frontiers in Climate

REVIEW
published: 09 June 2021
doi: 10.3389/fclim.2021.645120

Seven Primary Data Types in Citizen Science Determine Data Quality Requirements and Methods

Robert D. Stevenson[1]*, *Todd Suomela*[2], *Heejun Kim*[3] and *Yurong He*[4]

[1] Department of Biology, University of Massachusetts Boston, Boston, MA, United States, [2] Digital Pedagogy and Scholarship Department, Bucknell University, Lewisburg, PA, United States, [3] Department of Information Science, University of North Texas, Denton, TX, United States, [4] College of Information Studies, University of Maryland, College Park, MD, United States

Data quality (DQ) is a major concern in citizen science (CS) programs and is often raised as an issue among critics of the CS approach. We examined CS programs and reviewed the kinds of data they produce to inform CS communities of strategies of DQ control. From our review of the literature and our experiences with CS, we identified seven primary types of data contributions. Citizens can carry instrument packages, invent or modify algorithms, sort and classify physical objects, sort and classify digital objects, collect physical objects, collect digital objects, and report observations. We found that data types were not constrained by subject domains, a CS program may use multiple types, and DQ requirements and evaluation strategies vary according to the data types. These types are useful for identifying structural similarities among programs across subject domains. We conclude that blanket criticism of the CS data quality is no longer appropriate. In addition to the details of specific programs and variability among individuals, discussions can fruitfully focus on the data types in a program and the specific methods being used for DQ control as dictated or appropriate for the type. Programs can reduce doubts about their DQ by becoming more explicit in communicating their data management practices.

Keywords: citizen science, data quality, data type, data quality requirment, data quality methods

OPEN ACCESS

Edited by:
Alex de Sherbinin,
Columbia University, United States

Reviewed by:
Maria Rosa (Rosy) Mondardini,
University of Zurich, Switzerland
Martie Van Deventer,
University of Pretoria, South Africa

***Correspondence:**
Robert D. Stevenson
robert.stevenson@umb.edu

Specialty section:
This article was submitted to Climate Risk Management, a section of the journal Frontiers in Climate

Received: *22 December 2020*
Accepted: *05 May 2021*
Published: *09 June 2021*

Citation:
Stevenson RD, Suomela T, Kim H and He Y (2021) Seven Primary Data Types in Citizen Science Determine Data Quality Requirements and Methods. Front. Clim. 3:645120. doi: 10.3389/fclim.2021.645120

INTRODUCTION

Citizen science encompasses a variety of activities in which citizens are involved in doing science (Shirk et al., 2012; Haklay, 2013; Thiel et al., 2014; Cooper, 2016). Part of the excitement about CS is the number of scientific disciplines that have adopted a citizen science approach. For instance, astronomy has used CS to map galaxies (Galaxy Zoo), chemistry to understand protein folding (FoldIt), computer science to refine algorithms (SciPy), ecology to document coral reef biodiversity (REEF), environmental science to monitor water quality (Acid Rain Monitoring Project), and geography to map features of cities (OpenStreetMap). CS is a rapidly expanding field involving over 1,000 advertised projects (Scistarter websites). Pocock et al. (2017) identified over 500 CS projects in the ecology and environmental area alone.

At the center of many citizen science programs is the contribution citizens make to gathering and/or scoring observations (Miller-Rushing et al., 2012; Shirk et al., 2012; Bonney et al., 2014, 2016), but concerns regarding citizen contributions arise for several reasons (Cohn, 2008; Riesch and Potter, 2014; Burgess et al., 2017). By definition, participants share a common interest to participate but are not trained experts (Thiel et al., 2014; Cooper, 2016; Eitzel et al., 2017) leading to inherent doubt about their abilities (Cohn, 2008; Bonney et al., 2014, 2016). Citizen science participants may be trained for the specific tasks of the programs in which they participate, but there is often no requirement for them to have formal training, accreditation, or a degree (Freitag et al., 2016). Furthermore, there may be no requirement for participants to regularly practice the skills needed.

In our experience, CS program managers are well aware that the quality of the scientific data their programs produce is paramount to success. A survey by Hecker et al. (2018) suggests that after funding considerations, data quality is the most important concern for program managers (also see Peters et al., 2015). Significant progress is being made in understanding and improving DQ in citizen science. Many papers have been written assessing the DQ of a specific project, and papers starting around 2010 have provided broader context (Alabri and Hunter, 2010; Haklay et al., 2010; Sheppard and Terveen, 2011; Wiggins et al., 2011; Goodchild and Li, 2012; Crowston and Prestopnik, 2013; Hunter et al., 2013; Thiel et al., 2014; Kosmala et al., 2016; Lukyanenko et al., 2016; Muenich et al., 2016; Blake et al., 2020; López et al., 2020). Also, there have been efforts to compare data quality across projects (Thiel et al., 2014; Aceves-Bueno et al., 2017; Specht and Lewandowski, 2018).

A number of papers have focused on DQ as part of the process of data collection/data life cycle (Wiggins et al., 2011; Kelling et al., 2015a; Freitag et al., 2016; Parrish et al., 2018a), and some have examined the variability of individual contributors (Bégin et al., 2013; Bernard et al., 2013; Kelling et al., 2015b; Johnston et al., 2018). Kosmala et al. (2016) and Parrish et al. (2019) emphasized the importance of individual program's protocols for DQ. In this paper, we examined citizen science programs from the point of view of the kinds of data they produce with the goal of informing the strategies of DQ control. This reasoning leads to the questions addressed here, "Are there primary types of data produced by citizen science projects?" and if so, "What are the ramifications of these types for DQ analysis and project design?"

METHODS

Scopus literature searches were performed using the term "data quality" in combination with the terms "citizen science," "volunteered geographic information," or "volunteer monitoring." A total of 293 papers were found from the published literature between the years 1994 and 2020. Papers were reviewed and discussed among our team using the general data quality framework provided by Wiggins et al. (2011). Investigations were performed using categorical analysis and decision trees. Additional efforts were made to collect the needed information from project web sites, but these sites proved difficult to navigate from the perspective of locating information about data quality methods. It was often unclear whether the information we sought was available or not. Our lack of success in searching on project websites leads us to look more carefully into the heterogeneity of citizen science projects, and specifically into the heterogeneity of data produced by CS projects. An iterative process of re-reading the literature, investigating papers cited in the literature, and re-examining project web sites produced the categorization of the primary data types reported here.

RESULTS

Categories of Data From Citizen Science Projects

Our review identified seven primary categories of data contributions made by people to citizen science projects (**Table 1**). Citizens can carry instrument packages, invent or modify algorithms, sort and classify physical objects, sort and classify digital objects, collect physical objects, collect digital objects, and report observations. In the following paragraphs, we describe each of these types and then turn to the implications for DQ requirements and project design.

In the simplest data type, a citizen's designated role is limited to transporting and/or maintaining standard measurement devices (**Table 1**). People carry instrument packages (CIP) or pilot vehicles that carry instrument packages. There is no active role in monitoring or recording data once the instrument is in place. Citizens also bear the cost of carrying the sensors. Weather Underground is an example of such a program. The benefit to the project is that no investment is needed other than arranging the transport of or giving advice about device options, installation, and providing a data sharing and storage website. With this limited role for participants, there are fewer concerns about data quality. Projects can rely on strategies normally employed by scientists when monitoring instrument packages that are deployed.

The second category of participation involves the invention or modification of algorithms (IMA) such as the Foldit project in which citizens help discover the sequence of proteins folds or a search such as the Great Internet Mersenne Prime Search in which citizens help search for class of prime numbers. This kind of citizen science project may take the form of a game or contest. The contributions of participants are explicitly recorded and tested in a public arena. The success of algorithms is usually known to all, and the insights of a citizen or citizen team can often be incorporated by others in subsequent submissions. Data quality is not an issue for these projects. Keeping track of the history of the algorithm submissions is part of the process, so provenance is also inherently addressed.

The third type of project involves the sorting and classifying of physical objects (SCPO). In these projects, scientists already have an existing source of data but need help organizing the collection. Fossils or archeology artifacts are two examples of

TABLE 1 | Seven basic types of data contributions made to citizen science projects with examples.

Data category	Data contribution	Example 1		Example 2		Example 3	
		Project name	**Description**	**Project name**	**Description**	**Project name**	**Description**
Carry Instrument packages (CIP) or pilot vehicles that CIPs	Indirectly through deployment of instrument package	Air Quality Citizen Science	Monitor environmental air quality	SeaKeepers International	Works with NOAA and WMO and deploys Seakeeper Difters and Argo floats	Weather Underground	Connects consumer weather instruments in a network
Invent or modify algorithms, IMA	Algorithms, beat the best computer algorithms	Fold-It	Submit steps for protein 3-D folding to understand protein function	MATLAB Online Programming Contest	Develop and share code to solve computing challenge	EteRNA	Submit steps for RNA 3-D folding to understand RNA function
Sort and classify physical objects, SCPO	Object categorized	Passport in Time	Contribute to field archeology program with the USFS	Field Museum Collection Center Volunteers	Count, sort and digitize artifacts and specimens	American Museum of Natural History	Volunteering in the Division of Paleontology
Sort and classify digital objects, SCDO	Digital object categorized	Galaxy Zoo	Classify galaxies from digital images	EyeWire	Map neurons in the eye of *Drosophila*	Old Weather	Transcribe weather records from ships' logs
Collect physical objects, CPO	Sample obtained and submitted, collection process documented	Florida LakeWatch	Collect water samples for analysis	School of Ants	Collect ants around schools that are submitted for identification	The Bighorn Basin Dinosaur Project	Find and collect dinosaur fossils
Collect digital objects, CDO	Digital object obtained and submitted, collection process documented	Juneau Humpback Whale Flukes	Collect images of whale flukes	BatME	Collect audio recordings of bats with mobile devices	PicturePost	Contribute digital images of landscape
Report observations, RO	Text from instrument readings, counts, classifications, and/or descriptions	Great Sunflower Project	Record pollinator activity in gardens	CoCoRaHS	Submit data about rainfall, hail events, and snow fall	Feeder Watch	Counts bird species that visit bird feeders

physical objects that can be organized in this type of project. The projects are location-specific, and citizens are usually part of the local science team. Citizens and scientists work together closely, and questions about data quality are quickly resolved because people with appropriate expertise can be easily consulted.

In the fourth type, the digital cousin of the third category, participants sort and classify digital objects (SCDO). Objects are in the form of photographs, audio recordings, or videos that were collected and organized by scientists, and they need to be sorted and classified. These data can be easily shared electronically using the internet. This approach has greatly expanded the opportunity for participation because the activities of the citizens and scientists no longer need to be tightly coordinated. Indeed, this category has some of the largest and best-known citizen science projects in existence such as GalaxyZoo and EyeWire. The Zooniverse platform that evolved from the GalaxyZoo program now hosts dozens of projects that require the classification or interpretation of digital objects collected by scientists.

For SCDO projects, scientists are no longer nearby to review the data classification. In fact, the scale of the project may prevent systematic review because the large classification task that scientists alone were unable to complete is what motivated the use of the citizen science approach in the first place. The digital nature of the project allows scientists to engage a much larger audience and allows multiple people to complete the same task. Scientists can verify the abilities of participants by asking them to classify objects that have been previously classified by experts. If the results from participants disagree, then software can increase the number of replications to get a statistically confident classification, define the object as unclassifiable, or flag the results for review by experts. Hybrid models have arisen in recent years because of the rapid advances in the success of deep learning algorithms.

In the fifth type, citizens help scientists find and collect physical objects (CPO) at temporal and spatial scales that cannot be achieved through other methods. The objects are typically submitted to a science team for further analysis and archiving. Data quality issues may arise regarding sampling location and time or the collection and processing procedures. Scientists can address data quality issues by making citizens provide information about the collecting event or submit duplicate samples.

The sixth category is the digital equivalent of the fifth category. Citizens collect digital objects (CDO) instead of physical objects. Mobile smartphones, with their internal clocks and GPS units, make it easier to record the time and location for all digital objects collected. The digital record of what the observer saw may bolster data quality. The advantage of this category is that electronic samples can be easily shared, thereby allowing multiple people to classify and review the same observation. Thus, the statistical approaches for data quality used in other types that use digital objects, such as category four, can also be applied to this category.

In the seventh and last category of contribution, citizens report observations (RO), including quantitative measurements, counts, categorical determinations, text descriptions, and metadata. The skill of the participants directly affects data quality because more sophisticated tasks and judgments are required. Because these observations are typically numeric or text data, it is easier to store and collect them than it would be for physical or multimedia objects. The inexpensive recording of these observations via the web makes these projects easy to start and support over the long term.

Data Type and Data Quality Strategies

The different categories of data contribution to CS (**Table 1**) are subject to different types of data quality issues (**Table 2**). When carrying an instrument package or creating new algorithms (CIP, IMA), data quality controls and procedures would be very similar to or the same as in scientific study without citizens. When sorting, characterizing, and categorizing objects (SCPO, SCDO), the objects have already been collected using standard scientific protocols, so their origin and provenance is not in question. If the citizens are working on physical items (SCPO), they are usually working with teams of scientists so when questions arise with a particular item, they can be referred to more experienced team member. Classification of digital objects (SCDO) collected and managed by scientists offers the great advantage that they can be scored by more than one person, which means that statistical techniques can be used to assess data quality and find outliers. The Galaxy Zoo/Zooniverse team has offered several approaches to check data quality (Lintott et al., 2008; Willett et al., 2013).

The collection of specimens for scientific analysis (CPO) seems that it could be very easy if one can accurately record the time, place, and method of collection. In some instances, this can be challenging (Chapman, 2005), and it can be more challenging if the specimens need to be processed in the field. A noted case with a long history of such challenges is the collection of water samples. Here duplicate samples are sometimes used to help ensure data quality, and the US Environmental Protection Agency developed the Quality Assurance Project Plan (QAPP) approach to help bring standard procedures to the process. When people collect digital samples (CDO) (photographs, videos, sound recordings, etc.), there seem to be fewer concerns because collecting digital objects has become so much easier with the growth of smartphones. Today's smartphones commonly time-and-place stamp digital objects automatically with high degrees of accuracy and precision. Time and location, outside of the object itself and the collector, are the most valuable pieces of metadata.

The last instrument type (RO) includes the input of data and metadata by humans and is, therefore, the most prone to data quality issues. Because of the large number and varied protocols and requirements of these projects, it is more difficult to make specific comments about data quality. However, using cell phones when recording data is having a large impact because it allows people to record data as they observe using forms based on pick lists that significantly reduce data input errors. Data can then be shared almost immediately because it can be uploaded directly from the cell phone, reducing chances that data will not be shared or that errors will creep in before data is shared.

The Galaxy Zoo project stands out in its ability to measure observer errors and bias (**Table 2**). The high-quality analyses by the Galaxy Zoo project are possible because they have large data sets, a small number of objects to classify, a large number of

TABLE 2 | Characteristics of seven data types related to data quality.

Data category	Format of primary data	Data quality comments	Data quality approaches	Examples of papers about data quality strategies	Concerns about data quality
Carry Instrument packages (CIP) or pilot vehicles that CIPs	Digital files	Citizens may determine location of the instrument and some initial metadata	Calibration before and after deployment. Locality and quality of instruments employed can be ranked. Using time series and other data to check sensors over time	Bell et al. (2013)	Minimal, because the citizen's contribution to each observation is minimal. Sensor placement and sensor aging are issues
Invent or modify algorithms, IMA	Calculation result, Algorithm	The interactive nature of the process controls data quality. Algorithms are usually tested in a standard environment	The openness of the process allows others to see what is happening and duplicate results	None found	Minimal concerns, because the process is self-correcting. Testing, sharing and archiving solutions along the way is important
Sort and classify physical objects, SCPO	Tagging and describing objects	These projects are usually situated in a collection facility such as a museum or as part of scientific team making it easy for citizens and scientists to interact frequently and for citizens to be incorporated into the scientific team	Because they work closely with experts, it is relatively easy for volunteers to be given tasks that are appropriate for their skill level and for any questions about sorting or classification to be answered by experts within a short period of time	Obrecht et al. (1998), Herron et al. (2004)	Minimal, because volunteer's work is closely integrated within a scientific team
Sort and classify digital objects, SCDO	Groupings, lists or tags	Citizens are only responsible for determining what the object is. It is relatively easy to crowdsource the task using the internet to a large number of interested people	Calibrate each citizen with known objects, classification of real and test object by multiple citizens, statistical evaluation of classification by multiple citizens, expert review, use AI to narrow the range of possible choices	Lintott et al. (2008), Hansen et al. (2011), Fortson et al. (2012), Swanson et al. (2016), Willett et al. (2016), Jiménez et al. (2019), Walmsley et al. (2020)	High to low, will depends on the difficult of the classification task, the experience of the participants and the number of experienced participants who view each object
Collect physical objects, CPO	Physical object or sample	Some objects such as pottery chards or a feather are very stable and the interpretation depends on the circumstances of discovery. Other objects such as water or soil samples may also depend critically on the sampling, storage and transport methods	Replicate samples, for lab processing use splits, blanks and standards for water analysis, expert review	Obrecht et al. (1998), Williams (2000), EPA (2002)	High to low, depending on the documentation of the provenance of the object, specific documentation of the sampling, storage and transport methods and examination by experts can resolve questions
Collect digital objects, CDO	Image, video, or sound recordings	Recent technological advances, especially embodied in smart phones, have allowed citizens to readily capture still images, video, and sounds and share them via the internet	Digital objects without accompanying metadata are almost worthless but cell phones or simple digital camera usually record time and place, making it relatively easy for projects to automate collection of the most salient metadata	None found	High to low depending on the contextual data provided; minimal, when digital objects come with the time and location of the observation based on embedded sensors in the recording instrument
Report observations, RO	Text	The observer provides the description of the observation and the data that describe the context of the observation	Pseudo-replication, technical difficult of the history of individual contributors, project specific knowledge, machine review, expert review	Yu et al. (2010, 2012), Kelling et al. (2012, 2015a,b)	High to low, will depends on the skill required for the observation, extent of training of the observer, knowledge about the skill of the observer

Primary data denotes the focus of the project, the what of the study. Participants may often also report the who, when, where, how, and by whom. These supporting data can be essential to the value of the observation.

classifications per object (>30), reference images to test users, and expert reference datasets to compare with participant results. Calibrating projects without repeated measures is more difficult, but the eBird project is making progress by analyzing individuals capabilities based on the total number of birds they see and their cumulative sampling records (Yu et al., 2010, 2012; Kelling et al., 2012, 2015a,b). Program leaders are aware of these issues and have practiced improving data quality approaches (Wiggins et al., 2011), but it is not always clear in papers or on project websites what steps have been taken or corrections made.

DISCUSSION

Data Types

Wiggins et al. (2011) gave an overview of many approaches used in citizen science for data quality and validation. However, the seven types of data contributions defined here indicate a more refined approach is possible (**Table 1**). The lens of data types offers a new dimension to understand DQ and to compare projects. In the following paragraphs, we offer suggestions about what this typing can offer to the discussion of data quality and project design.

Criticism of Data Quality in Citizen Science

As described in the introduction, DQ has been a major concern in CS programs. Scientists and others naturally question DQ because of minimal training and a lack of formal accreditation by citizen participants (Freitag et al., 2016). Our findings of different data types (**Table 1**), however, suggest that CS activities that involve carrying instrument packages or inventing or modify algorithms will not have data quality issues beyond what scientists normally encounter. We also believe that projects that sort and classify physical objects are unlikely to have significant data quality issues because of the close physical presence and access to collection managers and experts during the sorting process. The very nature of a physical collection requires collection infrastructure in the form of museum facilities and collection managers to maintain it.

Our analysis suggests that the general criticism about data quality in CS programs is more of a concern in the four remaining data types (sort and classify digital objects, collect physical objects, collect digital objects, and report observations). For instance, collecting physical objects such as water samples for water quality programs often requires a special collection process to prevent contamination and/or special storage procedures to reduce deterioration of the samples. In the case of reporting observations, there are a wide range of DQ issues that stem from the complexity of procedures and human judgment required of specific programs. Unlike the collection of physical or digital objects or the classification of digital objects, there is no direct way to judge the quality of the observation. One must use pseudo-replication techniques or knowledge about the history of an individual contributor. Scientists and others have leveled general criticism of the DQ of CS programs, but consideration of these different types makes it clear that DQ assurance is closely tied to the type of data being gathered, and thus criticism should be more specific now.

It is important to note that the seven data types discussed above, in themselves, do not constitute an exhaustive list for information sharing within projects. Project organizations may use multiple forms of communication, including personal conversations, telephone calls, websites, email, email servers, blogs, and chat rooms to guide projects and monitor the collection of data. These auxiliary information channels may play a critical role in triangulating on data quality but may not be part of the formal records of the project or linked to the scientific data.

Single Projects May Use More Than One Primary Data Type

It is also important to observe that a single project can include more than one of these basic instrument types. For instance, OpenStreetMap participants can collect data by using a hand (RO), a GPS unit, and more advanced instruments (CDO). They use these data and data from satellites to map additions, corrections, and annotations onto the OpenStreetMap map layers (SCDO) (OpenStreetMap Wiki, 2016). COASST has an extensive protocol to monitor seabirds that includes observation data (RO) but can also include submitting photographs (CDO) and dead birds for archiving (CPO) (Parrish et al., 2018b). eBird was initially designed to collect text reports of people's observations (RO) but since 2015 also supports submissions of digital recordings of sounds, images, and videos (CDO) (Weber, 2019). iNaturalist combines the collection of digital objects (CDO), and the classification of digital objects (SCDO), with the possibility to simply report observations (RO) (Saari, 2021).

Data Types Are Not Unique to a Scientific Discipline

Different projects within a science discipline may use different types of citizen science data to advance their research. For instance, BatME has recruited citizens to collect audio recordings of bat calls (CDO), while Bat Detective uses citizens to classify bat calls (SCDO). Marshall et al. (2014) give an overview of the multiple ways that citizens contribute to astronomy, focusing on the original observations of amateurs (RO) and the contributions and classification of digital images (CDO & SCDO). St. Fleur (2016) reported that citizens are working with scientists to collect meteors (SCPO). One way for citizen science projects to grow within a scientific discipline would be to develop projects that contribute classes of data that have not been applied to that discipline before. For example, in astronomy, scientists and citizens could work together to catalog meteors and micrometeorites (CPO), or perhaps astronomers would include instrument packages to SpaceX launches (CIP).

Data Types and Implications for Project Design

What is the implication of these data categories for the design of citizen science projects? One obvious answer is that data categories will define the requirements for handling data for a project. This suggests that a single software platform dedicated to one instrument type could serve the needs of other projects

that share the same data type and accelerate the growth of similar citizen science programs.

The clearest example of reusing project software is for the classification of digital objects in which the Galaxy Zoo project has been generalized into the Zooniverse platform. Zooniverse is designed to be readily customized, and it now supports the classification of digital objects from many domains. An example of the lateral transfer of citizen science approaches is the adoption by the eButterfly platform of the eBird sampling protocols (Kelling, personal communication). eBird is an example of a general text collection instrument, but it was designed specifically for bird biodiversity surveys. It is likely that the eBird structure could be generalized for biodiversity surveys of other taxonomies but not other citizen science tasks. A number of efforts, including Anecdata, ArcCollector, BioCollect, CitSci.org, Cybertraker, EpiCollect, FieldScope, GIS Cloud, and OpenDataKit, were built with the goal of allowing people to customize the software for specific field projects. These platforms have been used for numerous projects that collect text and images, but it seems unlikely they would be a good choice to support other instrument types we have outlined.

A general strategy for improving data quality in field collection is to check for errors as early in the process as possible. Specific strategies include (1) requiring users to choose from pick lists rather than using free form input fields, (2) using electronic input via mobile devices (3) checking input immediately from users to give feedback if values seem questionable given the context of the situation (4) taking input such as time and location from sensors when possible, etc.

Another widely accepted approach for data quality is provenance tracking. iNaturalist keeps track of the history of identification for its observations and CoCoRaHS keeps track of instances in which original observations are updated.

Sorting and classifying and/or finding and archiving physical objects (SCPO, CPO) requires a sophisticated infrastructure to manage the objects. Although they may exist, we are not aware of any examples of citizen science platforms that specialize in helping citizens find and archive or sort and classify physical samples most likely because collection management software tools are common in science and largely domain-specific. Instead, citizen science programs would be likely to adapt to interface with established collections software such as Specify (Specify Collections Consortium, 2020), which is used in natural history collections. The scale of these projects is currently bound by citizen proximity to the collection and the space that is needed for work. Sorting and classifying or finding and archiving digital objects (SCDO, CDO) are much more scalable than projects based on physical objects because citizens can be recruited from a larger pool, and expert involvement is not required to assert data quality.

CONCLUSION

This review of the literature and program websites identified seven primary types used in CS programs (**Table 1**). We conclude that blanket criticism of the CS data is no longer appropriate because data types vary widely in their requirements for DQ needs (**Table 2**). DQ is not needed in the invention or modification of algorithms type because DQ is inherent in the process while plans from a variety of approaches are needed and being employed.

Ultimately citizen science has been practiced in a societal context in which there are tradeoffs with DQ (Anhalt-Depies et al., 2019), but at the moment, we believe that significant progress can be made with a simple focus on DQ. We conclude that discussions about the data types in a program and the specific methods being used for DQ control as dictated or appropriate for the type will be fruitful. Information scientists, domain scientists as well as program designers and managers can use the data types as a lens to compare DQ practices and DQ issues across domains. The seven primary data-type lenses can reduce doubts about DQ for funders, participants, and third party data consumers and help managers be more explicit in communicating their data management practices.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

The authors thank the DataONE project for financial support to work as a team through NSF-0830944 and the PPSR committee, on which RDS served, for their insights about citizen science.

REFERENCES

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., et al. (2017). The accuracy of citizen science data: a quantitative review. *Bull. Ecol. Soc. Am.* 98, 278–290. doi: 10.1002/bes2.1336

Alabri, A., and Hunter, J. (2010). "Enhancing the quality and trust of citizen science data," in *E-Science (e-Science), 2010 IEEE Sixth International Conference* (Brisbane, QLD: IEEE), 81–88. doi: 10.1109/eScience.2010.33

Anhalt-Depies, C., Stenglein, J. L., Zuckerberg, B., Townsend, P. M., and Rissman, A. R. (2019). Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. *Biol. Conserv.* 238:108195. doi: 10.1016/j.biocon.2019.108195

Bégin, D., Devillers, R., and Roche, S. (2013). "Assessing volunteered geographic information (VGI) quality based on contributors' mapping behaviours," in *8th International Symposium on Spatial Data Quality* (Hong Kong), 149–154. doi: 10.5194/isprsarchives-XL-2-W1-149-2013

Bell, S., Cornford, D., and Bastin, L. (2013). The state of automated amateur weather observations. *Weather* 68, 36–41. doi: 10.1002/wea.1980

Bernard, A. T. F., Götz, A., Kerwath, S. E., and Wilke, C. G. (2013). Observer bias and detection probability in underwater visual census of fish assemblages measured with independent double-observers. *J. Exp. Mar. Biol. Ecol.* 443, 75–84. doi: 10.1016/j.jembe.2013.02.039

Blake, C., Rhanor, A., and Pajic, C. (2020). The demographics of citizen science participation and its implications for data quality and environmental justice. *Citizen Sci. Theory Pract.* 5:21. doi: 10.5334/cstp.320

Bonney, R., Cooper, C., and Ballard, H. (2016). The theory and practice of citizen science: launching a new journal. *Citizen Sci. Theory Pract.* 1:1. doi: 10.5334/cstp.65

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., et al. (2014). Citizen science. next steps for citizen science. *Science* 343, 1436–37. doi: 10.1126/science.1251554

Burgess, H. K., DeBey, L. B., Froehlich, H. E., Schmidt, N., Theobald, E. J., Ettinger, A. K., et al. (2017). The science of citizen science: exploring barriers to use as a primary research tool. *Biol. Conserv.* 208, 113–120. doi: 10.1016/j.biocon.2016.05.014

Chapman, A. D. (2005). *Principles and Methods of Data Cleaning. Report for the Global Biodiversity Information Facility 2004*. Copenhagen: GBIF.

Cohn, J. P. (2008). Citizen science: can volunteers do real research? *Bioscience* 58, 192–197. doi: 10.1641/B580303

Cooper, C. (2016). *Citizen Science: How Ordinary People Are Changing the Face of Discovery*. New York, NY: The Overlook Press.

Crowston, K., and Prestopnik, N. R. (2013). "Motivation and data quality in a citizen science game: a design science evaluation," in *Proceedings of the Annual Hawaii International Conference on System Sciences* (Wailea, HI), 450–59. doi: 10.1109/HICSS.2013.413

Eitzel, M. V., Cappadonna, J. L., Santos-Lang, C., Duerr, R. E., Virapongse, A., West, S. E., et al. (2017). *Citizen Science Terminology Matters: Exploring Key Terms*. Gfzpublic.Gfz-Potsdam.De. doi: 10.5334/cstp.96

EPA (2002). "Guidance for quality assurance project plans," in *Guidance for Quality Assurance Project Plans*, Vol. QA/G-5, EPA/240/R-02/009. Available online at: https://www.epa.gov/sites/production/files/2015-06/documents/g5-final.pdf

Fortson, L., Masters, K., Nichol, R., Edmondson, E. M., Lintott, C., Raddick, J., et al. (2012). "Galaxy Zoo," in *Advances in Machine Learning and Data Mining for Astronomy*, eds M. J. Way, J. D. Scargle, K. M. Ali, and A. N. Srivastava (CRC Press), 213–236.

Freitag, A., Meyer, R., and Whiteman, L. (2016). Strategies employed by citizen science programs to increase the credibility of their data. *Citizen Sci. Theory Pract.* 1:2. doi: 10.5334/cstp.91

Goodchild, M. F., and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spat. Stat.* 1, 110–120. doi: 10.1016/j.spasta.2012.03.002

Haklay, M. (2013). "Citizen science and volunteered geographic information: overview and typology of participation," in *Crowdsourcing Geographic Knowledge* (Berlin: Crowdsourcing Geographic Knowledge; Springer), 105–22. doi: 10.1007/978-94-007-4587-2_7

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How many volunteers does it take to map an area well? The validity of linus' law to volunteered geographic information. *Cartogr. J.* 47, 315–322. doi: 10.1179/000870410X12911304958827

Hansen, D. L., Jacobs, D. W., Lewis, D., Biswas, A., Preece, J., Rotman, D., et al. (2011). "Odd leaf out: improving visual recognition with games," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom).* 87–94. doi: 10.1109/PASSAT/SocialCom.2011.225

Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., and Bonn, A. (2018). *The European Citizen Science Landscape-a Snapshot*. London: Citizen Science; Innovation in Open Science, Society and Policy. doi: 10.2307/j.ctv550cf2

Herron, E., Green, L., Stepenuck, K., and Addy, K. (2004). *Building Credibility: Quality Assurance and Quality Control for Volunteer Monitoring Programs*. South Kingstown, RI: University of Rhode Island; University of Wisconsin.

Hunter, J., Alabri, A., and Ingen, C. (2013). Assessing the quality and trustworthiness of citizen science data. *Concurr. Comput. Pract. Exp.* 25, 454–466. doi: 10.1002/cpe.2923

Jiménez, M., Triguero, I., and John, R. (2019). Handling uncertainty in citizen science data: towards an improved amateur-based large-scale classification. *Inf. Sci.* 479, 301–320. doi: 10.1016/j.ins.2018.12.011

Johnston, A., Fink, D., Hochachka, W. M., and Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods Ecol. Evol.* 9, 88–97. doi: 10.1111/2041-210X.12838

Kelling, S., Fink, D., La Sorte, F. A., Johnston, A., Bruns, N. E., and Hochachka, W. M. (2015a). Taking a 'Big Data' approach to data quality in a citizen science project. *Ambio* 44, 601–611. doi: 10.1007/s13280-015-0710-4

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W. K., Yu, J., et al. (2012). A human/computer learning network to improve biodiversity conservation and research. *AI Magazine* 34:10. doi: 10.1609/aimag.v34i1.2431

Kelling, S., Johnston, A., Hochachka, W. M., Iliff, M., Fink, D., Gerbracht, J., et al. (2015b). Can observation skills of citizen scientists be estimated using species accumulation curves? *PloS ONE* 10:e0139600. doi: 10.1371/journal.pone.0139600

Kosmala, M., Wiggins, A., Swanson, A., and Simmons, B. (2016). Assessing data quality in citizen science. *Front. Ecol. Environ.* 14, 551–560. doi: 10.1002/fee.1436

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., et al. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Month. Notices R. Astronom. Soc.* 389, 1179–1189. doi: 10.1111/j.1365-2966.2008.13689.x

López, M. P., Soekijad, M., Berends, H., and Huysman, M. (2020). A knowledge perspective on quality in complex citizen science. *Citizen Sci. Theory Pract.* 5:15. doi: 10.5334/cstp.250

Lukyanenko, R., Parsons, J., and Wiersma, Y. F. (2016). Emerging problems of data quality in citizen science. *Conserv. Biol.* 30, 447–449. doi: 10.1111/cobi.12706

Marshall, P. J., Lintott, C. J., and Fletcher, L. N. (2014). Ideas for citizen science in astronomy. *Annu. Rev. Astron. Astrophys.* 53, 247–278. doi: 10.1146/annurev-astro-081913-035959

Miller-Rushing, A., Primack, R., and Bonney, R. (2012). The history of public participation in ecological research. *Front. Ecol. Environ.* 10, 285–290. doi: 10.1890/110278

Muenich, R. L., Peel, S., Bowling, L. C., Haas, M. H., Turco, R. F., Frankenberger, J. R., et al. (2016). The wabash sampling blitz: a study on the effectiveness of citizen science. *Citizen Sci. Theory Pract.* 1:3. doi: 10.5334/cstp.1

Obrecht, D. V., Milanick, M., Perkins, B. D., Ready, D., and Jones, J. R. (1998). Evaluation of data generated from lake samples collected by volunteers. *Lake Reserv. Manag.* 14, 21–27. doi: 10.1080/07438149809354106

OpenStreetMap Wiki (2016). *Category:Data Collection Technique*. Available online at: https://wiki.openstreetmap.org/wiki/Category:Data_collection_technique (accessed December 15, 2020).

Parrish, J. K., Burgess, H., Weltzin, J. F., Fortson, L., Wiggins, A., and Simmons, B. (2018a). Exposing the science in citizen science: fitness to purpose and intentional design. *Integr. Comp. Biol.* 58, 150–160. doi: 10.1093/icb/icy032

Parrish, J. K., Jones, T., Burgess, H. K., He, Y., Fortson, L., and Cavalier, D. (2019). Hoping for optimality or designing for inclusion: persistence, learning, and the social network of citizen science. *Proc. Natl. Acad. Sci. U.S.A.* 116, 1894–1901. doi: 10.1073/pnas.1807186115

Parrish, J. K., Litle, K., Dolliver, J., Hass, T., Burgess, H. K., Frost, E., et al. (2018b). "Defining the baseline and tracking change in seabird populations," in *Citizen Science for Coastal and Marine Conservation* (London: Routledge), 19–38. doi: 10.4324/9781315638966-2

Peters, M. A., Eames, C., and Hamilton, D. (2015). The use and value of citizen science data in New Zealand. *J. R. Soc. N. Z.* 45, 151–160. doi: 10.1080/03036758.2015.1051549

Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D., and Roy, H. E. (2017). The diversity and evolution of ecological and environmental citizen science. *PLOS ONE* 12:e0172579. doi: 10.1371/journal.pone.0172579

Riesch, H., and Potter, C. (2014). Citizen science as seen by scientists: methodological, epistemological and ethical dimensions. *Public Understand. Sci.* 23, 107–120. doi: 10.1177/0963662513497324

Saari, C. (2021). *Getting Started INaturalist*. Available online at: https://www.inaturalist.org/pages/getting+started (accessed December 11, 2020).

Sheppard, S. A., and Terveen, L. (2011). "Quality is a verb: the operationalization of data quality in a citizen science community," in *WikiSym 2011 Conference Proceedings - 7th Annual International Symposium on Wikis and Open Collaboration* (New York, NY: ACM Press), 29–38. doi: 10.1145/2038558.2038565

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., et al. (2012). Public participation in scientific research: a framework for deliberate design. *Ecol. Soc.* 17:29. doi: 10.5751/ES-04705-170229

Specht, H., and Lewandowski, E. (2018). Biased assumptions and oversimplifications in evaluations of citizen science data quality. *Bull. Ecol. Soc. Am.* 99, 251–256. doi: 10.1002/bes2.1388

Specify Collections Consortium (2020). *Software for Biological Collections and Samples*. Retrieved from: https://www.specifysoftware.org/ (accessed May 18, 2021).

St. Fleur, N. (2016). How an Amateur Meteorite Hunter Tracked Down a Fireball. *NY Times*. Available online at: http://www.nytimes.com/2016/03/11/science/

how-an-amateur-meteorite-hunter-tracked-down-a-fireball.html. (accessed December 15, 2020).

Swanson, A., Kosmala, M., Lintott, C., and Packer, C. (2016). A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conserv. Biol.* 30, 520–531. doi: 10.1111/cobi.12695

Thiel, M., Angel Penna-Díaz, M., Guillermo, L. J., Sonia, S., Javier, S., and Wolfgang, S. (2014). Citizen scientists and volunteer participants, their contributions and their projection for the future. *Oceanogr. Mar. Biol. Annu. Rev.* 52, 257–314. doi: 10.1201/b17143-6

Walmsley, M., Smith, L., Lintott, C., Gal, Y., Bamford, S., Dickinson, H., et al. (2020). Galaxy Zoo: probabilistic morphology through Bayesian CNNs and active learning. *Monthly Notices R. Astron. Soc.* 491, 1554–1515. doi: 10.1093/mnras/stz2816

Weber, D. (2019). *A New Way to Upload and Tag Photos and Sounds - EBird*. Available online at: https://ebird.org/news/a-new-way-to-upload-and-tag-photos-and-sounds (accessed December 1, 2020).

Wiggins, A., Newman, G., Stevenson, R. D., and Crowston, K. (2011). "Mechanisms for data quality and validation in citizen science," in *E-Science Workshops (EScienceW), 2011 IEEE Seventh International Conference* (Stockholm: IEEE), 14–19. doi: 10.1109/eScienceW.2011.27

Willett, K. W., Galloway, M. A., Bamford, S. P., Lintott, C. J., Masters, K. L., Scarlata, C., et al. (2016). Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging. *Monthly Notices R. Astron. Soc.* 464, 4176–4203. doi: 10.1093/mnras/stw2568

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., et al. (2013). Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Month. Notices R. Astronom. Soc.* 435, 2835–2860. doi: 10.1093/mnras/stt1458

Williams, K. F. (2000). "Oregon's volunteer monitoring program," in *Sixth National Volunteer Monitoring Conference: Moving into the Mainstream* (Austin, TX), 62–66.

Yu, J., Kelling, S., Gerbracht, J., and Wong, W. K. (2012). "Automated data verification in a large-scale citizen science project: a case study," In *E-Science (e-Science), 2012 IEEE 8th International Conference* (Chicago, IL: IEEE), 1–8. doi: 10.1109/eScience.2012.6404472

Yu, J., Wong, W. K., and Hutchinson, R. A. (2010). "Modeling experts and novices in citizen science data for species distribution modeling," in *Data Mining (ICDM), 2010 IEEE 10th International Conference* (Sydney, NSW: IEEE), 1157–1162. doi: 10.1109/ICDM.20 10.103

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.