Citizen Science 2.0: Overcoming the Participation and Data Quality Tradeoff

Roman Lukyanenko[1], Jeffrey Parsons[2], Yolanda Wiersma[3],
1 Faculty of Business Administration, Memorial University, roman.lukyanenko@mun.ca
2 Faculty of Business Administration, Memorial University, jeffreyp@mun.ca
3 Department of Biology, Memorial University, ywiersma@mun.ca

Citizen science enables non-scientists to participate in research, creating opportunities to collect and analyze data in ways that are not possible for individual researchers. Applications of citizen science are growing: volunteers are now engaged in a variety of scientific projects – from folding proteins to finding interstellar dust; from identifying birds to classifying galaxies. A central challenge in harnessing the power of human contributors is developing a technological environment capable of facilitating participation and accommodating individual perspectives of volunteers. Despite the potential for online engagement with citizen science, we argue that the prevailing assumptions and practices underlying data collection in these projects inhibit participation and data quality. We outline the root of these problems and suggest a solution designed to increase participation and data quality simultaneously.

The success of citizen science projects depends on the willingness of volunteers to report information, as well as the technological ability to represent such data faithfully. For example, popular projects such as eBird (ebird.org) or iSpot (ispot.org.uk) encourage volunteers to provide detailed information about their plant or animal sightings. In these and similar projects, citizen scientists who wish to be involved must know how to classify species. However, positive species identification can be done reasonably well only by participants with a substantial level of domain knowledge. Those non-experts who are uncertain may choose not to participate or make incorrect guesses about observed species. The result is what appears to be an unavoidable tradeoff between data quality and level of participation.

Furthermore, the focus on classification necessarily undermines data quality of citizen science projects. Reality is infinitely diverse and each sighting is unique. By abstracting from this diversity, classification limits the potential richness of communicated information for the sake of simplicity and economy. And while classification does not preclude humans from retaining details of an object, recording individual information as a member of some class in a computer system means that everything that does not "fit" into a preexisting class definition will escape structured storage. This may not be a concern for highly structured, "closed" domains, but it can undermine valuable indigenous observations of citizen scientists. The prevailing "closed world" data storage paradigm is not only contrary to the spirit of citizen science, but may severely limit its potential gains.

We contend that it is possible to create an environment that allows *broader participation by people with all levels of domain expertise while maintaining higher quality of collected information.* To achieve this dual objective, the way information is collected and stored needs to be changed. We propose an approach to data collection and storage that does not require users to identify and classify observed phenomena. Instead, they should be given the option to record observable attributes associated with the information they are contributing. The attribute-centered solution offers a new approach to data collection that removes a psychological and technological barrier to data quality and participation. The outcome is an improved ability to harness human ingenuity which, coupled with better data quality, will increase the relevance of citizen contributions to scientific research.