quality approach is then employed, whereby the observation is automatically flagged for review and routed to a regional expert who evaluates the plausibility of the observation, and may consult additional data sources and elicit further evidence to make a judgment as to whether the data should be included in the research data set. The reviewer may reject the observation if there is lack of sufficient plausibility or evidence, although the user still retains this record to support personal interests and uses.

While the emergent filters effectively constrain the number of unusual observations, the volume of flagged records (4% of all the bird observations) and the reliance on human experts for validation present scalability problems, especially at the exponential growth rate eBird is currently experiencing. This has motivated research on strategies

[3] Of course, internet access and competency with a web browser is assumed; a barrier to participation for some.

that would reduce this reliance on the human review process by automatically identifying observer variability in the ability to detect species. To better understand observer variability in eBird we have applied a probabilistic machine-learning approach called the Occupancy Detection Experience (ODE) model to provide an objective measure of experience for all eBird observers (Yu, Wong, & Hutchinson, 2010). We can use the ODE model to distinguish the difference between expert observers, who typically find more birds and are more likely to detect both species and counts that fall outside of the emergent filter limits, as compared to novice birders, who are more likely to misidentify common species. In this manner, we will better automate a significant proportion of observation verification and only invest the limited supply of human intelligence where it is most needed.

eBird Data Products

The combination of a high volume of participation and attention to principles of data quality results in a number of high quality data products that are accessible through the eBird website and at other locations, to anyone interested worldwide. Aspects of this data usage community are described in the next section. This section is dedicated to a description of the products themselves. In addition to the data products described here, it should be noted that the primary consumption of eBird data takes place via the data exploration, visualization and analysis tools that are available on the eBird website and that facilitate access to eBird data. These tools are accessed by more than one million unique visitors annually.

We note that a critical component of curation and use is a set of policies that ensure proper attribution and acknowledgment for data providers. The data access policy for eBird, instituted in November 2012, attends to this through a series of steps. First, anyone downloading an eBird data product must first register with eBird. Downloaders must provide a set of basic metadata about their intended use, including their name, country of residence, project type, title and abstract project, as well as their affiliation. Finally, downloaders must agree to the eBird Terms of Use policy, which prohibits commercial use of the data and requires them to properly attribute the Cornell Lab of Ornithology in whatever results from the use of the data (e.g., academic papers).

Data Available for Download from Clearinghouses (DataONE and GBIF)

The eBird Observational Dataset (EOD) contains primary species-occurrence data defined as a record of a particular taxon in a particular place at a particular point in time (Soberón & Peterson, 2009). This data format is optimized for integration with other observational data and natural history collections data for estimating patterns in biodiversity. The 2012 version contains more than 100 million bird observation records. The use of the EOD in ecological services is widespread (Davis, Malas, & Minor, 2013; Lait, Friesen, Gaston, & Burg, 2012). Data usage reports of the EOD are not provided by the data clearinghouses from which it is available.

Data Available for Download Directly from eBird According to Data Access Policy

The eBird Basic Dataset (EBD) contains checklist data. Checklists are defined as counts of all bird species observed during a single search event (Sullivan et al., 2009). Each record corresponds to a species observation, containing the metadata about the observation as described earlier (e.g., species, number of individuals). An observation record also contains the key of adjacent records (observations) that should be grouped within a single checklist. The checklist metadata (e.g., time, date, location, etc.) is duplicated across the set of observations in the checklist.

The eBird Reference Dataset (ERD) is a value-added data product that combines the EBD with two additional types of information necessary for more detailed distributional analysis. The first is species absence information, which can be inferred from data supplied by checklists where "all species reported" was indicated. Combined with participant-recorded information on search effort, these apparent absences add valuable information that is used to capture and control for sources of variation associated with the detection process (Fink et al., 2010; Fink, Damoulas, & Dave, 2013). The second piece of added information is a large suite of variables that describe the local environment where searches took place. These variables include descriptions of land cover and elevation, climate, human population density and so on.

Download metadata is collected for EBD and ERD.

Data Available with Restrictions

The Spatial-Temporal Exploratory Models (STEM) data set is a comprehensive, model-based data product derived from the ERD via statistical models that produce high-resolution, weekly distribution estimates across the continental United States for several hundred species using STEM and AdaSTEM modeling processes (Fink et al., 2010; 2013). This model analysis effectively "fills in" the data in areas where observations are sparse or have not been made, and adjusts observations to account for variation in observer effort statistically demonstrated to affect probability of detection. The STEM data set is only available on special request and typically requires that eBird personnel be listed as co-authors on the paper making use of this data set.

Characteristics of eBird Data Use Community

Anyone choosing to download the publicly available EOD and ERD data products must register with eBird and supply basic metadata about themselves and the projects in which they wish to use the data. In this section, we provide descriptive analysis of that metadata and what it reveals about the data use community. While these data are currently sparse, in future research we hope to enhance these data via interviews or surveys that will reveal more about the eBird data user community.

Since the download registration process was instituted in November 2012, there have been over 1,100 downloads of eBird data products. The large majority of these are by unique users. The distribution of the countries of origin of these downloaders is illustrated in Figure 3. As would be expected for a data set that currently contains predominantly North American data, the majority of downloads come from the United States and other North American countries (e.g., Canada and Mexico). Notably, this distribution of countries roughly corresponds to the national distribution of eBird contributions.

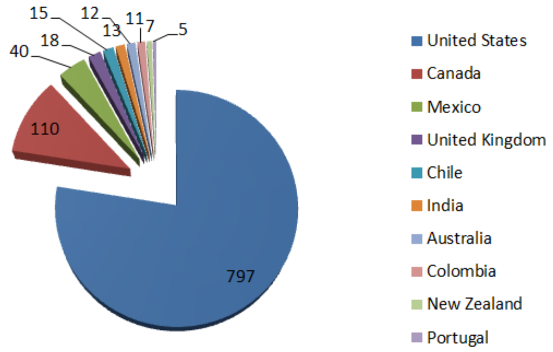

Figure 3. The top ten countries of origin for downloaders of eBird data set.

The word cloud in Figure 4, which shows term frequency in project title and description fields, gives a snapshot of the topic distribution of projects that are consuming eBird data. Topics such as migration, conservation, modeling, mapping, climate, population, conservation, habitat, and a number of others stand out as primary areas of interest.

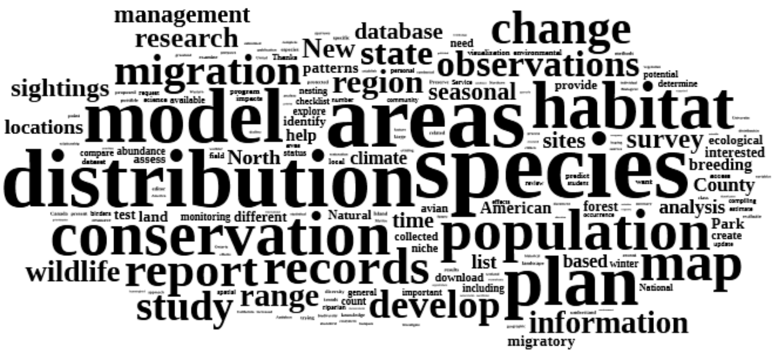

Figure 4. Word cloud of downloading project titles and descriptions.[4]

The word cloud in Figure 5 illustrates the frequency of terms present in the organization names for users registered to download eBird data. As shown, there is a fairly prominent and equal distribution of academic (terms such as college, university, department, Universidad), government (terms such as national, Canada, California), private (terms such as society, association, museum), and personal use of the data.

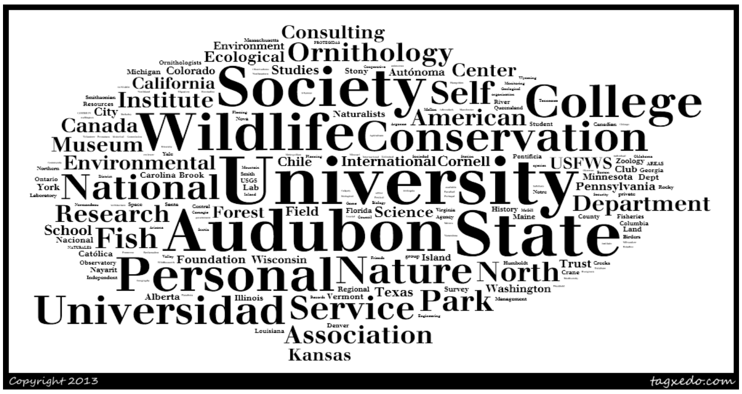

Figure 5. Word cloud of downloading organization names.

[4] This and other word clouds in this paper were generated at http://tagxedo.com

This distribution of organization name terms is reflected in Figure 6, which illustrates the distribution of self-described occupation types of eBird product downloaders. While academic use predominates, the pie is fairly evenly distributed amongst academic, governmental, and general (which can be inferred to be private organizations or individuals). Downloading by commercial organizations is obviously quite small, constrained by the prohibition for commercial use of the data.

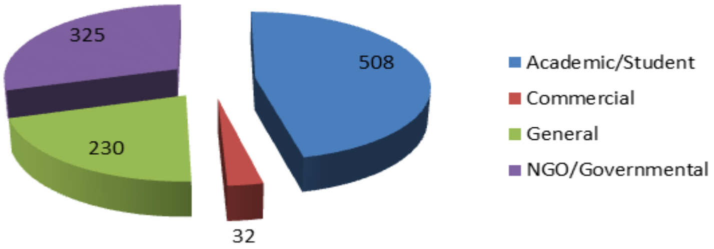

Figure 6. Occupations of downloaders.

eBird Data Use Examples

The ultimate test of the effectiveness of the eBird curation strategy lies in the quality and quantity of the results from the projects that download eBird data products. The remainder of this section describes exemplars of these projects in the areas of research, policy and decision support, and education.

Research: Avian Migration Ecology

eBird data enables investigation of avian ecology questions in unprecedented spatial and temporal detail (Hochachka et al., 2012). This is particularly true for studies of migration patterns, where typically only the movements of individual birds have been studied using tracking devices or banding (Bairlein, 2003). eBird makes it possible to extend these studies beyond the individual into the dynamics of the population as a whole. This has allowed the testing of key predictions originating from optimal migration theory (La Sorte, Fink, Hochachka, DeLong, & Kelling, 2013) and quantification of the structure, seasonal dynamics and determinants of migration flyways. These studies are increasingly important to understand the impact of global climate change on migratory birds (Carey, 2009) and the decline of many species of long-distance migratory birds (Sanderson, Donald, Pain, Burfield, & Van Bommel, 2006).

Policy: Conservation

A number of examples demonstrate the outstanding utility of eBird data for conservation policy and research. The partnership between eBird and the North American Bird Conservation Initiative has produced a series of comprehensive analyses of the state of the nation's birds. These analyses produce sobering indicators of ecological health and biodiversity status (Butchart et al., 2010; Pereira & David Cooper, 2006). But, on the optimistic side, they give evidence that birds can respond quickly and positively to conservation action. This optimistic outcome led to an effort that integrated the eBird STEM data product with the US Protected Areas Database to estimate weekly stewardship responsibilities for 370 species on private and public lands within the contiguous United States. By combining these data, researchers identified species moving seasonally between public lands under management by different government units, providing a compelling rationale for inter-agency cooperation to develop full lifecycle conservation plans for birds. Finally, a partnership with The Nature Conservancy of California is identifying and prioritizing critical habitats for migratory birds and designing evidence-based strategies for their conservation. This is been particularly effective in California's Central Valley, which is one of the most altered landscapes in the world, but still contains existing and vital refuges that support millions of migrating and wintering waterfowl and shorebirds (Central Valley Joint Venture, 2006). STEM species distribution models are used to identify time-dependent relationships between bird occurrence and land tenure, management, agricultural crop types and water availability. The goal is to identify high-value areas for birds and then work with landowners and managers to better manage these key sites for waterbirds.

Education: Big Data for K-12

All aspects of eBird – the data entry tools, data analysis tools, and the data products – provide an outstanding context for students, whether they are in grade school or in an undergraduate program, to pose scientific questions, design investigations, analyse real-world data from their own home town, and interpret the results. A number of curriculum resources have been created to support teachers who wish to build science units based on eBird (Schaus, Bonney, Rosenberg, & Phillips, 2007). Before or after observing local birds, students can use eBird data to determine which species are common in their community, discover trends and develop hypotheses. Students in schools that make use of eBird over a number of years can research multi-year trends (Fee, Curley, & Trautmann, 2013). The openly accessible eBird data gives students a chance to experience what it is like to work with "big data" and understand ecological science beyond stereotypical images of individual scientists working in the lab or field.

Conclusions and eBird Futures

In this paper we described eBird, a citizen science project that is not only notable for its popularity among volunteers and hobbyists, but also for its utility for high quality and influential science, decision support and educational value. Many facets of eBird's success derive from the careful attention to data curation that extends over the entire eBird data lifecycle, from the manner in which data are collected, the strategies by which the quality of the data is ensured, the storage and management of those data, and the broad availability of both the data and value-added data products. eBird's successes provide an aspirational model of best practices for a variety of citizen science efforts.

The increasing and disturbing effects of climate change will inevitably increase the importance and popularity of eBird and related citizen science efforts, which can support public understanding of and participation in addressing global-scale concerns. Our work on improving and automating data quality factors will continue to legitimize citizen science as a valid foundation for scientific knowledge production, while further supporting evidence-based policy and education initiatives for our changing world.

Acknowledgements

We thank the thousands of eBird participants around the world who have contributed to the success of eBird. Portions of this work for funded by the National Science Foundation under CISE-SOCS:66281.

References

Bairlein, F. (2003). The study of bird migrations: Some future perspectives. *Bird Study, 50,* 243–253. doi:10.1080/00063650309461317

Butchart, S., Walpole, M., Collen, B., Van Strien, A., Scharlemann, J. P. W., & Almond, R.E.A. (2010). Global biodiversity: Indicators of recent declines. *Science, 328,* 1164. doi:10.1126/science.1187512

Carey, C. (2009). The impacts of climate change on the annual cycles of birds. *Philosophical Transactions Of The Royal Society Biological Sciences, 364,* 3321–3330. doi:10.1098/rstb.2009.0182

Central Valley Joint Venture. (2006). Central Valley Joint Venture implementation plan: Conserving bird habitat. Sacramento.

Davis, A.Y., Malas, N., & Minor, E.S. (2013). Sustainable habitats? The biophysical and anthropogenic drivers of an exotic birds distribution. *Biological Invasions, 1*(13). doi:10.1007/s10530-013-0530-z

Fee, J., Curley, L., & Trautmann, N.M. (2013). Connecting with students through birds. In *Citizen Science: 25 Lessons between Biology to Life* (pp. 23–32). Arlington, VA: NSTA Press.

Fink, D., Damoulas, T., & Dave, J. (2013). Adaptive spatial-temporal exploratory models: Hemisphere-wide species distributions from massively crowd sourced eBird data. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI-13)*. Palo Alto, CA: AAAI Press.

Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., … Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications, 20*(8), 2131–2147. doi:10.1890/09-1340.1

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 and Evolution, 27,* 130–137. doi:10.1016/j.tree.2011.11.006

Kelling, S., Yu, J., Gerbracht, J., & Wong, W.-K. (2011). Emergent filters: Automated data verification in a large-scale citizen science project. In *2011 IEEE Seventh International Conference on eScience Workshops,* (pp. 20–27). doi:10.1109/eScienceW.2011.13

La Sorte, F.A., Fink, D., Hochachka, W.M., DeLong, J.P., & Kelling, S. (2013). Population-the scaling of avian migration speed with body size and migration distance for powered fliers. *Ecology, 94,* 1839–1847. doi:10.1890/12-1768.1

Lait, L.A., Friesen, V.L., Gaston, A.J., & Burg, T.M. (2012). The post-Pleistocene population genetic structure of a western North American passerine: The chestnut-backed chickadee Poecile rufescens. *Journal of Avian Biology, 43*(6), 541-552. doi:10.1111/j.1600-048X.2012.05761.x

Pereira, H.M., & Cooper, D.H. (2006). Towards the global monitoring of biodiversity change. *Trends in Ecology & Evolution, 21,* 123–9. doi:10.1016/j.tree.2005.10.015

Sanderson, F.J., Donald, P F., Pain, D.J., Burfield, I.J., & Van Bommel, F.P.J. (2006). Long-term population declines in Afro-Palearctic migrant birds. *Biological Conservation, 131,* 93–105. doi:10.1016/j.biocon.2006.02.008

Sauer, J.R., Peterjohn, B.G., & Link, W.A. (1994). Observer differences in the North American breeding bird survey. *The Auk, 111*(1), 50–62. doi:10.2307/4088504

Savage, N. (2012). Gaining wisdom from crowds. *Communications of the ACM, 55,* 13. doi:10.1145/2093548.2093553

Schaus, J. M., Bonney, R., Rosenberg, A. J., & Phillips, C. B. (2007). Investigating evidence [Web page]. Retrieved from Cornell University, BirdSleuth website: http://www.birdsleuth.org/investigating-evidence-2/

Soberón, J., & Peterson, A.T. (2009). Monitoring biodiversity loss with primary species-occurrence data: toward national-level indicators for the 2010 target of the convention on biological diversity. *Ambio, 38,* 29–34. doi:10.1579/0044-7447-38.1.29

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation, 142*(10), 2282–2292. doi:10.1016/j.biocon.2009.05.006

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday.

Yu, J., Wong, W.-K., & Hutchinson, R.A. (2010). Modelling experts and novices in citizen science data for species distribution modelling. In *2010 IEEE International Conference on Data Mining* (pp. 1157–1162). doi:10.1109/ICDM.2010.103