Connecting Diverse Communities Through
Citizen Science Data Infrastructure

Jen Hammock

Encyclopedia of Life Secretariat

Smithsonian Institution

10th and Constitution Ave, NW

Washington, DC 20560 USA

hammockj@si.edu

Andrea Wiggins

College of Information Studies

University of Maryland

4121G Hornbake Bldg, South Wing

College Park, MD 20742 USA

wiggins@umd.edu

The BioCubes project is a citizen science biodiversity observation and reporting protocol designed by Smithsonian for group participation in classroom and non-formal education settings. A two-year study was recently funded to develop a proof-of-concept information technology infrastructure to support distribution and use of data from BioCubes across the US. The goal of this work is promoting novel STEM learning opportunities for diverse student populations while demonstrating the feasibility of exchanging and integrating citizen science data with ongoing biodiversity research, a key step in advancing citizen science (Bonney et al., 2014). In this abstract, we briefly introduce key features of our project and identify several considerations for the design of the sociotechnical systems to support both data contributors and data users.

BioCubes participants explore biodiversity through a census of organisms living within a cubic foot of space, mostly small insects. Participants select a cube location, document it, and then collect and sort the inhabitants, photographing them, identifying them if possible, and finally reporting them online. The one cubic foot model is useful for sampling because nearly every group of learners can access an urban garden, park, or corner of the schoolyard. BioCubes leverages the accessibility of nature in miniature to appeal to participants of all backgrounds, including those with limited bioliteracy and exposure to nature. We plan to distribute BioCubes kits to up to 300 educators, engaging up to 900 students from diverse populations.

Once BioCubes data are recorded, the technical infrastructure will create new opportunities for using and interacting around the data. These include basic data summary and exploration functionality, but the primary focus is on connecting disparate communities that have common interests in the data. The online platform where the observation records will be posted, iNaturalist, already has a vibrant community of naturalists with a collaborative culture, most of whom have substantial expertise in at least one taxon. If BioCubes records are sufficiently detailed and interesting, the iNaturalist community can be expected to interact with the BioCubes learners through online comments and assistance with identifying the BioCubes wildlife. In addition, iNaturalist members are important for data validation; every observation that achieves consensus on the identification from two members is rated as "research grade" data, and we anticipate this volunteer-based validation will be critical for data consumers.

In addition, we will develop a bidirectional data notification service. Potential biodiversity data users, including researchers and conservation professionals, will be able to make ad hoc queries or create standing alerts for new records as data becomes available. These subscribers will be sent records matching their geographic region and organisms of interest (e.g., earthworms in the Great Lakes region) with links to the online records, so they can post questions or comments about the data. The service will also notify data contributors when one of their records has been delivered to an interested data subscriber.

Ideally, learners will generate primary biodiversity data, receive guidance from experienced naturalists to help them identify the wildlife in their BioCubes, contribute to biodiversity science, receive notifications when researchers use their data, and potentially communicate directly with researchers. But for some learners, this exercise will fail at the first hurdle. The research plan therefore includes a social science component focused on identifying barriers to participation and production of reliable scientific data, as well as assessing the needs and concerns of data users, who may view citizen science data with some skepticism.

The design of tools and documentation will be critical to the success of the system and its users. We have identified several potential barriers to participation:

- *Technical*: registration and data entry on iNaturalist were a barrier in test groups; we will develop a COPPA-compliant classroom-level login management system that may help mitigate this issue.
- *Social*: producing observation records that are interesting enough to elicit helpful reactions from the iNaturalist community may be constrained by several limitations, such as lack of "interesting" wildlife in areas with poor biodiversity, the difficulty of high level organism identification, and producing useful, usable photographs given varying access to camera equipment and experience.
- *Feedback*: this infrastructure is designed to facilitate a gratifying feedback loop that informs data producers when their data are used by researchers, which depends in part on timely notifications and interactions. If new BioCube records prompt no initial response, will the experience be unsatisfying for the observers? If a naturalist or researcher responds to an observation record several days or months after the fact, will the observers still care or even notice the feedback?
- *Scale*: the feedback system could be either motivating or overwhelming. If one record is delivered to a data consumer and the observers are notified, they become aware of their contribution to science, which we expect will be a rewarding experience. Should an individual or group produce a larger number of records of interest to researchers, how can we prevent the reciprocal data usage notifications from becoming an annoyance? For most users, we expect only one or two notices will be delivered to data users, so we want to maximize their opportunity to learn that their data has reached a consumer. However, for a few highly productive data contributors, prominent notifications could amount to spam, which could become a disincentive.

In summary, our system has potential to promote productive social structures by connecting learners, naturalists, and researchers with tools and data that appeal to the interests of each group. The embedded social science research will focus on identifying how to promote data consumers' trust of citizen science contributions and how to systemically identify and address data quality problems. Anticipating challenges around ease of use, prompting timely interactions, and the design of the feedback system, we welcome suggestions and ideas for creating a more useful infrastructure.

References

Bonney, R.; Shirk, J.; Phillips, T.; Wiggins, A.; Ballard, H.; Miller-Rushing, A.; and Parrish, J. 2014. Next Steps for Citizen Science. *Science* 343:1436-1437.