
© zlikovec | Fotolia

# Data quality in citizen science – a research study

**Yolanda Wiersma**, **Jeffrey Parsons** and **Roman Lukyanenko** discuss the findings of their research into how data quality can be improved for participants of citizen science.

On a bright spring day, a lifelong birdwatcher travels to her favourite green space and spends the day observing spring migrants, taking notes of the species she observes and their abundance. When she gets home, she logs into eBird.org and enters her sightings into a database that already contains millions of records from around the world.

On the same day, a group of Boy Scouts hikes along an urban river. Under the supervision of their leader, they meticulously collect water samples from different points. They note where the samples come from on labels and take data on the time of day and the water temperature. Their leader (who in his day job is a chemistry professor at the local university) takes the samples into work the next day, where he hands them to a colleague who is studying urban water quality.

That evening a teen with ambitions to be a wildlife biologist finishes her homework and logs on to her computer. Rather than killing time watching YouTube videos, she navigates to the "Snapshot Serengeti" website[1] and spends the evening identifying photos of wildebeest, gazelles and other large African mammals.

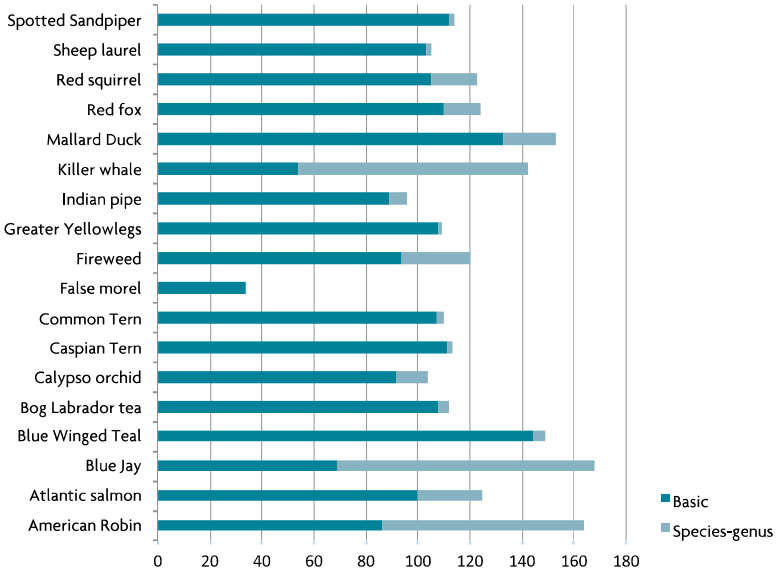
## THE BENEFITS OF CITIZEN SCIENCE

The above scenarios are all examples of 'citizen science', and the birdwatcher, Boy Scouts and teen are all 'citizen scientists'. Citizen science can involve lay people participating in many types of research projects, including medicine, environmental science, astronomy, geology, biochemistry, ecology and earth science, and has many benefits. These include educating and engaging the public on scientific issues, as well as the generation of large data sets for scientific work. Citizen science projects vary in terms of the tasks they ask the public to complete, but most projects involve citizens collecting, processing or analysing data.

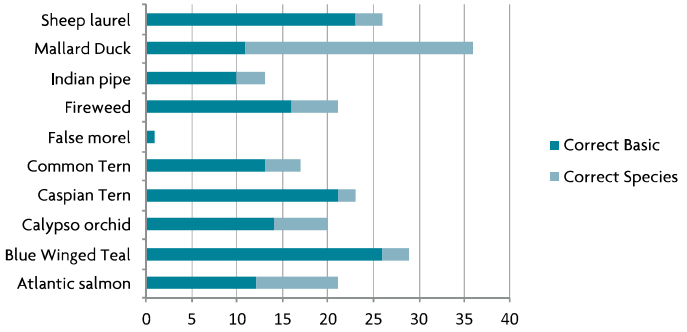
Citizen science seems like a natural win-win for all involved – citizens get the fun and learning experience by being engaged with science and scientists get the free labour. However, some people question whether lay people can actually contribute meaningfully to science. A concern frequently expressed is the quality of the data. This is particularly true when all or part of the citizen science project is online (e.g., eBird and Snapshot Serengeti), as participants may be anonymous and the risk of sabotage is higher than when scientists are in direct contact with citizen participants.

Thus, there is a need for research on the process of doing citizen science, as well as on the dimensions that affect its success; this includes motivation, types of participation and data quality. The interdisciplinary group – an ecologist working with two information system scientists – have been conducting research on the latter topic. Data quality is usually assumed to mean accuracy, but can have over 100 different dimensions[2] including precision, timeliness, completeness and believability. This research has focused on accuracy, completeness and, to a lesser extent, fitness-for-use.

**"A concern frequently expressed is the quality of the data."**




▲ **Figure 1. Number of responses in an experiment where non-biology undergraduates were asked to identify photographs of plants and animals (listed on the y-axis) in the province of Newfoundland and Labrador. The x-axis shows number of responses at the "basic" level (e.g., bird, flower and fish) in dark blue, and the species-genus (species name or general group such as salmon or orchid) in light blue. Respondents were only able to give a specific response for highly common and charismatic species (Killer Whale, Blue Jay and American Robin, for example).**




▲ **Figure 2. Number of correct species-level responses vs. predicted basic-level responses in the second experiment when participants were presented with a constrained-choice list of choices (at species and basic levels), by which to identify photographs of plants and animals in the province of Newfoundland and Labrador**

## DATA QUALITY IN CITIZEN SCIENCE

Accuracy refers to how well a data contribution matches reality. This sounds simple, but it is complicated by the fact that the measurement depends on whose 'reality' you are assessing against. For example, for a dedicated birder, an accurate observation of a given bird might be to identify it as a Eurasian Blue Tit, because they recognise that it is the species to which the bird belongs. Many natural-history themed citizen science projects (including eBird) usually require identification to the species level. But what if the person observing the bird is a beginner? Describing the sighting as a small blueish garden bird with a yellow breast would match that person's reality, and thus should be deemed accurate according to the above definition. However, a data point of "blue-yellow garden bird" is obviously not as precise as "Eurasian Blue Tit", and for avian ecologists analysing millions of bird sightings from eBird, such a data point may not be useful. On the other hand, if the beginner birder is frustrated by their lack of ability to properly log the species identification in a site like eBird, they may simply opt out of participating and the sighting would go unreported, thus rendering the data of lower quality on the dimension of completeness.

In this research study, the effects on data quality of different data collection approaches in citizen science projects, have been experimentally examined[3]. The first experiment simulated a natural history citizen science project, but in a classroom setting. University students (non-biology) were shown images of flora and fauna from the local area (the province of Newfoundland and Labrador, Canada) on a large screen. In the first experiment, students were divided into two groups; one group was asked to name the organism in the photograph (i.e., to classify it), and then describe it, while the second group was asked to simply describe the organism. It was found that, other than for very common and/or charismatic species (for example, American Robin, Blue Jay and a Killer Whale), most participants were only able to identify organisms at what cognitive psychologists call the "basic level"[4], and which represents classifications that mirror terms in common speech, or words that children first learn. A summary of answers provided at species vs. basic levels is shown in **Figure 1**. For example, the basic-level category for "American Robin" or "Blue Tit" is simply "bird". We noted, however, that in many cases participants were able to classify at levels intermediate to the species-level that scientists might desire and the basic levels which very small children might use. For example, in many cases participants were able to classify a bird more specifically as "gull", "duck" or "shorebird". In a second constrained-response experiment, where the same images were used, participants were offered correct and incorrect options at basic, sub-basic and species levels. Again, a significant number of the non-biology students preferred to describe the images at levels above the species level, and were more accurate when reporting at higher levels, as shown in **Figure 2**.

While the traditional definition of accuracy in citizen science is the extent to which an observation provided by the citizen scientists matches that needed by the scientists, a more suitable definition is "agreement with reality as perceived by the data contributor (citizen scientist)". Under this proposed definition, the results suggest accuracy in citizen science data is improved when citizens are allowed to contribute data at the level they feel comfortable, rather than when scientists impose a requirement to contribute in a way that adheres to scientific standards of accuracy. This work has also shown that allowing such flexibility in data contribution also increases data completeness[5]. In a parallel experiment, a real online environment[6] was used that allowed members of the public to contribute sightings of plants and animals in the province of Newfoundland and Labrador. Participants were again divided into two groups; those in one group were required to classify their sighting by species (the interface was constrained such that non-species names were not permitted; however, participants had the option to select "I don't know"), whereas participants in the other group used a more flexible interface in which they were allowed to describe a sighting in any way they wished. There was a significant difference in the total

number of contributions between the two groups, as well as in the number of observations per person. This suggests that a flexible interface approach facilitates a more complete data set.

**"Accuracy refers to how well a data contribution matches reality. This sounds simple, but it is complicated by the fact that the measurement depends on whose 'reality' you are assessing against."**

### NON-TRADITIONAL APPROACHES TO DATA

This work suggests that freeing citizen scientists (non-experts) from the data entry constraints imposed by scientists/experts may increase the data quality dimensions of accuracy and completeness. An important question that follows is whether such (rather unconventional) data can actually have utility for scientists. The preliminary results suggest that they can. Most ecologists will require species-level identifications, so this means the data requires some post-processing to be useful. An additional study was conducted whereby the attributes that the citizen scientists used to describe their sightings to natural history experts in a sort of "guessing-game" experiment, were presented. Most of the time, the experts were able to use this information to infer the species (or at least infer a probable species), thus rendering the data more useful. Had citizen scientists been required to provide species names, many of the participants would have been unable to participate and these observations would not have been provided.

Outside of the directed experiments on data quality, the website, nlnature.com, has serendipitously contributed to science by facilitating the reporting of a new mosquito species to the province[7], which may be a possible vector for the West Nile virus. The ability of citizen scientists to spot something novel is documented most famously in astronomy in the citizen science project "Galaxy Zoo," where a Dutch school teacher, Hanny Van Arkel, identified a new type of celestial body in classifying objects in images taken by the Hubble space telescope. The Galaxy Zoo project directed citizens to group images of galaxies into one of three shapes[8], but Van Arkel used the online forum to communicate a sighting that did not fit the pre-defined categories. This further illustrates the impact that pre-defined categories can have on data quality[9]. Had Van Arkel not taken the initiative to alert the project sponsors to this new object, Hanny's Voorwerp[10] might have gone unknown to science.

Through this project's experimental work in citizen science, it has been shown that citizens are capable of providing accurate and complete information, as long as scientists adopt a more inclusive view of data quality. ES

**Yolanda Wiersma** is an Associate Professor in Biology at Memorial University of Newfoundland.

**Jeffrey Parsons** is a University Research Professor in the Faculty of Business Administration at Memorial University of Newfoundland.

**Roman Lukyanenko** is an Assistant Professor of Information Systems at Edwards School of Business, University of Saskatchewan.

### REFERENCES

1. Snapshot Serengeti (2016) Snaptshot Serengeti – a Zooniverse project. www.snapshotserengeti.org
2. Wang, R.Y. and Strong, D.M. (1996) Beyond accuracy: what data quality means to data consumer. Journal of Management Information Systems, 12, pp.5-33.
3. Lukyanenko, R., Parsons, J. and Wiersma, Y.F (2014) The IQ of the crowd: understanding and improving information quality in structured user-generated content. Information Systems Research, 25, pp. 669–689.
4. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. and Boyesbraem, P. (1976) Basic objects in natural categories. Cognitive Psychcology, 8(3), pp. 382–439.
5. Lukyanenko, R., Parsons, J. and Wiersma, Y.F. (2014) The impact of conceptual modelling on dataset completeness: a field experiment. Proceedings of the International Conference on Information Systems 2014. Association for Information Systems, Atlanta. http://aisel.aisnet.org/icis2014/proceedings/GeneralIS/29/
6. NLNature (2016) Newfoundland Nature – a Citizen Science project to document nature in Newfoundland and Labrador, www.nlnature.com
7. Fielden, M.A., Chaulk, A.C., Bassett, P.K., Wiersma, Y.F., Erbland, M., Whitney, H. and Chapman, T.W. (2015) Aedes japonicas (Diptera: Culicidae) arrives at the most easterly point in North America. The Canadian Entomologist 147, pp. 737-740.
8. Hopkin, M. (2007) See new galaxies – without leaving your chair. Nature News Online, 11th July.
9. Lukyanenko R., Parsons, J., Wiersma, Y.F. (2016) Editorial: Emerging problems of data quality in citizen science. Conservation Biology 30(3): 447-449.
10. Lintott, C.J., Schawinski, K., Keel, W., Van Arkel, H., Bennert, N., Edmondson, E., Thomas, D., Smith, D.J.B., Herbert, P.D., Jarvis, M.J., Virani, S., Adreescu, D., Bamford, S.P., Land, K., Murray, P., Nichol, R.C., Raddick, M.J., Slosar, A., Szalay, A. and Vandenberg, J. (2009). Galaxy Zoo: Hanny's Voorwerp, a quasar light echo? Monthly Notices of the Royal Astronomical Society 399, pp. 129–140.