Conservation Biology

Editorial

Emerging problems of data quality in citizen science

The role of citizen science in research and natural resource monitoring and management is increasing, as evidenced by the growing number of peer-reviewed publications (including a special section in this journal) and calls for involving citizens in monitoring and governance (through, for example, "participatory research" [Danielsen et al. 2014] and "participatory monitoring" [Kennett et al. 2015]). Citizen science projects can be targeted to a specific research question (and thus involve very specific data-collection protocols) or can be more open-ended (giving rise to a need to collect data for which the uses may be unknown or changing) (Wiersma 2010). Advances in online content production and sharing technologies (i.e., Web 2.0), mobile computing, and sensor-equipped devices have contributed to a dramatic rise in online citizen science projects, in which citizens contribute sightings (e.g., eBird [Sullivan et al. 2009]), transcribe data (e.g., Old Weather [Eveleigh et al. 2013]), or classify phenomena (e.g., Galaxy Zoo [Hopkin 2007]). It is these online projects, also referred to as crowdsourcing (Franzoni & Sauermann 2014), which have been the focus of our research and that inform the opinions presented here.

Galaxy Zoo exemplifies an initiative that began as a targeted project in which citizens were engaged in the relatively simple task of classifying images of galaxies as one of 3 shapes (Hopkin 2007). The goal was to distribute a large workload among a large number of people. Citizen participation grew quickly, which led project sponsors to create an online forum to accommodate the large volume of comments and questions. Through this forum, a number of unanticipated categories of celestial bodies arose, including 2 from Dutch school teacher Hanny Van Arkel, who noted the "green peas" phenomena (Cardamone et al. 2009) and a new body that became known as "Hanny's Voorwerp" (Lintott et al. 2009).

The Galaxy Zoo story provides an example of the different dimensions of data quality in citizen science. The researchers anticipated a small, fixed set of categories of galaxy shapes and designed the data-collection interface accordingly. One dimension of data quality (Lewandowski & Specht 2015) is data accuracy; others include data completeness and timeliness. (For a complete discussion of the many dimensions of data quality, see Wang and Strong [1996]). In the case of Galaxy Zoo, data accuracy is measured as the proportion of images correctly classified by galaxy shape. Had it not been for the attentiveness of one person who went beyond the task of classifying galaxies into predetermined categories and was able to communicate this to the researchers via the online forum, what turned out to be important new phenomena might have gone undiscovered. Failure to discover these phenomena would have affected the data-quality dimension of completeness because not all celestial bodies in the images would have been cataloged. Thus, the data quality would be diminished.

Lewandowski and Specht (2015) describe 4 dimensions of data quality in their broad review of biology-themed citizen science: data accuracy and precision; sufficient sample size; and standardized sampling procedures (including sufficient spatial and temporal representation). These dimensions are congruent with good scientific practice and thus suggest that the criteria used to measure the quality of citizens' data should fit the standards of professional science. In this sense, citizen science amounts to asking citizens to fill in the blanks in a story written by scientists.

Although it is helpful for citizen scientists to adhere to standards of scientific practice, the process of doing science includes more than simply collecting and processing data. As Stevens et al. (2014:21) admitted: "Often . . . participants might be viewed as sensors or data collectors, but they're rarely invited to decide what data to collect or to contribute to the data analysis or interpretation, even though they . . . might have valuable insights," a view echoed in a recent *Nature* commentary by Kennett et al. (2015).

The online forum created by the Galaxy Zoo project manifests a design decision that allowed for participants to provide valuable new insights and contribute beyond simply classifying images. Because discoveries resulted from one individual going beyond the assigned task, an open question is how many discoveries went undetected because other participants failed to notice particular features (given the prescribed task) or noticed but failed to post on the site's forum. Based on examples such as this one, we argue that data quality in citizen science is much more than data accuracy.

Because citizens generally lack formal scientific training, they view problems and issues in light of their own knowledge and interests, creating fertile ground for discoveries. This perspective – that citizen scientists view problems differently than scientists – means that the quality of data should be defined as more than simply

Conservation Biology, Volume 30, No. 3, 447–449

DOI: 10.1111/cobi.12706

consistency with data collected under scientific protocols. Quality of data also includes the extent to which the design of a specific project facilitates citizens' abilities to spot something interesting, unexpected, or novel. Rather than seeing inexperience and lack of formal scientific training as threats to data quality, we suggest these characteristics improve data quality, provided that researchers are able to understand how to take full advantage of them. It is also important to consider that citizen scientists are not a homogenous group. He and Wiggins (2015) characterize citizen scientists as members of "communities." Such communities are thought to be at the opposite end of the spectrum of the larger "crowd" that is referred to in discussions on crowd sourcing. This community is a subset of the public with specific interests, whereas the crowd usually refers to a broader citizenry. These community members may have some training and expertise; thus, we consider them "expert amateurs" (Van Arkel is an example; she self-identifies as an amateur astronomer.) and thus not representative of the full suite of potential participants in online citizen science projects.

This broader view of data quality in citizen science is consistent with prevailing conceptualizations within the information-based fields of computer science, information systems, and philosophy of information (Wang & Strong 1996; Floridi 2012). Research in these areas treats data quality as a multidimensional construct (e.g., Wang and Strong [1996] identify hundreds of dimensions). Consequently, caution is warranted in emphasizing a particular dimension of data quality in citizen science projects; trade-offs in different dimensions of data quality are inevitable (Pipino et al. 2002; Scannapieco et al. 2005; Batini & Scannapieca 2006). Recent empirical evidence shows that data completeness (because most people are excluded) may be compromised to increase data accuracy (if the task can only be completed by a member of the community or an expert amateur). Analysis of participation patterns and data collected by citizens suggests that one reason that accuracy in the identification of objects does not differ between experts and citizen scientists (Crall et al. 2011; Jordan et al. 2012; Nagy et al. 2012) is often because citizen scientists in these projects already have expertise and a high level of interest in the topic. For example, to participate in eBird, one must already have facility with, or at least interest in, bird identification; many dedicated birders have as good (or better) field identification skills than professional ornithologists. Many citizen science projects, therefore, may actually inhibit widespread participation because of the requirement to provide data at a level that matches the expertise of the project sponsor, thus resulting in a trade-off in 2 dimensions of data quality (Parsons et al. 2011).

We contend that to truly engage a broad array of citizenry in science, projects should be designed to be as inclusive as possible, rather than limited to expert amateurs. Our proposed solution is to consider how project design influences quality. We have shown through laboratory and field experiments that data contributed through a flexible approach that allows nonexperts to provide descriptions of the observed organism (e.g., plant and animal) can have higher classification accuracy (Lukyanenko et al. 2014*a*) and greater numbers of observations reported (Lukyanenko et al. 2014*b*) than data contributed through traditional approaches to citizen science that require citizens to report observations using predetermined categories (e.g., species). We have explored alternative approaches to citizen science data collection in hopes of minimizing what seems to be an inevitable trade-off between data-quality dimensions. Specifically, we propose a flexible, instance-based approach to data collection that allows a contributor to classify data at the level at which they feel competent, rather than requiring participants to meet scientific standards that only expert amateurs may be capable of (Lukyanenko et al. 2014*a*).

We contend that in trying to hold amateurs to scientific standards, researchers not only ask nonexperts to perform often unrealistic tasks, but also risk missing the opportunity to fully engage with people in the core objective of discovery. The emerging problem of quality in citizen science is, therefore, writing a story in which citizens contribute to the plot.

Acknowledgments

Many of the ideas in this paper were formulated while R.L. was a PhD candidate and funded by an NSERC Canada Graduate Scholarship. We are grateful to our colleagues, especially R. Sieber and G. Wachinger, for insightful discussions and to A. Wals and one anonymous reviewer for comments that helped improve this manuscript.

Roman Lukyanenko,*† Jeffrey Parsons,‡ and Yolanda F. Wiersma§

*Department of Decision Science and Information Systems, Florida International University, Miami, FL 33199, U.S.A.
†Finance and Management Science Department, Edwards School of Business, University of Saskatchewan
‡Faculty of Business Administration, Memorial University, St. John's, NL A1B 3X5, Canada
§Department of Biology, Memorial University, St. John's, NL A1B 3X9, Canada, email ywiersma@mun.ca

Literature Cited

Batini C, Scannapieca M. 2006. Data quality: concepts, methodologies and techniques. Springer, New York.

Cardamone C, et al. 2009. Galaxy Zoo green peas: discovery of a class of compact extremely star-forming galaxies. Monthly Notices of the Royal Astronomical Society **399:**1191–1205.

Crall AW, Newman GJ, Stohlgren TJ, Holfelder KA, Graham J, Waller DM. 2011. Assessing citizen science data quality: an invasive species case study. Conservation Letters **4:**433–442.

Danielsen F, et al. 2014. Linking public participation in scientific research to the indicators and needs of international environmental agreements. Conservation Letters **7:**12–24.

Eveleigh A, Jennet C, Lynn S, Cox AL. 2013. "I want to be a captain! I want to be a captain!": gamification in the Old Weather citizen science project. ACM International Conference Proceeding Series. Proceedings of the first international conference on gameful design, research, and applications. Gamification **13:**79–82.

Floridi L. 2012. The road to the philosophy of information. Springer, New York.

Franzoni C, Sauermann H. 2014. Crowd science: the organization of scientific research in open collaborative projects. Research Policy **43:**1–20.

He Y, Wiggins A. 2015. Community-as-a-service; data validation in citizen science. University of Maryland Open Knowledge Lab, College Park. Available from http://openknowledge.umd.edu/wp-content/uploads/2015/11/Method2015.pdf (accessed January 2016).

Hopkin M. 2007. See new galaxies – without leaving your chair. Nature News Online 11 July. DOI: 10.1038/news070709-7.

Jordan RC, Ballard HL, Phillips TB. 2012. Key issues and new approaches for evaluating citizen-science learning outcomes. Frontiers in Ecology and the Environment **10:**307–309.

Kennett R, Danielsen F, Silvius KM. 2015. Conservation management: citizen science is not enough on its own. Nature **521:** 161.

Lewandowski E, Specht H. 2015. Influence of volunteer and project characteristics on data quality of biological surveys. Conservation Biology **29:**713–723.

Lintott CJ, et al. 2009. Galaxy Zoo: Hanny's Voorwerp, a quasar light echo? Monthly Notices of the Royal Astronomical Society **399:** 129–140.

Lukyanenko R, Parsons J, Wiersma YF. 2014*a*. The IQ of the crowd: understanding and improving information quality in structured user-generated content. Information Systems Research **25:**669–689.

Lukyanenko R, Parsons J, Wiersma YF. 2014*b*. The impact of conceptual modeling on dataset completeness: a field experiment. Proceedings of the International Conference on Information Systems 2014. Association for Information Systems, Atlanta. Available from http://aisel.aisnet.org/icis2014/proceedings/GeneralIS/29/ (accessed February 2016).

Nagy C, Bardwell K, Rockwell RF, Christie R, Weckel M. 2012. Validation of a citizen science-based model of site occupancy for eastern screech owls with systematic data in suburban New York and Connecticut. Northeastern Naturalist **19**(sp6)**:**143–158.

Parsons J, Lukyanenko R, Wiersma Y. 2011. Easier citizen science is better. Nature **471:**37.

Pipino LL, Lee YW, Wang RY. 2002. Data quality assessment. Communications of the ACM **45:**211–218.

Scannapieco M, Missier P, Batini C. 2005. Data quality at a glance. Datenbank-Spektrum **14:**6–14.

Stevens M, Vitos M, Altenbuchner J, Conquest G, Lewis J, Haklay M. 2014. Taking participatory science to extremes. Pervasive Computing, IEEE **13:**20–29.

Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Daniel F, Kelling S. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation **142:**2282–2292.

Wang RY, Strong DM. 1996. Beyond accuracy: what data quality means to data consumer. Journal of Management Information Systems **12:**5–33.

Wiersma YF. 2010. Birding 2.0: citizen science and effective monitoring in the Web 2.0 world. Avian Conservation and Ecology **5**(2)**:**13. [online] URL: http://www.ace-eco.org/vol5/iss2/art13/http://dx.doi.org/10.5751/ACE-00427-050213