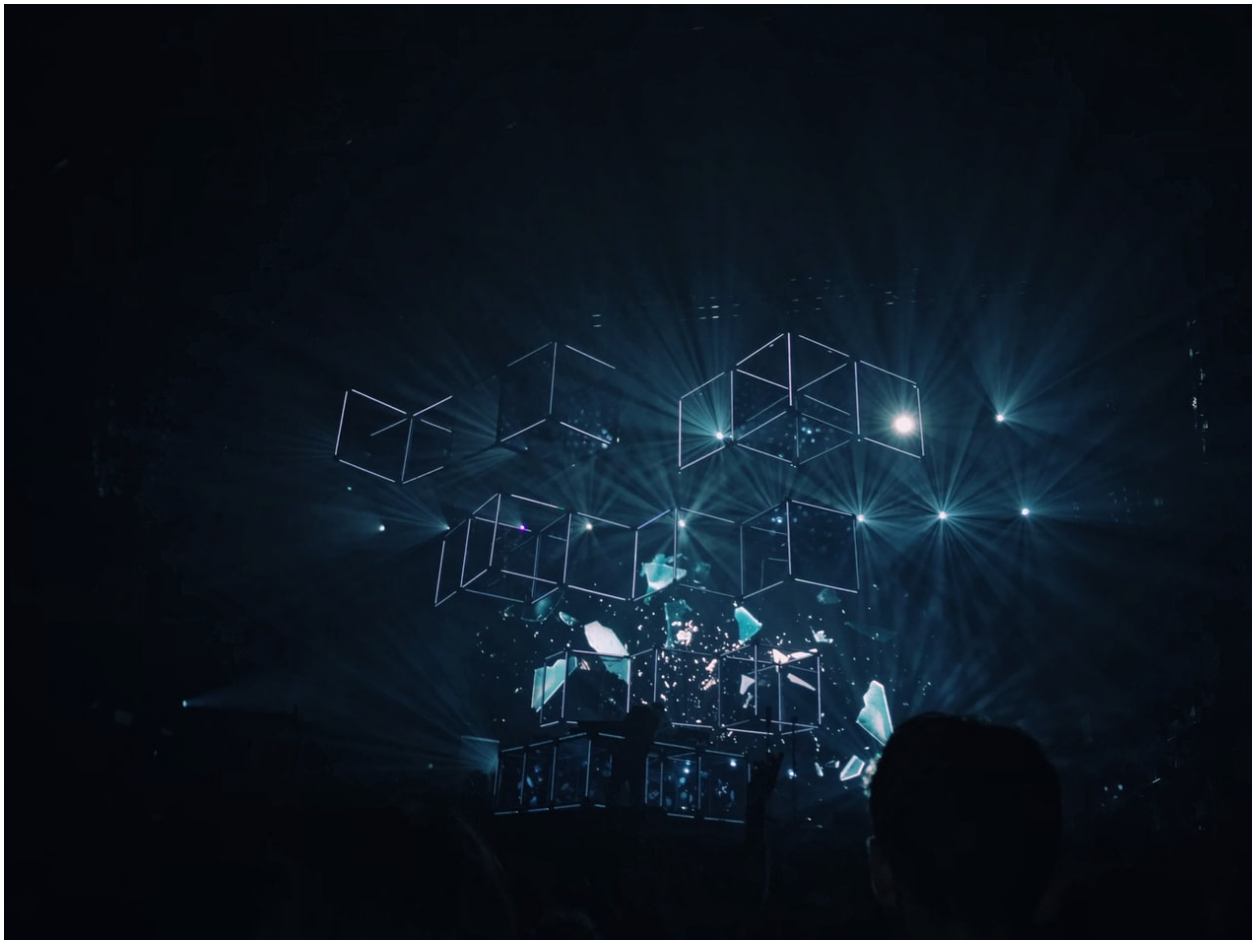
Citizen Science Projects (MOOC) 3.3

Woods, Mel; Coulson, Saskia; Ajates, Raquel; Amditis, Angelos ; Cobley, Andy; Domian, Dahlia

Publication date:
2020



Data quality is paramount in citizen science. Many organisations and researchers have been asking: Can citizens provide data of the same quality as professional scientists? In this step, we'll find out why data quality matters. We'll also look at how to manage data quality when you are using sensors, like smartphones, and collecting observations from citizen scientists.

##Reliable and valid data

We can look at data quality in terms of both reliability and validity. **Data reliability** (or replicability) is about whether you can get the same results when you repeat an experiment or make an observation. For example, let's say that you see a bird and identify that bird as a robin. At the same time, someone else sees the same bird and comes to the same conclusion. The data is reliable because multiple observations have given the same result.

Data validity, on the other hand, is about how credible or trustworthy the data are. The data collected are valid if they correctly represent the real world. For example, maybe you measure air quality using a low-cost sensor that has not been calibrated correctly. You may get reliable (consistent) measurements from the sensor but the data will not be valid.

##So how can we make sure the data quality is good?

The Scent Citizen Observatory handles quality assurance in different ways, depending on the types of data that citizen scientists produce.

Using the [Scent Explore](https://scent-project.eu/scent-toolbox) mobile application, citizens are guided to areas where we need environmental information. For those locations, they can collect images of land cover/land use and add text descriptions.

Once a photo is taken, it is stored in a crowdsourcing platform and is available to the Scent Intelligence Engine (SIE). This tool uses machine learning to automatically detect land cover types and objects in an image according to Scent's classification system. This information is fed back to the crowdsourcing platform, where the tool decides whether the annotation's quality is good enough. If it is not, the image is sent for further annotation, this time by people, through [Scent Collaborate](http://collaborate.scent-project.eu/).

The system works by assigning a score to each annotation tag. The score is based on two factors:
+ if the annotation came from the user who created the image in Scent Explore
+ if the Scent Intelligence Engine added the annotation and how confident it was in its results.

If the score is high enough, the annotation is considered valid. If not, the annotation is invalid and has to be manually updated in Scent Collaborate.

Another data quality control system is also available for measurements collected by portable soil and air temperature sensors. The first step in this system is to identify measurements that are potentially invalid before they get into the system. This could be air temperature values outside of a predefined range, for example. The selection criteria are strict so that we are only examining valid measurements, rather than letting incorrect values into the system at all.

A user rating scheme is also in use, allowing the system to judge the users' data quality. This numbered score increases or decreases depending on the quality of measurements that the user has collected. Each user receives a score for each collected measurement. If a measurement is evaluated as a faulty one, points are deducted, decreasing the user's total score, which indicates how trustworthy the user is.

LandSense also uses a quality assurance system to keep data collected for land cover detection, agricultural monitoring and habitat monitoring campaigns clean. First, the system checks for overlaps in areas drawn by users and flags them so that users can correct them. Next, the system looks for problems with photographs. Many citizen science projects have mobile apps that ask citizens to take photographs as part of the data collection process, but this can create issues of personal privacy. For example, faces and license plates are automatically blurred out to comply with the EU General Data Protection Regulation (GDPR). Photographs

are also checked to make sure they are not too dark or blurry. Next, the service checks for accuracy of position, using mobile phone GPS to make sure an observation is geographically accurate. It also checks against reference data from a ‘gold standard’ dataset produced by professional scientists to make sure there is an overall agreement, and nothing is omitted or entered incorrectly. Finally, the service compares answers from the same location given by multiple contributors to provide a level of confidence in the data.

Data quality is only one important factor in citizen science. Keeping data collectors engaged and motivated is another. In the next step, we discuss ways of stimulating and sustaining participation, both of which are challenges you might face in your own observatory or campaign.

##Share your thoughts!

Think about an environmental concern of your interest that could be the focus of a citizen science project. What do you think are some ways that you could ensure the data collected were high-quality?

Please share your ideas with us in the discussion area below.