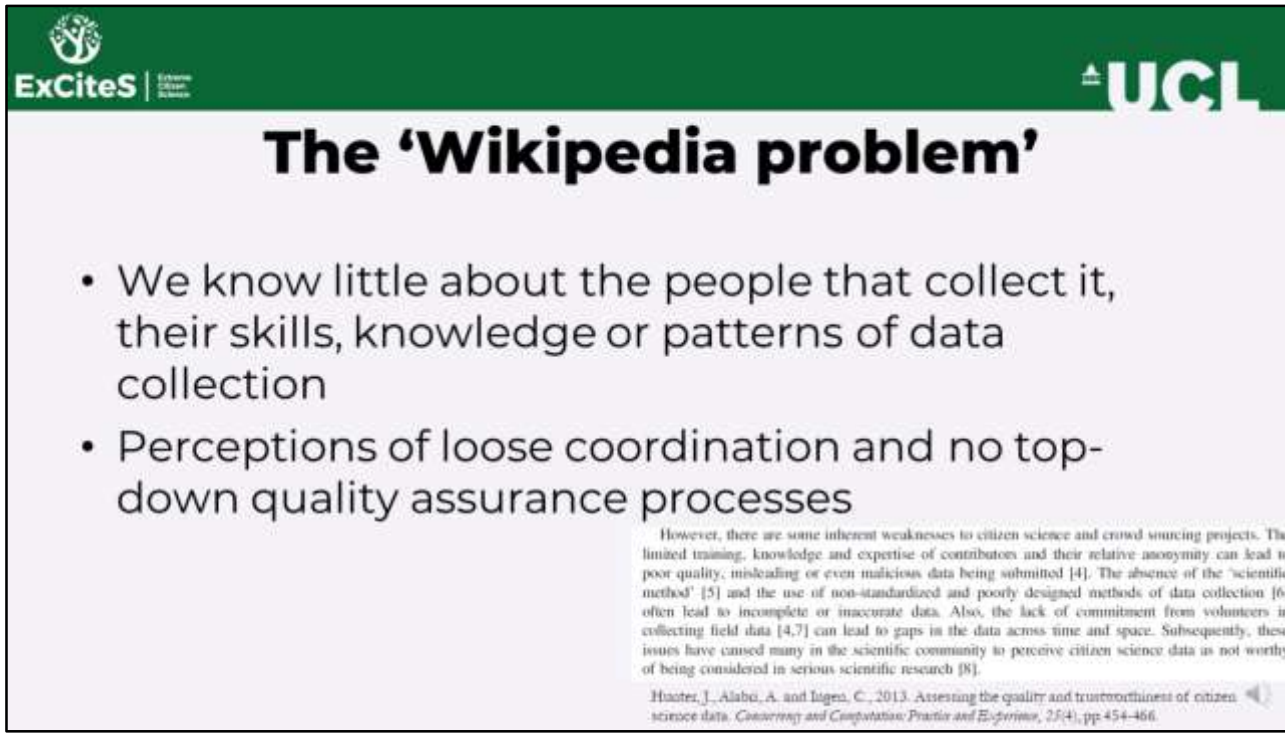
Today, we turn to data management. Without good data, we cannot do science…



When it comes to citizen science, it is very common to hear suggestions that the data is not good enough and that volunteers cannot collect data at a good quality, because unlike trained researchers, they don't understand who they are – a perception that we know little about the people that are involved and therefore we don't know about their ability. There are also perceptions that like Wikipedia, it is all very loosely coordinated and therefore there are no strict data quality procedures. However, we know that even in the Wikipedia case that when the scientific journal *Nature* showed over a decade ago (2005) that Wikipedia has similar quality to Encyclopaedia Britannica.

In citizen science where sensing and data collection from instruments is included, there are also concerns over the quality of the instruments and their calibration – the ability to compare the results with high end instruments.

The opening of the Hunter et al. paper (which offers some solutions), summarises the concerns that are raised over data

ExCiteS | Extreme Citizen Science

UCL

Underlying concerns

- Data quality concerns linked to professional standing and roles – science version of "the cult of the amateur"
- Special concerns when citizen science linked to activism
- Basic unfamiliarity with crowdsourcing mechanisms

20 AUGUST 2015 | VOL 524 | NATURE | 263

THIS WEEK

EDITORIALS

Rise of the citizen scientist

From the oceans to the soil, technology is changing the part that amateurs can play in research. But this greater involvement raises concerns that must be addressed.

Critics have raised concerns about data quality, and some studies do find that volunteers are less able to identify plant species than are academics and land managers. And there are issues around how to reward and recognize the contribution of volunteers, and around ensuring that data are shared or kept confidential as appropriate. But these problems seem relatively simple to address — not least because they reflect points — from authorship to data quality and access — that the professional scientific community is already wrestling with.

Based on conversations with scientists and concerns that are appearing in the literature, there is also a cultural aspect at play which is expressed in many ways – with data quality being used as an outlet to express them. This can be similar to the concerns that were raised in the cult of the amateur to protect the position of professional scientists and to avoid the need to change practices. There are also special concerns when citizen science is connected to activism, as this seems to "politicise" science or make the data suspicious – we will see next lecture that the story is more complex. Finally, and more kindly, we can also notice that because scientists are used to top down mechanisms, they find alternative ways of doing data collection and ensuring quality unfamiliar and untested.

Against this background, it is not surprising to see that checking data quality in citizen science is a popular research topic. Caren Cooper have identified over 50 papers that compare citizen science data with those that were collected by professional – as she points: "To satisfy those who want some nitty gritty about how citizen science projects actually address data quality, here is my medium-length answer, a brief review of the technical aspects of designing and implementing citizen science to ensure the data are fit for intended uses. When it comes to crowd-driven citizen science, it makes sense to assess how those data are handled and used appropriately. **Rather than question whether citizen science data quality is low or high, ask whether it is fit or unfit for a given purpose.** For example, in studies of species distributions, data on presence-only will fit fewer purposes (like invasive species monitoring) than data on presence *and absence*, which are more powerful. Designing protocols so that citizen scientists report what they do *not* see can be challenging which is why some projects place special emphasize on the importance of "zero data."

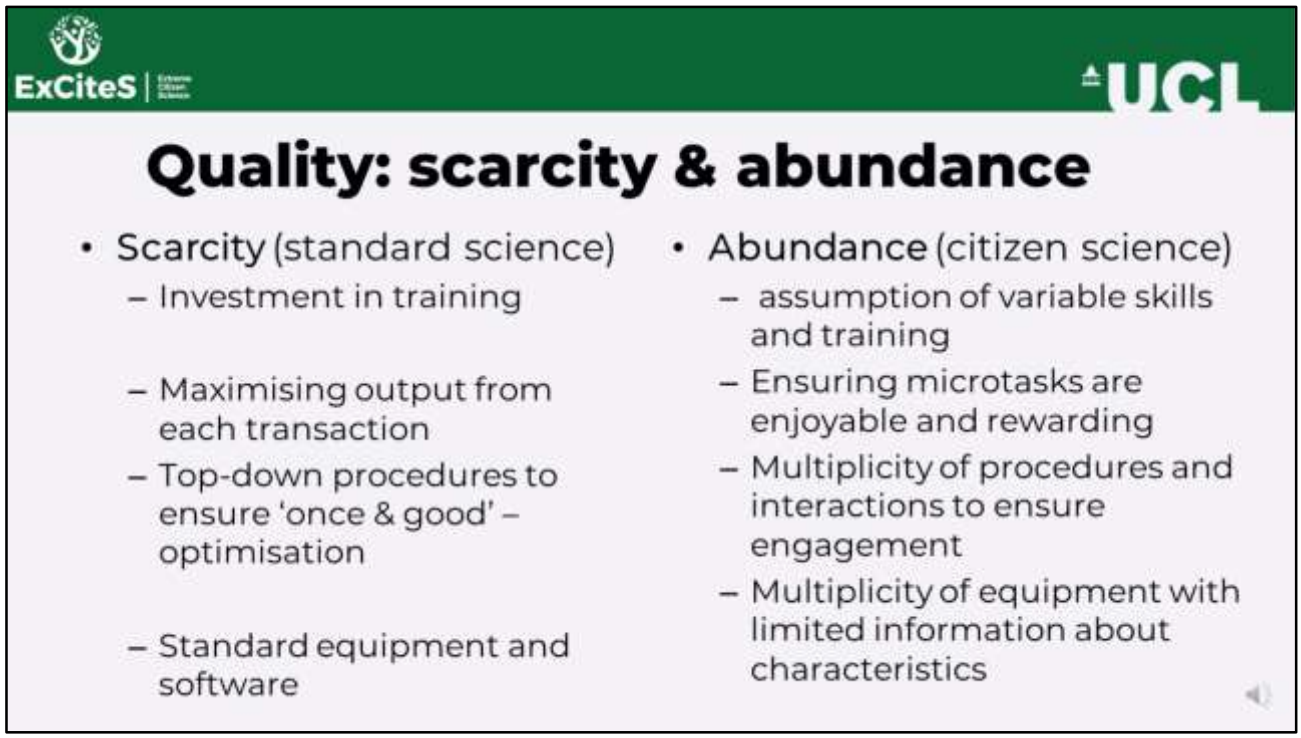
It is a misnomer that the quality of each individual data point can be assessed without context. Yet one of the most common way to examine citizen science data quality has been to compare volunteer data to those collected by trained technicians and scientists. Even a few years ago I'd noticed over 50 papers making these types of comparisons and the overwhelming evidence suggested that volunteer data are fine. And in those few instances when volunteer observations did not match those of professionals, that was evidence of poor project design. While these studies can be reassuring, they are not

always necessary nor would they ever be sufficient."
(http://blogs.plos.org/citizensci/2016/12/21/quality-and-quantity-with-citizen-science/)

One way to examine the issue with data quality is to think of the clash between two concepts and systems of thinking on how to address quality issue – we can consider the condition of standard scientific research conditions as ones of scarcity: limited funding, limited number of people with the necessary skills, a limited laboratory space, expensive instruments that need to be used in a very specific way – sometimes unique instruments.

The conditions of citizen science, on the other hand are of abundance – we have a large number of participants, with multiple skills, but the cost per participant is low, they bring their own instruments, use their own time, and are also distributed in places that we usually don't get to (backyards, across the country – we talked about it in week 2).

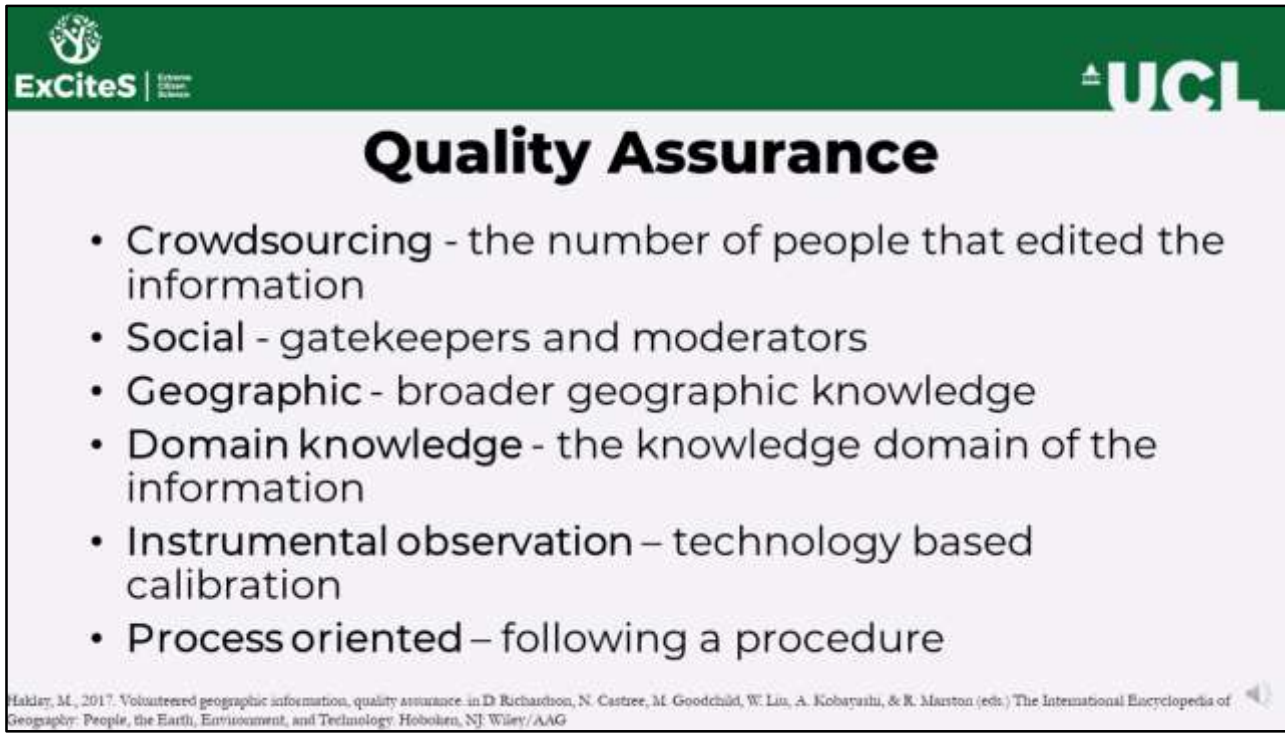
Conditions of abundance are different and require different thinking for quality assurance



Here some of the differences. Under conditions of scarcity, it is worth investing in long training to ensure that the data collection is as good as possible the first time it is attempted, since time is scarce. Also we would try to maximise the output from each activity that our researcher carried out, and we will put procedures and standards to ensure "once & good" or even "once & best" optimisation. We can also force all the people in the study to use the same equipment and software, as this streamlines the process.

On the other hand, in abundance conditions we need to assume that people are coming with a whole range of skills and that training can be variable – some people will get train on the activity over a long time, while to start the process we would want people to have light training and join it. We also thinking of activities differently – e.g. conceiving the data collection as micro-tasks. We might also have multiple procedures and even different ways to record information to cater for different audience. We will also need to expect a whole range of instrumentation, with sometimes limited information about the characteristics of the instruments.

Once we understand the new condition, we can come up with appropriate data collection procedures that ensure data quality that are suitable for this context



There are multiple ways of ensuring data quality in citizen science data. Let's briefly look at each one of these. The first 3 methods were suggested by Mike Goodchild and Lina Li in a paper from 2012

ExCiteS | UCL

Crowdsourcing

- Using collective wisdom, or aggregating individual responses
- Each piece of data is being evidenced by multiple observers/analysers

Crowdsourced science: sociotechnical epistemology in the e-research paradigm

David Watson, Luciano Floridi

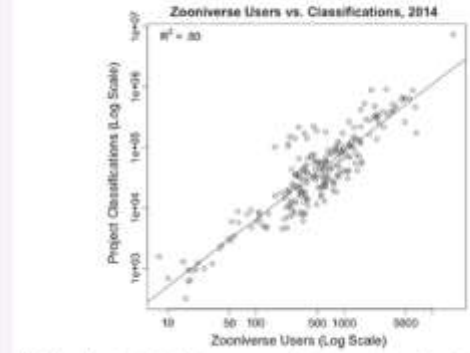


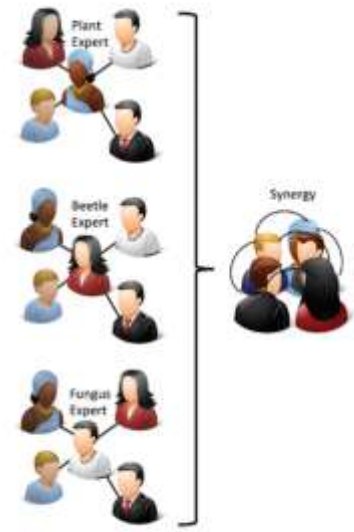
Fig. 3 Log–log scatterplot of Zooniverse users versus classifications, with an ordinary least squares regression line fit to the data

The first method for quality assurance is crowdsourcing – the use of multiple people who are carrying out the same work, in fact, doing peer review or replication of the analysis which is desirable across the sciences. As Watson and Floridi argued, using the examine of Zooniverse, the approaches that are being used in crowdsourcing give these methods a stronger claim on accuracy and scientific correct identification because they are comparing multiple observers who work independently

ExCiteS | UCL

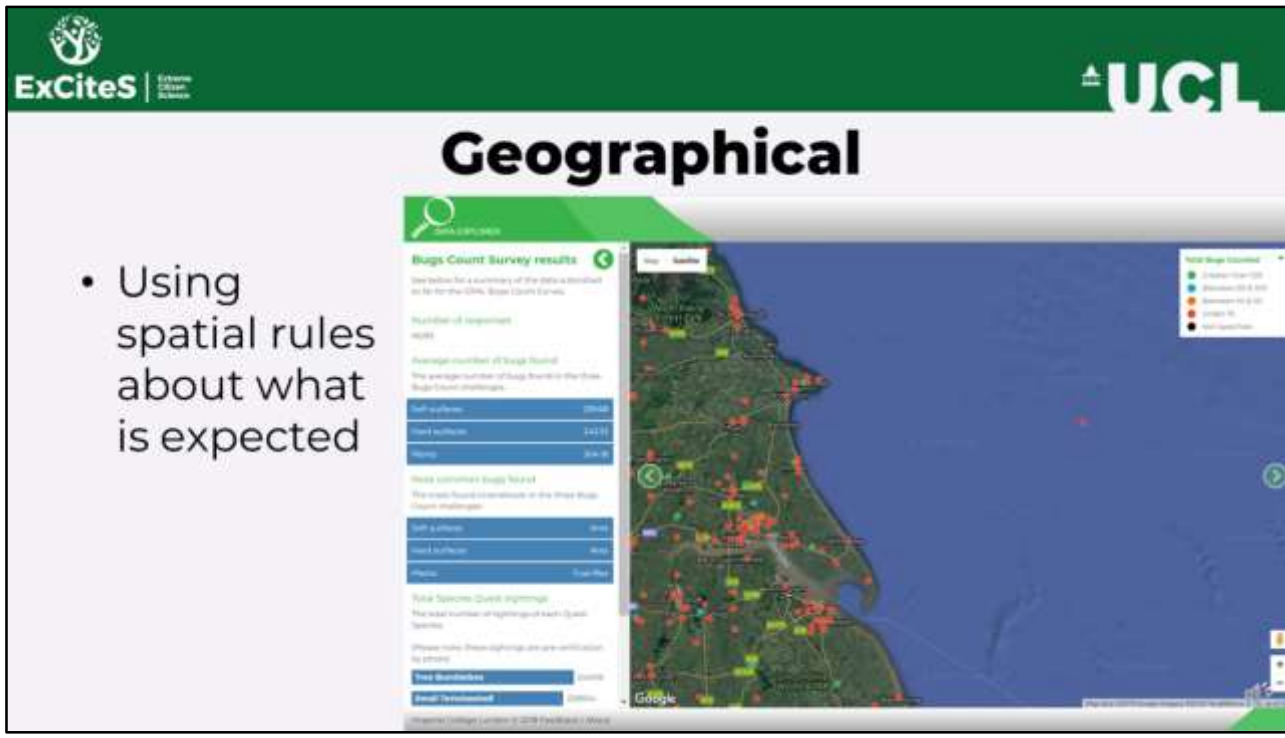
Social

- Social quality assurance use a hierarchy of participants, with those with known expertise checking and assisting other participants
- iSpot was designed as a social network that support this



Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J. and McConway, K., 2015. Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys*, (480), p.125.

The social form of quality assurance is using more and less experienced participants as a way to check the information and ensure that the data is correct. This is fairly common in many areas of biodiversity observations and integrated into iSpot, but also exist in other areas, such as mapping, where some information get moderated (we've seen that in Google Local Guides, when a place is deleted).

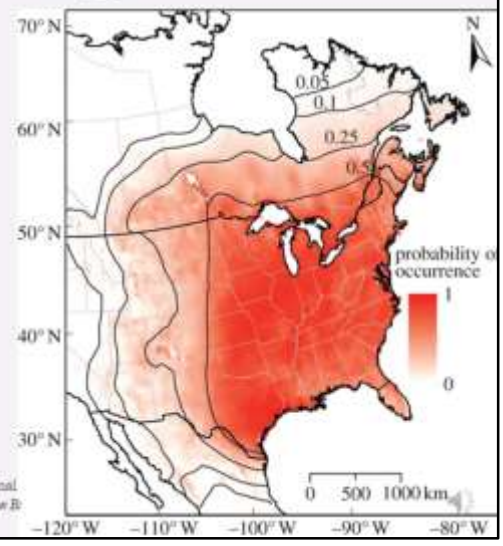


The geographical rules are especially relevant to information about mapping and locations. Because we know things about the nature of geography – the most obvious is land and sea in this example – we can use this knowledge to check that the information that is provided make sense, such as this sample of two bumble bees that are recorded in OPAL in the middle of the sea. While it might be the case that someone seen them while sailing or on some other vessel, we can integrate a rule into our data management system and ask for more details when we get observations in such a location. There are many other such rules – about streams, lakes, slopes and more.

ExCiteS | UCL

Domain knowledge

- In many scientific domains, we can use knowledge from what we already know about geographical, temporal, and other characteristics



Flockhart, D.T., Wassenaar, L.I., Martin, T.G., Hobson, K.A., Wunder, M.B. and Norris, D.R., 2013. Tracking multi-generational colonization of the breeding grounds by monarch butterflies in eastern North America. *Proceedings of the Royal Society of London B: Biological Sciences*, *280*(1768), p.20131087.

The 'domain' approach is an extension of the geographic one, and in addition to geographical knowledge uses a specific knowledge that is relevant to the domain in which information is collected. For example, in many citizen science projects that involved collecting biological observations, there will be some body of information about species distribution both spatially and temporally. Therefore, a new observation can be tested against this knowledge, again algorithmically, and help in ensuring that new observations are accurate. If we see a monarch butterfly within the marked area, we can assume that it will not harm the dataset even if it was a mistaken identity, while an outliers (temporally, geographically, or in other characteristics) should stand out.

ExCiteS | UCL

Instrumental observation

- Instrumental observation provide us with an evidence from a sensor about the measurement

The ‘instrumental observation’ approach remove some of the subjective aspects of data collection by a human that might made an error, and rely instead on the availability of equipment that the person is using. Because of the increased in availability of accurate-enough equipment, such as the various sensors that are integrated in smartphones, many people keep in their pockets mobile computers with ability to collect location, direction, imagery and sound. For example, images files that are captured in smartphones include in the file the GPS coordinates and time-stamp, which for a vast majority of people are beyond their ability to manipulate. Thus, the automatic instrumental recording of information provide evidence for the quality and accuracy of the information. This is where the metadata of the information become very valuable as it provides the necessary evidence.

ExCiteS | UCL

Process oriented

- Ensuring that participants follow an exact protocol that ensure standardised data collection

Finally, the ‘process oriented’ approach bring citizen science closer to traditional industrial processes. Under this approach, the participants go through some training before collecting information, and the process of data collection or analysis is highly structured to ensure that the resulting information is of suitable quality. This can include provision of standardised equipment, online training or instruction sheets and a structured data recording process. For example, volunteers who participate in the US Community Collaborative Rain, Hail & Snow network (CoCoRaHS) receive standardised rain gauge, instructions on how to install it and an online resources to learn about data collection and reporting.

Combination of methods

- When project organisers are asked, they describe multiple methods (notice that these can be grouped in the above groups)
- Many times methods are combined (75% of the time)

Table I
VALIDATION METHODS REPORTED

Method	n	Percentage
Expert review	46	77%
Photo submissions	24	40%
Paper data sheets submitted along with online entry	20	33%
Replication or rating, by multiple participants	14	23%
QA/QC training program	13	22%
Automatic filtering of unusual reports	11	18%
Uniform equipment	9	15%
Validation planned but not yet implemented	5	8%
Replication or rating, by the same participant	2	3%
Rating of established control items	2	3%
None	2	3%
Not sure/don't know	2	3%

Table II
COMBINATIONS OF MECHANISMS REPORTED

Methods	n	Percentage
Single method	10	17%
Multiple methods, up to 5 (average of 2.5)	45	75%
Expert review + Automatic filtering	11	18%
Expert review + Paper data sheets	10	17%
Expert review + Photos	14	23%
Expert review + Photos + Paper data sheets	6	10%
Expert review + Replication, multiple	10	17%

Wiggins, A., Newman, G., Stevenson, R.D. and Crowston, K., 2011, December. Mechanisms for data quality and validation in citizen science. In *e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on* (pp. 14-19). IEEE.

What is important to be aware of is that methods are not being used alone but in combination. The analysis by Wiggins et al. in 2011 includes a framework that includes 17 different mechanisms for ensuring data quality. It is therefore not surprising that with appropriate design, citizen science projects can provide high quality data