CITIZEN SCIENCE:
THEORY AND PRACTICE

Bowser, A, et al. 2020. Still in Need of Norms: The State of the Data in Citizen Science. *Citizen Science: Theory and Practice*, 5(1): 18, pp. 1–16. DOI: https://doi.org/10.5334/cstp.303

RESEARCH PAPER

Still in Need of Norms: The State of the Data in Citizen Science

Anne Bowser[*], Caren Cooper[†], Alex de Sherbinin[‡], Andrea Wiggins[§], Peter Brenton[‖,¶], Tyng-Ruey Chuang[**], Elaine Faustman[††], Mordechai (Muki) Haklay[‡‡] and Metis Meloche[*]

This article offers an assessment of current data practices in the citizen science, community science, and crowdsourcing communities. We begin by reviewing current trends in scientific data relevant to citizen science before presenting the results of our qualitative research. Following a purposive sampling scheme designed to capture data management practices from a wide range of initiatives through a landscape sampling methodology (Bos et al. 2007), we sampled 36 projects from English-speaking countries. The authors used a semi-structured protocol to interview project proponents (either scientific leads or data managers) to better understand how projects are addressing key aspects of the data lifecycle, reporting results through descriptive statistics and other analyses. Findings suggest that citizen science projects are doing well in terms of data quality assessment and governance, but are sometimes lacking in providing open access to data outputs, documenting data, ensuring interoperability through data standards, or building robust and sustainable infrastructure. Based on this assessment, the paper presents a number of recommendations for the citizen science community related to data quality, data infrastructure, data governance, data documentation, and data access.

Keywords: citizen science; crowdsourcing; data; data management; FAIR; data quality

Introduction

Citizen science refers to a spectrum of activities where scientists and members of the public collaborate in scientific work. While conversations to more concretely define and bound "citizen science" are underway (Eitzel et al. 2017), we consider citizen science inclusive of projects across domains and scales, to include both local, place-based initiatives and broader, crowdsourcing solutions.[1] Though the phrase citizen science entered the vernacular in the mid-1990s (Bonney 1996; Irwin 1995), members of the lay public have been involved in science for centuries. Driving and enabling factors for the current proliferation of activities include the rise of the Internet; increased smartphone penetration along with the spread of other information and communication technologies (ICT); recognition from scientists that involving volunteers can support and augment their work; funder requirements for public engagement or outreach; and the rapid increase in global education (Silvertown 2009; Cooper 2016). Millions of people now contribute to citizen science each year. SciStarter, a United States (US) based directory of citizen science projects and related activities, recorded an average of 30 projects added per month during the course of 2018.

Educators in formal and informal settings introduce citizen science with the goal of enhancing topical knowledge and public understanding of science (Bonney et al. 2016). Scientists in academic institutions incorporate citizen science into their research programs, with bibliometric analysis demonstrating the exponential growth of publications referencing citizen science in recent years (Follett and Strezov 2015). Citizen science is also enjoying increased attention on the policy level, as seen in Europe and the US (Nascimento et al. 2018). Members of professional and public communities engage in diverse citizen science activities for a wide range of reasons. Some seek to advance the research enterprise, for example, by enabling data collection on scales and resolutions not possible through professional activities alone (Cooper et al. 2012). Others seek to bridge the science-society gap by making professional researchers and citizens more accountable to each other (Irwin 1995).

The growth and formalization of citizen science is supported by professional associations based in Australia,

[*] Woodrow Wilson International Center for Scholars, US
[†] North Carolina (NC) State University, US
[‡] Center for International Earth Science Information Network (CIESIN), The Earth Institute, Columbia University, US
[§] University of Nebraska Omaha, US
[‖] Atlas of Living Australia, AU
[¶] CSIRO, AU
[**] Institute of Information Science, Academia-Sinica, TW
[††] Institute for Risk Analysis and Risk Communication (IRARC), School of Public Health, University of Washington, Seattle, WA, US
[‡‡] Extreme Citizen Science (ExCiteS), Department of Geography, UCL, GB
Corresponding author: Anne Bowser (anne.bowser@wilsoncenter.org)

Europe, and the US, as well as emerging associations in Asia, South America, and Africa. These organizations provide convening power, and help collect and distribute best practices on the science of citizen science, including through conferences and a peer-reviewed journal (Storksdieck et al. 2016). As further evidence for global reach, the Citizen Science Global Partnership was launched in collaboration with United Nations Environment Programme as a network-of-networks supporting global coordination and linking citizen science to the UN. Sustainable Development Goals (SDGs). Beyond the establishment of new organizations, existing governments and NGOs are developing resources for their employees, grantees, and partners to conduct citizen science. For example, the US. Federal Government and partners launched the CitizenScience.gov platform in 2016, which included a toolkit, a catalogue of federal citizen science projects, and a community page (Nascimento et al. 2018).

One common theme across these citizen science initiatives is the central importance of data collected or generated by the efforts of volunteers who are not typically from scientific professions. As the common denominator in nearly all citizen science projects, data are the foundation of citizen science: Without proper handling of such data, projects will have limited access. However, the potential to generate knowledge through primary research and the reuse of data, and to inform evidence-based decision-making, will be limited if the field does not further advance norms around high-quality data collection and management. Several researchers have offered case studies of individual citizen science projects that excel at various aspects of data collection, management, and use. These case studies generally document effective practices within a specific project, and sometimes offer more generalized recommendations in areas including avian presence and distribution (Sullivan et al. 2017), marine debris (van der Velde et al. 2017), urban tree inventories (Roman et al. 2016), and invasive species (Crall et al. 2011).

Other researchers have identified and analyzed, for example, data quality practices and fitness-for-use assessments across citizen science initiatives (see for example Specht and Lewandowski 2018; Kelling 2018; Aceves-Bueno et al. 2017; Kosmala et al. 2016; Lukyanenko et al. 2016; Sheppard, Wiggins, and Terveen 2014; Wiggins et al. 2011). Still others delved into issues related to standardized data collection (Higgins et al. 2018; Sturm et al. 2017), data management (Schade et al. 2017; Bastin et al. 2017) or concepts like fitness to purpose or fitness for use (Parrish et al. 2018). But with the exception of Schade et al. (2017), who collected data focused on citizen science data access, standardization, and preservation via an online survey, little published work in the context of citizen science evaluates practices related to the full data lifecycle as defined in **Box 1**.

Box 1: Data Lifecycle and Data Management.

This box provides definitions of different aspects of the data lifecycle and data management. The purpose is to provide a high-level overview for citizen science researchers who may be less familiar with terminology and approaches taken by the research data community.

Data acquisition: Collection, processing, and curation of scientific information. Acquisition can occur through human observation or automated sensors.

Data quality: Quality assurance/quality control (QA/QC) checks taken across the data lifecycle, from acquisition to archiving to dissemination. These include validation, cleaning, and checks for data integrity.

Data infrastructure: Tools and technologies including hardware and software that support data collection, management, and access.

Data security: Methods of protecting data from unauthorized access, modification, or destruction through proper system security and staff training.

Data governance: Rules for the control of data including provisions for stewardship, privacy, and ethical use, including ensuring the protection of personally identifiable information (PII).

Data documentation: Discovery metadata (structured descriptive information about data sets used by catalog search tools) and documents describing data inputs and methods used to develop data sets.

Data access: The conditions required for users to find and use data, including metadata and licensing. The research community has variously adopted standards of open access or FAIR (Findable Accessible, Interoperable, and Reusable) data. This includes long-term preservation.

Data services: Tools and web-based applications built with data sets and computer code.

Data integration: The process of combining data from different sources, which requires interoperability enabled through the use of data and service standards.

For additional information on any of these aspects, visit the World Data System training resources page (https://www.icsu-wds.org/services/training-resources-guide) or the ESIP Federation data management training clearinghouse (http://dmtclearinghouse.esipfed.org/).

We sought to advance conversations about the state of the data in citizen science through structured interviews with 36 citizen science projects around the world, representing many scientific domains, and to provide recommendations for improved practice. This research was conducted by citizen science and data experts working under the auspices of the International Science Council Committee on Data (CODATA) and World Data System (WDS). Together, CODATA and WDS formed a task group on citizen science and etc the Validation, Curation, and Management of Crowdsourced Data in 2016. The objectives of the task group were to better understand the ecosystem of data-generating citizen science, scientific crowdsourcing, and volunteered geographic information (VGI) to characterize the potential and challenges of these developments for science as a whole, and data science in particular.

Following this introduction, we review current trends in science and scientific data relevant to citizen science, and then examine current issues around data quality and fitness for use in citizen science. The first contribution of this paper is an exploratory empirical investigation into the state of the data in citizen science. We present our methods and results of the survey of practices before discussing the results. This paper also contributes practical and research-oriented recommendations. As an initial step toward offering concrete guidelines, we identify a list of good data management practices that may be helpful for citizen science projects to consider, particularly if they wish to elevate the value of their data for reuse. We also suggest areas where more research is needed to understand more about our findings, and maximize the impact of this steadily growing field.

Trends in Science and Scientific Data

Shifting norms around open and FAIR

Norms and practices governing data management are still emerging in conventional science, and are not yet firmly established across disciplines. One important development in scientific research is the emergence of open and FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Broadly, open science is research conducted in a way that allows others to collaborate and contribute (OECD 2020). As a movement or paradigm, open science can be traced to the Scientific Revolution of the late 16th and early 17th centuries when rapid dissemination of knowledge became a guiding principle for scientific research (David 2008). Contemporary advocates argue that open science strengthens research by facilitating reproducibility through transparency (Munafò et al. 2017), and makes science more accessible to stakeholders including the general public, though important power differences often remain (Levin and Leonelli 2017). Recently, open science has been accelerated by policy initiatives in Australia, the European Union, the United Kingdom, and the US (Tenopir et al. 2015).

As an umbrella term, open science encompasses a range of components, including participatory research; open access to research publications and pre-prints; open access to data and methodologies, including processes such as lab notes and code; open peer review; and open access dissemination of results and data. Within open science, much of the emphasis to date has been on open data sharing, with a strong focus on licensing. Clear data licensing helps enable open data by clarifying to third party users the status of a data set and their ability to apply the data for different purposes and under different conditions. Common ways to release open data include the Creative Commons Public Domain Dedication (CC0), the Creative Commons Attribution license (CC BY), the Creative Commons Attribution-NonCommercial license (CC-BY-NC), or the Creative Commons Attribution-ShareAlike (CC-BY-SA). These latter licenses include restrictions that can be problematic, an issue we discuss further in Section 5.

A second, related movement is emerging around making data more FAIR. Many of the ideals behind FAIR match the rhetoric around open science; guiding principles include transparency, reproducibility, and reusability (Wiklinson et al. 2016). Calls for open and FAIR data differ on a few key points. First, all FAIR data do not necessarily need to be open. FAIR is about enabling, rather than securing, access to information. Whereas open data are necessarily free of charge, FAIR data could be accessible but behind a paywall. Second, while open science can be described as a paradigm, or an approach to scientific research, FAIR is more prescriptive, offering concrete guidelines and even checklists for researchers to follow (Wilkinson et al. 2016). Practices around cataloguing and metadata documentation help make data FAIR.

The state of the data in scientific research

Understanding the current state of data management is critical for understanding and charting progress moving forward. Notably, the larger scientific community has only recently begun to adopt practices related to open and FAIR data. One benchmark study of 1,329 researchers across scientific domains explored practices and perceptions of data sharing (Tenopir et al. 2011).[2] At the time of publication in 2011, 29% of respondents had data management plans, while 55% did not and 16% were uncertain. Regarding data access, 38.5% of respondents stored their data in an organization-specific system. A follow-up study conducted shortly after National Science Foundation (NSF) policies went into effect reported mixed progress. Perceptions of the value of data sharing increased, but so did perception of threats, and progress on self-reported practices was mixed (Tenopir et al. 2015).

A number of factors contribute to suboptimal data management in scientific research. While researchers are generally satisfied with tools for short-term storage and documentation of their data, access to longer-term repositories may be lacking (Tenopir et al. 2011), and citizen science practitioners may not be familiar with the many domain-specific repositories—though in recent years open repositories such as Dryad and FigShare have grown in popularity. Beyond the provision of technical tools, "Barriers to effective data sharing and preservation are deeply rooted in the practices and culture of the research process as well as the researchers themselves" (Tenopir et al. 2011, p.1). Incentives are often missing for researchers

to invest the time and effort required to make their data open or FAIR, since data cleaning and documentation are time-consuming activities that lack the same incentives as, for example, publication. Further, an academic culture that tethers scholarly publication to professional milestones like the tenure process may actively disincentive openness and sharing if researchers fear getting scooped. And volunteer citizen scientists are not necessarily motivated by the same incentives as researchers, but rather factors such as personal interest, learning, creativity, socialization, and the desire to contribute to scientific research (Jennett et al. 2016; Rotman et al. 2012).

Researchers have also started to study data reuse, defined as the use of data by the original data collector or third-party users, sometimes by combining the data with other data, for the same or different purposes for which they were originally collected. One study found that perceived utility of a data set was the single strongest factor leading to reuse, and concluded that the value of reuse should be more widely demonstrated to the academic community (Curty et al. 2017). Efforts to make data discoverable, promote the use of strong metadata, and improve norms and practices around data attribution and citation could all lead to more data reuse. Regarding citation, the use of persistent identifiers (e.g., Digital Object Identifiers [DOIs]) can ensure that researchers are able to refer to a unique data set produced at a given point in time by providing persistent URLs to data that are retained even if, for example, data moves from a project website to a longer-term repository. This is important for traceability in scientific findings as well as for appropriate attribution.

The state of the data in citizen science

The White House memorandum Addressing Society and Scientific Challenges through Citizen Science and Crowdsourcing (Holdren 2015) offers three core principles for citizen science: Contributions of volunteers should be 1) fully voluntary, 2) meaningful, and 3) acknowledged. Similarly, the European Citizen Science Association (ECSA)'s 10 Principles of Citizen Science (ECSA 2016) include "citizen science project data and metadata are made publicly available and where possible, results are published in an open access format." These codes suggest that data sharing, including through publication, may be necessary to fulfill a core best practice of citizen science. Some researchers document the importance of report-backs, or the process of sharing individual and collective results with volunteers in ways that are meaningful and useful to them (Bonney et al. 2009; Morello-Frosch et al. 2009; Gallo and Waitt 2011). Related, there is often a noticeable commitment within citizen science projects to publish academic publications in open access journals (although fees can be a barrier to follow-through). However, the realities of data sharing may suggest differently: One study of open biodiversity data available through the Global Biodiversity Information Facility (GBIF) found that citizen science datasets were among the least open (Groom et al. 2016).

Beyond open data, a significant portion of research on data practices addresses data quality. The topic of data quality is a key concern in the scientific enterprise because perceptions of poor data quality can influence the willingness of scientists or policy makers to trust the results of citizen science. In the context of a research project, the construct of data quality means that data are high enough quality to serve a project's goals: there are no universal criteria to establish quality in scientific data because it is inherently contextual. In acknowledgement of this reality, the concept of fitness for use is frequently applied in citizen science (Kosmala et al. 2016), with the focus on designing project processes with the end in mind (Parrish et al. 2018). For example, in air-quality monitoring, low-cost sensors cannot currently compete with professional instruments for achieving precision and accuracy at the levels necessary for regulation (Castell et al. 2017). Therefore, one goal of citizen science air-quality projects may be to get regulators to take notice when systematically collected data indicates a potential problem meriting further investigation. Low-cost (including commercial or open-source/do-it-yourself) sensors are of suitable quality to be fit for this, and often other, purposes.

When used to describe an individual data record, data quality typically refers to the accuracy and precision with which a data value represents a measurable parameter of an entity or phenomenon. At a whole dataset level, data quality refers to all attributes being accurately measured using a standard/common protocol and accurate instrumentation.[3] Higher-quality data accurately and precisely represent reality, whereas low-quality data are a poor or inconsistent representation. Errors in measurement can be random (scattered) or systematic (always wrong or biased in the same direction), and they can arise owing to poor instrumentation (imprecise, poorly calibrated, or old) and operator errors, which usually introduce systematic biases in data. Therefore, measurement accuracy may be affected by several factors, including the training and competence of volunteers; sensitivity, calibration and construction quality of measuring instruments; establishment of a consistent sampling frame; the methods used in taking/determining measurements and their consistency over time and space and across volunteers; and delays between sample collection and measurement (in lab settings). With respect to field-based observational facts such as species occurrence recording, competency and attention to detail by citizen scientists can affect factors such as correct identification, spatial accuracy, precision and uncertainty, and date/time precision. In addition, third-party perceptions of data quality can be affected by whether records have been verified or validated by experts or if there are methods or additional data sets available to cross-validate or even triangulate results.

However, the actual quality of data has significance only in the context of usage. This is a relative concept that relates to fitness for use (Chapman 2005), i.e., for some applications, low-quality data may be acceptable. One of the underlying premises of citizen science in the field of biology, for example, is that scores of amateur scientists can collect data over much larger areas and longer periods than would ever be possible by highly trained biologists

alone. Thus, in some studies, the lower quality is balanced by a far wider scope, demonstrating that almost all data has value depending on the purpose for which it is to be used. In addition, citizen science data may be analyzed along with other scientific or instrumental observations as a method of either validating or cross-validating the data, or complementing data of known quality with a larger sample size.

Researchers typically consider data quality and fitness for use in individual project design, explaining the factors affecting data quality within the text of research papers. However, such explanations are not always documented in metadata accompanying the primary raw and processed datasets used in the research, and if they are documented, it is rarely in structured, standardized formats. These cases both create a number of significant problems and constraints for secondary users of the primary data.

For a variety of reasons, researchers are increasingly turning to individual and aggregated datasets collected by other projects as primary or secondary data (e.g., to augment their own original datasets) for their research. These secondary applications of data are highly dependent on researchers having a clear understanding of the provenance, methods, data-quality constraints, and prior treatments of datasets in order to support decisions about fitness for use in their particular application of the data. For secondary users of data to be able to assess fitness for use, they must be able to efficiently filter, sort, and select particular datasets that satisfy the quality criteria for their purpose. To accomplish this, it is critical for dataset metadata to describe the quality aspects of the data as comprehensively as possible, including its provenance, treatment, constraints, and biases, in a structured, standardized way. Our results indicate that currently, well-documented data are not always the norm in citizen science.

In summary, while the citizen science community may lag slightly behind ideals, this is probably in part because of the rapid evolution of scientific norms of open data, data publication, metadata, data documentation, and data reuse over the past decade, which in fact means that many corners of the global scientific enterprise are rushing to catch up. We turn now to our methods and results, before turning to a discussion of what the evolution of norms means for the citizen science community and how the community can improve its data practices.

Methods

The level of detail we sought about data management practices was rarely conveyed on project websites. Therefore, to better understand the state of the data in citizen science, we conducted structured interviews with project managers or key personnel working on the data management aspects of 36 citizen science projects (see Appendix A for a full list).

Sampling framework

Members of the Task Group began by reviewing a range of literature on citizen science data practices across the data lifecycle to inform development of study methods. We reviewed citizen science typologies and other classification schemes to create the sampling framework. Typologies were largely drawn from academic research, and covered aspects of citizen science including governance model (Haklay 2013; Shirk et al. 2012) and scientific research discipline (Kullenberg and Kasperowski 2016; Follett and Strezov 2015). Other classification schemes included UN regions for capturing geographic distribution, and controlled vocabularies used to document variables including type of hosting organization (e.g., university, community-based group, etc.).

The sampling framework allowed us to search for projects representing different types of diversity (e.g., in governance, in scientific research discipline, and in geographic distribution). Using this framework, we recruited participants through a three-step process. Our participants were recruited following a purposive sampling scheme designed to capture data management practices from a wide range of initiatives through a landscape sampling methodology (Bos et al. 2007).[4] First, we pulled a random sample of citizen science projects from the SciStarter database, requesting the listed contact for each project to participate in our study. In this initial sample, we found that projects in environmental citizen science, particularly biodiversity, and projects based in the US were over-represented.

We then used our sampling framework to identify gaps in the sample and sought out projects not necessarily listed on SciStarter to fill the gaps. As gaps were filled, the research team met numerous times to discuss our evolving sample and early findings. The research team then conducted additional purposive sampling until theoretical saturation was reached (Weed 2006) at 36 interviews with citizen science projects and platforms. Note that while this sampling strategy appears successful in covering a wide range of citizen science projects, it is not intended to be statistically representative of the field as a whole, and only English-speaking projects were represented.

Data collection

We began our structured interview protocol with questions from our sampling framework. Interview questions addressed various practices related to data quality and data management (see Appendix B for the interview protocol). In addition to supporting our sampling methodology, these questions enabled us to collect valuable information to help characterize our sample. The second part of our interview protocol addressed practices related to data quality and data management. We focused on these practices because our review of the literature suggested that practices related to data quality and data management (as opposed to, for example, data security) may be unique to citizen science compared with other forms of scientific research. Grounding our protocol in the existing literature allowed us to create a structured protocol with multiple choice rather than open-ended questions. For example, rather than asking participants "Where can your data be accessed?" we asked, "Can your data be accessed from: a) Project website; b) Institutional repository; c) Top-

ical or field-based repository; and/or, d) Public sector data repository?".

Regarding data acquisition, we asked our participants to describe the full range of data collection or processing tasks that were used in their citizen science research. For data management (including data quality), we asked about quality assurance/quality control (QA/QC) processes, including those related to data collection but also human aspects such as targeted recruitment or training; instrument control such as the use of a standardized instrument; and, data verification or validation strategies, such as voucher collection (e.g., through a photo or specimen) or expert review. We asked questions on data access, including whether access to analyzed, aggregated, and/or raw data were provided, and how data discovery and dissemination retrieval were supported (if at all). Because they relied on known practices identified through existing literature, the vast majority of our questions were multiple choice, though participants were encouraged to elaborate on their answers or provide additional information.

Members of the research team conducted interviews or surveys, either in person, by phone, by Skype, or by email.[5] Each team member followed the same structured protocol during the interview process, although open-ended questions allowed for the collection of richer detail on selected cases.

Data analysis

Analysis was conducted through tallying responses, comparing responses with previous research, and augmenting structured responses with unstructured comments. We also compared results with prior quantitative assessments of citizen science data practices, including Schade et al. (2017) and Wiggins et al. (2011).

Results

Although we did not structure interviews directly following the data life cycle (**Box 1**), we solicited responses relevant to each step in the data lifecycle, except for data integration. Note that counts often exceed the total sample size because response categories are not mutually exclusive and many citizen science projects selected multiple response options for each item.

Early in our analysis, we found a number of discrepancies between self-reported information and actual practices. For example, our protocol asked project personnel to tell us, "Does the data set or access point include the name of a person to contact with questions?" A number of people we interviewed responded in the affirmative, and even suggested a specific name of their designated data point of contact, but a quick review of that project's digital presence in data catalogues, websites, and/or data repositories suggested that either no contact was given or the email listed was a generic one (e.g., info@projectname.org). In addition, the participants we interviewed, typically the scientific research leads, were not always familiar with the details of how their research was being supported by technological platforms or how their data were being managed. In some cases, an interviewee reached out to a colleague to provide follow-up information on data archiving. But in others, an interviewee offered information that was factually incorrect, for example, suggesting that a project launched with support from iNaturalist did not have the option to apply standardized data licenses when, in actuality, iNaturalist does offer this functionality. Because many of the details we asked about were not directly observable in projects' online presence, we were not able to systematically verify all of the data collected.

This finding informed our analysis and presentation of our results. For example, while our sample was large enough to support descriptive statistics such as tabulations, we believe that the format of statistical analysis implies a certainty and confidence in the findings that is not fully appropriate. A narrative reporting structure more closely aligns with the relatively exploratory nature of this study, and emphasizes the reliance of our methodology on self-report.

Characteristics of sample

The average start year of the projects in our sample was 2011, with the earliest year being 1992 and the most recent being 2017. Our sample was heavily weighted toward the environmental and biological sciences (n = 29, 81%), reflecting the early genesis of citizen science in these communities (Schade et al. 2017), but also included several health-related projects (n = 7, 19%), two VGI initiatives, a general-purpose crowdsourcing initiative, and a technology development project. Most of the projects were hosted in North America (n = 19, 53%). The remaining sample was from Europe (n = 7, 19%), Oceania (n = 7, 19%), Asia (n = 6, 17%), South America (n = 2, 6%), and Africa (n = 1, 3%). Host organizations included nonprofit organizations (n = 14, 39%); academic institutions (n = 12, 33%); government agencies, including federal, state, and tribal (n = 7, 19%); and for-profit companies (n = 3, 8%). Partnerships were plentiful, with ten projects (28%) designating one or more type of organization as host. Our sample included all participation models according to the Haklay (2013) typology, though not evenly. The sample included a majority of participatory science projects (n = 21, 58%), followed by crowdsourcing (n = 13, 36%), distributed intelligence (n=2, 6%), extreme citizen science (n = 2, 6%), and volunteered computing (n = 1, 3%). Several projects reported multiple participation models, for example offering options that included participatory science contributions as well as crowdsourcing tasks. In terms of geographic scope, 11 projects (31%) were global in reach, 11 (31%) were national, six (17%) were tied to a locality such as a city or specific site, five (14%) were regional, and three (8%) involved online-only participation with no geographic component. Most of the projects involved data collection at sites chosen by the contributors, but several involved assignments to work in specific locations.

Data life cycle

Data acquisition

Observational or raw data collection and/or interpretation tasks (e.g., bird watching or monitoring poaching patterns) were by far the most prevalent form of research (n =

27, 75%). Specimen or sample collection (e.g., water samples or animal scat) was also common (n = 13, 36%). Other projects engaged volunteers in cognitive work (e.g., self-reporting of dreams; n = 7, 19%); categorization or classification tasks (e.g., classifying images or labeling points of interest on a map; n = 4, 11%); digitization/transcription (n = 3, 8%); annotation (n = 2, 6%); and specimen analysis (including lab or chemical analysis; n = 2, 6%). Thirteen projects (36%) were classified as having only one general task type, typical of many crowdsourcing, distributed intelligence (Haklay 2013), and contributory-style citizen science projects (Shirk et al. 2012). Twenty-four projects (67%) involved volunteers in multiple research tasks, suggesting participatory science, extreme citizen science (Haklay 2013), collaborative, or co-created (Shirk et al. 2012) models.

Data quality

Interview participants reported a high number of QA/QC mechanisms (**Figure 1**). All projects used at least one QA/QC method, while 34 (94%) used more than one method, and 22 (61%) utilized five methods or more.

First, twenty projects (56%) conducted expert review, and six (17%) leveraged human expertise through crowdsourced review. Additional data validation strategies included voucher collection (n = 9, 25%), algorithmic filtering or review (n = 5, 14%), and replication or calibration across volunteers (n = 4, 11%). Fourteen projects (39%) removed data considered suspect or unreliable, while nine (25%) contacted volunteers to get additional information on questionable data.

Second, projects focused on the human aspects of data quality through training before data collection (n = 25, 69%) and/or on an ongoing basis (n = 11, 31%). Seven projects (19%) used targeted recruiting to find highly qualified volunteers. Four (11%) conducted volunteer testing or skill assessment.

Third, many projects approached data quality through standardizing data collection or analysis processes. Twenty-two (61%) used a standardized protocol. In addition, five (14%) used disciplinary data standards (e.g., Darwin Core for biodiversity data), and five (14%) used cross-domain standards (e.g., of the Open Geospatial Consortium [OGC]).

Fourth, many projects enabled data quality through instrument control. Fourteen (39%) used a standardized instrument for data collection or measurement. Five (14%) reported processes for instrument calibration.

Finally, a handful of projects documented their data quality practices. Seven projects (19%) shared what was classified as other documentation on a project website, while one project in our sample (3%) offered a formal QA/QC plan.

In addition to reporting on practices, many participants spoke at length about their data quality practices. Some indicated that data quality was secured through "very simple protocols and instructions." Upon reflection, one noted that the use of simple protocols led to data collection practices that were "standardized, but not deliberately standardized." Participants also taught us that data quality practices are often rich and contextual. One explained how data were vetted according to a six-pronged approach, where all published observations must be specific, complete, and appropriate ("the content is professional and for the purpose of education, not political or to further personal agendas"). A second described how traditional data quality metrics, such as temporal accuracy, were less relevant to their work than the ability to offer a detailed reporting of a phenomenon of interest.

Data infrastructure

Our survey did not delve heavily into infrastructure for two primary reasons. First, the topic of infrastructure did not emerge as significantly as other topics in our initial

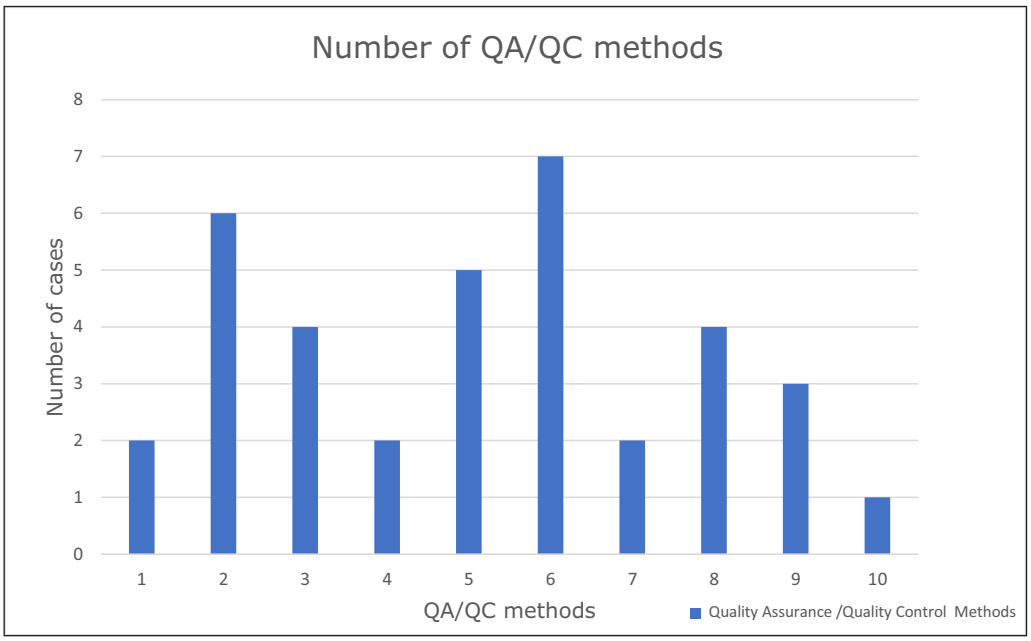

Figure 1: Number of quality assurance/quality control (QA/QC) methods per project.

literature review and scoping process. Second, during the interview process, many of the project principals we interviewed were not very familiar with the back end infrastructure supporting their projects. With this in mind, we noted that many projects adopted existing data collection applications and online communities, such as iNaturalist, BioCollect, CitSci.org, and Spotteron; leveraged existing crowdsourcing platforms, such as Zooniverse or Open Street Map; or developed their own fit-for-purpose platform with robust infrastructure, often including backups and redundancies. Smaller projects may rely on a volunteer technician to manage the data infrastructure, and here the lack of familiarity with the details of the IT backend of projects on the part of project managers may suggest some underlying fragilities.

Data security

For reasons similar to those offered above, the team did not pose specific questions related to the security of the systems used to store data (e.g., passwords, encryption, or two factor authentication), nor did we examine provisions for long-term data stewardship (e.g., archiving in trusted digital repositories). As with issues around data infrastructure, it is likely that citizen science projects vary in maturity levels with regards to adherence to standard security protocols, from relatively weak to very robust. And as with larger findings on data infrastructure, we noted that many citizen science project leaders struggled to articulate specifics around data security approaches when the topic arose organically through the interview process. Finally, while no projects reported data losses or breaches, it is conceivable that these may have occurred as a result of piecemeal approaches to infrastructure, which itself may reveal the often limited funding available for more robust approaches.

Data governance

In citizen science and other scientific research, sensitive data are often obscured. For the purpose of this study, data sensitivity addressed both data or information about citizen scientists or crowdsourcing volunteers who are contributing to research, and sensitive data that is collected by a citizen science community (Bowser et al. 2014; Bowser and Wiggins, 2015.) Twelve projects (33%) specified that they removed or anonymized personally identifiable information (PII), five projects (14%) obscured location information (typically for sensitive species), four projects (11%) reported obscuring other confidential information, and one specifically did not record individual-level information in the first place. One project had a social networking model, whereby only members could view identifying information about other members and the observations they had made: in essence members opted in and volunteered information about themselves. Six projects (17%) that made their data openly available deliberately avoided any obscuring, with one noting that an informed consent process was used to make sure participants understood and were comfortable with what was shared.

Data documentation

Regarding accompanying documentation about data collection activities, 13 (36%) included information on environmental conditions (e.g., weather, location details), 11 (31%) identified the methodology or protocol for data collection, three (8%) provided information about volunteers including characteristics or training levels, and, two (6%) included equipment details or device settings. In addition, eight projects (22%) included multiple pieces of information from the foregoing categories, most commonly environmental conditions and protocol details. Only twelve (33%) provided no additional information whatsoever. Participants were also asked a series of questions about documentation of the research study. Thirteen (36%) mentioned publishing information about the methodology or protocol, while eight (22%) documented limitations. Five projects (14%) offered fitness-for-use statements or use cases. Sometimes these were simply disclaimers, such as "data is provided as is." Participants also identified information on different types of documentation that might be helpful in fitness-for-use assessments, including whether a designated contact was available to answer additional questions.

Data access

Questions on data discovery were designed to probe whether potential users could find information on the project or data. Questions on access covered raw data, analyzed or aggregated data, and digital data services.

Eighteen projects made their data discoverable through the project website. Ten projects (28%) made data available through a topical or field-based repository (such as GBIF). Further, eight projects (22%) shared their data through an institutional repository, four (11%) through a public sector data repository, and two (6%) through a publication-based repository. Only nine projects (25%) did not easily enable secondary users to find their data. Notably, some projects' data were known to be redistributed by third parties, but interviewees were unable to specify the full range of discovery and access points (at least three projects).

Access to cleaned, aggregated data was mixed. Fourteen projects (39%) published open data, defined as "available for human or machine download without restriction." Thirteen (36%) offered data upon request, including by emailing the principal investigator (PI). Interestingly, one of the projects that made data available on request had actually developed a sophisticated data dashboard and gave permissions to 15 local government agencies, not advertising this because they lacked the capacity to handle more subscribers. Six projects (17%) published open data, but required processes like creating user accounts that effectively prohibited automated access. Seven projects (19%) stated that their data were never available, though one respondent commented that access "varies," and another indicated that data were available "only to project partners." An additional interviewee noted that "my priority is to publish first the results, and then I want to look for the ways that are in place to open those data as well."

Participants were asked about their use of a persistent and unique identifier, such as a GUID (globally unique identifier) or DOI (digital object identifier), and their use of a standardized data license. Eleven projects (31%) offered a persistent and unique identifier to support reuse and citation; the other 26 (72%) either did not offer one, or participants did not know. Only 16 projects (44%) had a standardized license to support data reuse. For those projects licensing their data, Creative Commons licenses were the most common. CC-BY and CC-BY-SA licenses, which require attribution, were most frequently adopted (n = 8, 22%), with five projects embracing CC0 public domain dedication (14%) and three projects (8%) using another license, such as CC BY-NC or CC BY-NC-SA, that prohibited commercial use. Beyond CC licenses, three projects (8%) reported holding or co-owning copyright, one project (3%) reported using an Open Database License (ODbL), one project (3%) reported another unnamed license. However, 18 participants **(**50%) did not identify any standardized license for their data, and two participants (6%) didn't know whether their project had a license or not. Numerous participants provided commentary. Some suggested that licensing was the responsibility of another team member. Others indicated a general desire to "keep it open access" or believed that even if a standardized license was not used, "the site has a FAQ that somehow addresses these questions." Notably, data provided without a license or explicit terms of use cannot really be considered open data, an important detail discussed in greater depth later on.

Projects were typically open to inquiries about their data: Twenty-six projects (72%) provided some form of contact information for data inquiries, although seven (19%) had a general project contact but no data-specific contact person, and eight (22%) provided no contact details at all.

Data services

Access to analyzed (cleaned, aggregated, summarized, visualized) data were provided in a variety of forms. Nineteen projects (53%) shared findings through project publications or whitepapers, while 16 (44%) shared findings through peer-reviewed publications. Many projects noted that scholarly publication was "a longer-term goal." Only six projects (17%) provided no access to analyzed data. Many projects used other mechanisms for sharing, some of which were specific to the audiences they served. For example, one project offered a dashboard for State government agencies with explicit partnership agreements to access data, but did not make this service available to others.

Twenty-three projects (64%) offered digital data services. Of these, 16 (44%) provided tools for user-specified queries or downloads (with several also providing application programming interfaces [APIs] for machine queries), 14 (39%) made data available through web services or data visualizations, including maps, 10 (28%) offered bulk download options, and 5 (14%) provided custom analyses or services. In addition, 1 project was willing to provide data "on request." However, 14 projects (39%) provided no specific tools for accessing data resources.

Discussion

Adoption of Best Practices

We found projects were generally implementing best practices with regard to data quality (as described by Wiggins et al. 2011), but were not implementing, and generally not aware of, best practices with regard to aspects of data management such as data documentation, discovery, and access.

In regard to data quality, we were encouraged to see the wide range of practices that projects employed. The majority of our sample (34 projects, 94%) used more than one method to ensure data quality, and 20 projects (56%) used five methods or more. That said, many could only articulate data quality methods when prompted, and only one had a systematic documentation of QA/QC through a formal plan. This suggests that, contrary to some external skepticism (e.g., Nature 2015), the issue with citizen science and data quality is not in actual practices, but with the documentation—or lack thereof—to describe the care and consideration taken with QA/QC.

Many projects demonstrated willingness to make their data available, for example by suggesting that data would be shared upon request. But we found that such de facto attitude to open access was not always backed by the appropriate licensing required to establish the legal (and ethical) conditions required for reuse, nor was provision of access in formats accessible to human and machine users alike a dominant practice. This finding supports prior research conducted within the field of biodiversity, which found that out of different types of data hosted in GBIF, citizen science data were among the worst documented and most restrictive (e.g., especially by prohibiting commercial reuse; Groom, Weatherdon, and Geijzendorffer 2016). While seemingly egalitarian, progressive, and in keeping with the community ethos of some citizen science initiatives, the restriction on commercial uses or the inappropriate application of share-alike licenses[6] can prevent third parties from providing value-added data and services based on raw data, and may stymie private sector research and innovation that could be in keeping with project and participant values. It may also hinder a project's goals; for example, a primary customer of citizen science data for mosquito-vector monitoring could be commercial mosquito control groups. In addition, if citizen science data are enhanced owing to significant investments by companies, they may represent a real value proposition for all data consumers, including citizen scientists themselves. In such cases, CC-BY-NC and CC-BY-SA licenses can be viewed as regressive and not in keeping with open science principles, though the debate is nuanced and open. For example, biodiversity observations shared under an NC license cannot be used on Wikipedia (which supports a broader open data policy) to illustrate articles about species for which citizen science data may be the primary or best available records.

Some citizen science projects implemented best practices in regard to data governance, including access and control. Many solutions, such as location obscuration or masking PII, were designed to protect the privacy of humans and/or sensitive species. Further, at least a handful of the projects that did not leverage these solutions had thought about implications like privacy and made a deliberate decision to prioritize, for example, principles like notice and informed consent (see also Bowser et al. 2017).

In respect to data provenance and traceability, the use of DOIs and appropriately explicit licensing statements is an issue for establishing scientific merits. One respondent indicated that "The data could have been referenced in publications, but we don't know about it," a situation that could be remedied by the use of DOIs. The global biodiversity informatics community has long recognized this issue too, and has made some progress on data archiving (Higgins et al. 2014). As one example, GBIF, together with its partners and members, implemented DOI minting and tracking mechanisms to link publications citing data sources with the original source data, which include datasets sourced from citizen science projects. While users are not required to use DOIs or even to attribute to the referenced data (CC-BY is a "requirement" that may not be enforced), it is becoming an increasingly prevalent practice in the science community. This is a persistent issue related to data citation practices: It's harder to establish impacts for fully open access data. And some practices, such as requiring registration for access to data can help to track usage, but may serve as an impediment for some users (Wiggins et al. 2018).

Finally, when datasets are not adequately described with relevant metadata, their potential for secondary uses is significantly compromised, frequently resulting in whole datasets being discounted as untrustworthy and reinforcing the perceptions of poor rigor in citizen science. Addressing this perception is critically important for citizen science–generated data to gain more trust within the research sector.

Across all aspects of data management, we found a few projects following best practices in every category, but most projects had a mishmash of practices and a clear work-in-progress narrative with respect to evolving practices as project activities progressed. As one respondent commented, "We really want scientists to use the data but we're not at a point where we would recommend that they use the data," and multiple projects reported plans to achieve higher levels of data management for several items we asked about. Further, many respondents, including project managers who had dedicated IT support or leveraged an external platform, often did not know details of their data management practices, as these duties were delegated to others (consistent with Wiggins et al. 2011). In a similar vein, several respondents noted that they had not written their project's data management plans nor designed the technological workflows themselves; these tasks had been outsourced, leaving our respondents unable to fully answer the questions asked.

This was particularly notable for projects whose data access, services, and persistent identifiers were provided by a platform that offered data hosting. While this may be a reasonable option, particularly for smaller or start-up citizen science projects, and whereas taking advantage of the expertise of an interdisciplinary team is often advocated, it can clearly lead to a lack of awareness about data practices, with potential consequences for data strategies. Outsourcing may lead to, or be a sign of, inattention to the importance of decisions made by the data host. This inattentiveness could lead to issues down the road if infrastructure should fail or security be lax. As one respondent explained, each project that was affiliated with the larger program made their own data sharing decisions, but deciding to make data openly available did not mean that that the lead researcher assumed responsibility for depositing the data into an open access database with a persistent identifier. Project managers were not always sure who was responsible to carry out policy-oriented dictates for data management and preservation. While not all data need to be archived, at present probably too little are being proactively preserved for the long term.

In some cases, the adoption of best practices in citizen science data management may be similar to or lagging only slightly behind those of conventional science. For example, we found that in regard to data discovery and access, ten projects (28%) made data available through a topical or field-based repository (such as GBIF), eight (22%) through an institutional repository, four (11%) through a public sector data repository, and two (6%) through a publication-based repository. In comparison, Tenopir et al. (2015) found that 27.5% of the researchers in their survey made their data available through a discipline-based repository, 32.8% through an institutional repository, and 18.4% through a publication-based repository. Comparing these studies suggests that both citizen and conventional science lag far behind the ideal. But the consequences are more significant for citizen science. Widespread adoption of best practices in data management in citizen science would provide much needed transparency about data collection and cleaning practices and could go a long way in advancing the reputation of the field. It could also help satisfy citizen science's commitment to ethical principles, as outlined in the Holdren Memorandum and ECSA's 10 principles of citizen science (Holdren 2015).

While the questions on data management and discovery practices often focused on a scientific user audience, it is important to recall that the scientific research community isn't always the primary audience for a citizen science project: Local communities, students, or other parties may be a target audience, for whom access through a project website is preferable and analyzed products may be preferred over raw data access. However, data access is also reflective of current archival practices and long-term stewardship choices. From this perspective, most of the projects in this study were not positioned to ensure long-term access to data, and in the majority of cases, data sustainability appears tenuous at best.

Infrastructure and technology impacts

Databases, software applications, mobile apps, and other e-infrastructures supporting citizen science have a significant role to play in facilitating improvements in data quality. Such infrastructures can, if they conform to appropriate standards and use good design principles, make the data more discoverable, more accessible, more reusable, more trusted, more interoperable with other systems, more accurate, and less prone to human-induced errors (Brenton et. al. 2018). Good design and open infrastructures enable efficient and simple data recording and management by using workflows, processes, and user-centered design to minimize the risk of user errors and ensure that consistent data formats and mandatory attributes are recorded correctly, along with consistent use of vocabularies, spatial referencing, and dates. At the same time, providing project managers with adequate and easily understood reference information about the default policies that apply to hosted data seemed to be a clear gap for our respondents.

At the global scale, and indeed in many countries, it would be fair to say that the e-infrastructures currently supporting the majority of citizen science projects are largely functioning independently of each other and are not often adequately ascribing metadata to describe the datasets and methods. In addition, very few e-infrastructures are currently implementing any commonly used data standards. This effectively isolates these systems from each other and from being able to share data in ways that can open doors to important new scientific insights through, for example, larger aggregated views and analyses based on spatially and temporally dense datasets.

However, there are examples in some countries where efforts are being made to bridge the e-infrastructure divide. Firstly, the Public Participation in Scientific Research-Core (PPSR-Core) project is an initiative of the citizen science associations (US, European, and Australian Citizen Science Associations) in partnership with the OGC and World Wide Web Consortium (W3C) to develop a set of standards for citizen science data, metadata, and data exchange protocols. Within each of the association regions there are separate third-party platform-based initiatives to support individual citizen science projects (e.g., CitSci.org, Zooniverse, iNaturalist and SciStarter [US]; BioCollect [Australia]; and Spotteron [Europe]). Some of these multi-project platforms are already implementing the PPSR-Core standards as they evolve and are already sharing project-level metadata amongst each other to improve the discoverability of citizen science projects. As a next step, researchers working with Earth Challenge 2020 and the Frontiers open access publication series are creating a metadata repository to facilitate the discovery and access of citizen science data.

Assuming that standards and best practices already exist in an accessible and usable form (which was not universally the case at the time of writing) to apply them in e-infrastructure and data management solutions, providers should codify them into their software to ensure consistency and offer guidance for users, particularly those inexperienced with such matters. However, one interviewee noted that adopting a third-party platform to manage their data did not allow them to direct data management practices because they didn't have control of the technical infrastructure to impose their own field-specific or project-specific preferences. This presents a significant challenge for infrastructure providers, as it suggests that software is expected to be both highly configurable around individual user needs while applying standards, rules, and workflows that assist users to apply best practices in data collection and management. At the extremes, these are diametrically opposed concepts, but it is possible to provide flexible solutions within a standards-constrained environment. Achieving the right balance between flexibility and appropriately structured constraints will require both project owners and infrastructure providers to be aware of standards and best practices, as well as for providers to be transparent as to if or how they are applied in their platforms.

The human dimension

A fundamental rationale for improving data management practices in citizen science is to ensure the ability of citizens, scientists, and policy makers to reuse the data for scientific research or policy purposes. Mayernik (2017) explores how hard and soft incentives can help support open data initiatives. Hard incentives include requirements by funders like the National Science Foundation (NSF) in the USA for researchers to supply data management plans or requirements from publishers that mandate publishing data in conjunction with a research article. Mayernik also uses the concepts of accountability and transparency to explore additional factors that may limit reuse. Transparency includes requirements for making data discoverable and can be charted on a spectrum. For example, providing a link to data online with brief textual descriptions is less transparent than registering data in a catalogue (metadata repository) with standardized descriptions and/or tags.

Culture also has a significant role to play. In line with broader discussions of open science (David 2008; Levin and Leonelli 2017; Munafò et al. 2017), traditional academic cultures often fail to incentivize researchers for good data management to enable reuse. Here, the use of DOIs can be a technical solution that also enables cultural change if researchers can get credit when other researchers are able to find, use, and ultimately cite their data. There is also an opportunity for cultural change specifically within the citizen science community. By evoking aspirational guidelines such as those outlined in the Holdren Memo and ECSA's 10 principles (Hodren 2015), linking good data management practices to already-articulated community values like transparency can create pressure for researchers to make their data more discoverable and accessible as an ethical imperative.

Conclusions and Recommendations

While citizen science has emerged as a promising means to collect data on a massive scale and is maturing in regard to data practices, there is still much progress to be made in approaches to the data lifecycle from acquisition to man-

agement to dissemination. This reflects the speed of development of scientific data management norms and the fact that the scientific community as a whole has difficulty keeping up. However, it may also reflect lack of resources, particularly for smaller or startup citizen science efforts that struggle to maintain staff and funding and perhaps find that data management falls to the bottom of the to-do list. Finally, the fact that many of those who start citizen science projects are motivated primarily by intellectual curiosity, educational goals, environmental justice, or the desire to inform society about significant challenges, may be reflected in project founders who may lack the background in data practices that could carry their work to the next level. The characterization of data practices in this paper is not intended as a criticism of the field, but rather an effort to identify areas where improvements are needed and to provide a call to action and greater maturation. We will have succeeded to the degree that we have educated the citizen science community about emerging practices that can help to improve the usability of their data for not only scientific research but also to solve important societal and environmental problems.

For projects that seek to elevate the value of their data for reuse, we propose a number of steps that could help to increase conformity to data management best practices (**Box 2**).

There are a number of limitations to this research, including the small sample size and the reliance on self-reported information by respondents. Reliance on self-reported information is particularly challenging given the discrepancy between self-reported information and actual practices, as described above.

These discrepancies offer significant opportunities for research and practical work. While the finding that project leaders do not necessarily understand their data management practices offers an important insight, there is a need for clarity regarding what actual practices are most and least common. A follow-up study could compare self-reported with actual practices by, for example, complementing self-report methodologies with desk research, perhaps developing profiles of projects with certain data management practices, or even quantifying the strength of data management approaches. There is a related opportunity to conduct studies of research role differentiation within citizen science projects, and map the different types of expertise, such as scientific, technological, or educational knowledge, represented on a project support team, which may be distributed across a number of departments or institutions.

Our landscape sampling framework sought to identify and characterize a wide range of practices across different types of citizen science projects. Others, including Schade and colleagues, have leveraged different methodologies, such as large-scale surveys, that attempt to gain a more representiative view (2017). Future research could leverage random or purposive sampling to build

Box 2: Recommendations.

This box provides key recommendations for improving data management practices that can be applied across a wide range of citizen science initiatives. Recommendations are offered for individual researchers, and for the field writ large. Additional helpful information may be found in a primer published by DataONE (Wiggins et al. 2013), though more work may be needed to identify an updated set of best practices for broad citizen science communities to use.

Data quality: While significant quality assurance/quality control (QA/QC) checks are taken across the data lifecycle, these are not always documented in a standardized way. Citizen science practitioners should document their QA/QC practices on project websites and/or through formal QA/QC plans. Researchers seeking to advance the field could help develop controlled vocabularies for articulating common data-quality practices that can be included in metadata for data sets and/or observations.

Data infrastructure: Citizen science practitioners should consider leveraging existing infrastructures across the data lifecycle, such as for data collection and data archiving, e.g., in large and stable data aggregation repositories. Researchers seeking to advance the field should fully document supporting infrastructures to make their strengths and limitations transparent and increase their utility, as well as develop additional supporting infrastructures as needed.

Data governance: Relevant considerations include privacy and ethical data use, such as ensuring the protection of sensitive location-based information, personally identifiable information (PII), and proper use of licensing. Citizen science practitioners should carefully consider tradeoffs between openness and privacy. Researchers seeking to advance the field could develop standard data policies, including privacy policies and terms of use, that clearly outline data governance practices.

Data documentation: Citizen science practitioners should make discovery metadata (structured descriptive information about data sets) available through data catalogues, and should share information on methods used to develop data sets on project websites. Researchers seeking to advance the field could develop controlled vocabularies for metadata documentation, particularly to enable fitness for purpose assessments.

Data access: In addition to discovery metadata, citizen science practitioners should select and use one or more open, machine-readable licenses like the Creative Commons licenses. Researchers seeking to advance the field should identify, share information about, and if necessary develop long-term infrastructures for data discovery and preservation.

on these studies and potentially investigate the role of a single variable, such as project governance model, in data management.

Finally, future work could expand across the data lifecycle to focus on such aspects as data infrastructure and data security, or seek to do a direct comparative study between citizen science and research conducted through other means. To the final point, we believe that given the ethical imperatives around good data practices that enable open and FAIR data, citizen science could play a strong leadership role in the broader community of scientific research.

Data Accessibility Statement

Because of the potentially sensitive nature of participant responses, qualitative data are not available for reuse.

Notes

[1] Although citizen science and crowdsourcing differ in some respects, here the authors collectively refer to projects gathering data principally through the engagement of volunteers as citizen science projects.

[2] The date of this study is notable, as 2011 marked the year that the US National Science Foundation (NSF) began mandating that principal investigators (PIs) must include a Data Management Plan as a core component of their proposal. The publication authored by Tenopir and colleagues in 2011, reporting on research activities conducted in 2010, can therefore be helpful as a benchmark for understanding norms before NSF policies took effect.

[3] Accuracy is the degree to which a measurement measures the actual or real value (proximity to reality), and precision is the degree to which measurements of the same parameter real value are close to each other and/or are consistent over time.

[4] Landscape sampling is not a methodology that seeks to produce a sample that fully and comprehensively reflects trends within a population—rather, the goal of landscape sampling is to uncover a wide diversity of practices within a population.

[5] NC State University's Institutional Review Board (IRB) classified this research as not involving human subjects and thus not requiring IRB review.

[6] Share-alike licenses require users of data to contribute to the community any newly developed data or value-added services that build upon the original raw data, with the same license as initially assigned. Depending on the initial license, this may or may not result in derivative products being made available as open and free of charge.

Supplementary Files

The supplementary files for this article can be found as follows:

- **Appendix A.** Citizen Science Projects Reviewed. DOI: https://doi.org/10.5334/cstp.303.s1
- **Appendix B.** Interview Protocol. DOI: https://doi.org/10.5334/cstp.303.s2

Ethics and Consent

NC State University's Institutional Review Board (IRB) classified this research as not involving human subjects and thus not requiring IRB review.

Acknowledgements

Rorie Edwards at WDS was a critical contributor who facilitated Task Group online meetings as well as preparing the initial Task Group proposal that was submitted to CODATA. In addition, Carolynne Hultquist of the Earth Institute at Columbia University provided feedback that helped improve our final manuscript.

Funding Information

The authors would like to acknowledge financial support from CODATA for a consultancy that greatly facilitated this research. Participation from AB and MM was supported by the Alfred P. Sloan Foundation. CC recognizes support from NSF #1835352, *Establishing Norms of Data Ethics in Citizen Science*. ADS recognizes support from NASA contract NNG13HQ04C for the continued operation of the Socioeconomic Data and Applications Center (SEDAC) and PB from the National Collaborative Research Infrastructure Strategy (NCRIS). TRC was funded in part by the Ministry of Science Technology, Taiwan (grant no. 108-2621-M-001-006 and 109-2621-M-001-001) and the Research Center for Information Technology Innovation, Academia Sinica. EF recognizes NIH, EPA, and Nippon Foundation. MH is funded by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 694767, ERC-2015-AdG.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

ADS, CC, and TRC co-chaired the CODATA Task Group and obtained funding. AB led development of the sampling frame and MH interviewed protocol. CC hosted AB on a research sabbatical. AB, CC, ADS, AW, PB, TRC, EMF, and MH, and interviewed project managers. AB, ADS, AW, CC, and PB wrote and edited the initial draft of the manuscript; CC, ADS, AW, PB, TRC, EMF, and MH advised on the research question and study design, and provided edits to the manuscript. MM contributed figures and copious edits.

References

Aceves-Bueno, E, Adeleye, AS, Feraud, M, Huang, Y, Tao, M, Yang, Y and **Anderson, SE.** 2017. The accuracy of citizen science data: a quantitative review. *The Bulletin of the Ecological Society of America*, 98(4): 278–290. DOI: https://doi.org/10.1002/bes2.1336

Bastin, L, Schade, S and **Schill, C.** 2017. Data and metadata management for better VGI reusability. *Citizen Sensor*, 249. DOI: https://doi.org/10.5334/bbf.k

Bonney, R. 1996. Citizen science: a lab tradition. *Living Bird*, 15(4): 7–15.

Bonney, R, Cooper, C, Dickinson, J, Kelling, S, Phillips, T, Rosenberg, KV and **Shirk, J.** 2009. Citizen science:

a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59: 977–984. DOI: https://doi.org/10.1525/bio.2009.59.11.9

Bonney, R, Phillips, T, Ballard, H and **Enck, J.** 2016. Can citizen science enhance public understanding of science? *Public Understanding of Science*, 25(1): 2–16. DOI: https://doi.org/10.1177/0963662515607406

Bos, N, Zimmerman, A, Olson, J, Yew, J, Yerkie, J, Dahl, E and **Olson, G.** 2007. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, 12(2): 652-672. DOI: https://doi.org/10.1111/j.1083-6101.2007.00343.x

Bowser, A, Shilton, K, Preece, J and **Warrick, E.** 2017. Accounting for privacy in citizen science: Ethical research in a context of openness. In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2124–2136. Seattle WA, February 2017. DOI: https://doi.org/10.1145/2998181.2998305

Bowser, A, Wiggins, A, Shanley, L, Preece, J and **Henderson, S.** 2014. Sharing data while protecting privacy in citizen science. *Interactions*, 21(1): 70–73. DOI: https://doi.org/10.1145/2540032

Bowser, A and **Wiggins, A.** 2015. Privacy in participatory research: Advancing policy to support human computation. *Human Computation*, 2(1): 19–44. DOI: https://doi.org/10.15346/hc.v2i1.3

Brenton, P, von Gavel, S, Vogel, E and **Lecoq, ME.** 2018. Technology infrastructure for citizen science. In: Hecker, S, Haklay, M, Bowser, A, Makuch, Z, Vogel, J and Bonn, A (eds). *Citizen Science: Innovation in Open Science, Society and Policy*. London: UCL Press. DOI: https://doi.org/10.14324/111.9781787352339

Castell, N, Dauge, FR, Schneider, P, Vogt, M, Lerner, U, Fishbain, B, Broday, D and **Bartonova, A.** 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99: 293–302. DOI: https://doi.org/10.1016/j.envint.2016.12.007

Chapman, AD. 2005. *Principles of Data Quality, version 1.0*. Copenhagen: Global Biodiversity Information Facility ISBN 87-92020-03-8.

Cooper, C. 2016. *Citizen science: How ordinary people are changing the face of discovery*. New York: The Overlook Press, Peter Mayer Publishers, Inc.

Cooper, CB, Hochachka, WM and **Dhondt, AA.** 2012. The opportunities and challenges of Citizen Science as a tool for ecological research. In: Dickinson, JL and Bonney, R (eds). *Citizen Science: Public Collaboration in Environmental Research*. Ithaca, NY: Cornell University **Press.**

Crall, A, Newman, G, Stohlgren, T, Holfelder, K, Graham, J and **Waller, D.** 2011. Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, 4(6): 1–10. DOI: https://doi.org/10.1111/j.1755-263X.2011.00196.x

Curty, RG, Crowston, K, Specht, A, Grant, BW and **Dalton, ED.** 2017. Attitudes and norms affecting scientists' data reuse. *PloS one*, 12(12): e0189288. DOI: https://doi.org/10.1371/journal.pone.0189288

David, PA. 2008. The Historical Origins of 'Open Science': An essay on patronage, reputation and common agency contracting in the scientific revolution. *Capitalism and Society*, 3(2): 1. DOI: https://doi.org/10.2202/1932-0213.1040

ECSA (European Citizen Science Association). 2016. 10 Principles of Citizen Science. 2016. Available at https://ecsa.citizen-science.net/engage-us/10-principles-citizen-science.

Eitzel, MV, Cappadonna, JL, Santos-Lang, C, Duerr, RE, Virapongse, A, West, SE, Kyba, CCM, Bowser, A, Cooper, CB, Sforzi, A, Metcalfe, AN, Harris, ES, Thiel, M, Haklay, M, Ponciano, L, Roche, J, Ceccaroni, L, Shilling, FM, Dörler, D, Heigl, F, Kiessling, T, Davis, BY and **Jiang, Q.** 2017. Citizen science terminology matters: exploring key terms. *Citizen Science: Theory and Practice*, 2(1): 1. DOI: https://doi.org/10.5334/cstp.96

Follett, R and **Strezov, V.** 2015. An analysis of citizen science based research: usage and publication patterns. *PloS one*, 10(11): e0143687. DOI: https://doi.org/10.1371/journal.pone.0143687

Gallo, T and **Waitt, D.** 2011. Creating a successful citizen science model to detect and report invasive species. *BioScience*, 61(6): 459–465. DOI: https://doi.org/10.1525/bio.2011.61.6.8

Groom, Q, Weatherdon, L and **Geijzendorffer, IR.** 2016. Is citizen science an open science in the case of biodiversity observations? *Journal of Applied Ecology*, 54(2): 612–617. DOI: https://doi.org/10.1111/1365-2664.12767

Haklay, M. 2013. Citizen science and volunteered geographic information: Overview and typology of participation. In: Sui, D, Elwood, S, and Goodchild, L (eds). *Crowdsourcing Geographic Knowledge*, 105–122. Dordrecht: Springer. DOI: https://doi.org/10.1007/978-94-007-4587-2_7

Higgins, JW, Baillie, SR, Boughey, K, Bourn, NA, Foppen, RP, Gillings, S, Gregory, RD, Hunt, T, Jiguet, F, Lehikoinen, A and **Musgrove, AJ.** 2018. Overcoming the challenges of public data archiving for citizen science biodiversity recording and monitoring schemes. *Journal of Applied Ecology*, 55(6): 2544–2551. DOI: https://doi.org/10.1111/1365-2664.13180

Holdren, J. 2015. *White House Memorandum Addressing Society and Scientific Challenges through Citizen Science and Crowdsourcing*. Washington, DC: White House Office of Science and Technology Policy.

Irwin, A. 1995. *Citizen Science: A Study of People, Expertise, and Sustainable Development*. New York: Routledge.

Jennett, C, Kloetzer, L, Schneider, D, Iacovides, I, Cox, A, Gold, M, Fuchs, B, Eveleigh, A, Mathieu, K, Ajani, Z and **Talsi, Y.** 2016. Motivations, learning and creativity in online citizen science. *Journal of Science Communication*, 15(3): A05. DOI: https://doi.org/10.22323/2.15030205

Kullenberg, C and **Kasperowski, D.** 2016. What is citizen science?–A scientometric meta-analysis. *PloS one*,

11(1): e0147152. DOI: https://doi.org/10.1371/journal.pone.0147152

Kelling, S. 2018. Improving Data Quality in eBird-the Expert Reviewer Network. *Biodiversity Information Science and Standards*, 2: e25394. DOI: https://doi.org/10.3897/biss.2.25394

Kosmala, M, Wiggins, A, Swanson, A and **Simmons, B.** 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10): 551–560. DOI: https://doi.org/10.1002/fee.1436

Levin, N and **Leonelli, S.** 2017. How does one "open" science? Questions of value in biological research. *Science, Technology, & Human Values*, 42(2): 280–305. DOI: https://doi.org/10.1177/0162243916672071

Lukyanenko, R, Parsons, J and **Wiersma, YF.** 2016. Emerging problems of data quality in citizen science. *Conservation Biology*, 30(3): 447–449. DOI: https://doi.org/10.1111/cobi.12706

Mayernik, MS. 2017. Open data: Accountability and transparency. *Big Data & Society*, 4(2): 2053951717718853. DOI: https://doi.org/10.1177/2053951717718853

Morello-Frosch, R, Brody, JG, Brown, P, Altman, RG, Rudel, RA and **Pérez, C.** 2009. Toxic ignorance and right-to-know in biomonitoring results communication: A survey of scientists and study participants. *Environmental Health*, 8(1): 6. DOI: https://doi.org/10.1186/1476-069X-8-6

Munafò, MR, Nosek, BA, Bishop, DV, Button, KS, Chambers, CD, Du Sert, NP, Simonsohn, U, Wagenmakers, EJ, Ware, JJ and **Ioannidis, JP.** 2017. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1): 0021. DOI: https://doi.org/10.1038/s41562-016-0021

Nascimento, S, Rubio Iglesias, JM, Owen, R, Schade, S and **Shanley, L.** 2018. Citizen science for policy formulation and implementation. In: Hecker, S, Haklay, M, Bowser, A, Makuch, Z, Vogel, J and Bonn, A (eds). *Citizen Science: Innovation in Open Science, Society and Policy*. London: UCL Press. DOI: https://doi.org/10.14324/111.9781787352339

Nature. 2015. Editorial: Rise of the citizen scientist. *Nature*, 524(7565): 265. DOI: https://doi.org/10.1038/524265a

OECD. 2020. Open Science. *OECD,* 2020. Available at: https://www.oecd.org/science/inno/open-science.htm.

Parrish, JK, Burgess, H, Weltzin, JF, Fortson, L, Wiggins, A and **Simmons, B.** 2018. Exposing the science in citizen science: Fitness to purpose and intentional design. *Integrative and comparative biology*, 58(1): 150–160. DOI: https://doi.org/10.1093/icb/icy032

Roman, L, Scharenbroch, B, Ostberg, J, Mueller, L, Henning, J, Koeser, A, Sanders, J, Betz, D and **Jordan, R.** 2016. Data quality in citizen science urban tree inventories. *Urban Forestry & Urban Greening*, 22: 124–135. DOI: https://doi.org/10.1016/j.ufug.2017.02.001

Rotman, D, Preece, J, Hammock, J, Procita, K, Hansen, D, Parr, C, Lewis, D and **Jacobs, D.** 2012. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the 15th ACM conference on Computer supported cooperative work & social computing*, 217–226. Vancouver, Canada, on March 15–18, 2012. DOI: https://doi.org/10.1145/2145204.2145238

Schade, S, Tsinaraki, C and **Roglia, E.** 2017. Scientific data from and for the citizen. *First Monday*, 22(8). DOI: https://doi.org/10.5210/fm.v22i8.7842

Shirk, J, Ballard, H, Wilderman, CC, Phillips, T, Wiggins, A, Jordan, R, McCallie, E, Minarchek, M, Lewenstein, BV, Krasny, ME and **Bonney, R.** 2012. Public participation in scientific research: A framework for deliberate design. *Ecology and society*, 17(2): 29. DOI: https://doi.org/10.5751/ES-04705-170229

Sheppard, SA, Wiggins, A and **Terveen, L.** 2014. Capturing quality: Retaining provenance for curated volunteer monitoring data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 1234–1245. Baltimore, MD, on 15–18 February 2014. DOI: https://doi.org/10.1145/2531602.2531689

Silvertown, J. 2009. A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9): 467–471. DOI: https://doi.org/10.1016/j.tree.2009.03.017

Specht, H and **Lewandowski, E.** 2018. Biased Assumptions and Oversimplifications in Evaluations of Citizen Science Data Quality. *The Bulletin of the Ecological Society of America*, 99(2): 251–256. DOI: https://doi.org/10.1002/bes2.1388

Storksdieck, M, Shirk, JL, Cappadonna, JL, Domroese, M, Göbel, C, Haklay, M, Miller-Rushing, AJ, Roetman, P, Sbrocchi, C and **Vohland, K.** 2016. Associations for Citizen Science: Regional Knowledge, Global Collaboration. *Citizen Science: Theory and Practice*, 1(2). DOI: https://doi.org/10.5334/cstp.55

Sturm, U, Gold, M, Luna, S, Schade, S, Ceccaroni, L, Kyba, CCM, Claramunt, B, Haklay, M, Kasperowski, D, Albert, A and **Piera, J.** 2017. Defining principles for mobile apps and platforms development in citizen science. *Research Ideas and Outcomes*, 3: e21283. DOI: https://doi.org/10.3897/rio.3.e21283

Sullivan, B., Phillips, T, Dayer, AA, Wood, CL, Farnsworth, A, Iliff, MJ, Davies, IJ, Wiggins, A, Fink, D, Hochachka, WM and **Rodewald, AD.** 2017. Using open access observational data for conservation action: A case study for birds. *Biological Conservation*, 208: 5–14. DOI: https://doi.org/10.1016/j.biocon.2016.04.031

Tenopir, C. Allard, S, Douglass, K, Aydinoglu, A, Wu, L, Read, E, Manoff, M and **Frame, M.** 2011. Data Sharing by scientists: Practices and perceptions. *PLoS One*, 6(6): e21101. DOI: https://doi.org/10.1371/journal.pone.0021101

Tenopir, C, Dalton, ED, Allard, S, Frame, M, Pjesivac, I, Birch, B, Pollock, D and **Dorsett, K.** 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS One*, 10(8): e0134826. DOI: https://doi.org/10.1371/journal.pone.0134826

van der Velde, T, Milton, DA, Lawson, TJ, Wilcox, C, Lansdell, M, Davis, G and **Hardesty, BD.** 2017. Com-

parison of marine debris data collected by researchers and citizen scientists: Is citizen science data worth the effort? *Biological conservation*, 208: 127–138. DOI: https://doi.org/10.1016/j.biocon.2016.05.025

Wilkinson, M, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3: 160018.

Wiggins, A, Newman, G, Stevenson, RD and **Crowston, K.** 2011 December. Mechanisms for data quality and validation in citizen science. In *e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference*, 14–19. Stockholm, Sweden, on 5–8 December, 2011. DOI: https://doi.org/10.1109/eScienceW.2011.27

Wiggins, A, Bonney, R, Graham, E, Henderson, S, Kelling, S, LeBuhn, G, Litauer, R, Lots, K, Michener, W and **Newman, G.** 2013. *Data management guide for public participation in scientific research*. Albuquerque, NM: DataOne Public Participation in Scientific Research (PPSR).

Wiggins, A, Bonney, R, LeBuhn, G, Parrish, JK and **Weltzin, JF.** 2018. A Science Products Inventory for Citizen-Science Planning and Evaluation. *BioScience*, 68(6): 436–444. DOI: https://doi.org/10.1093/biosci/biy028

How to cite this article: Bowser, A, Cooper, C, de Sherbinin, A, Wiggins, A, Brenton, P, Chuang, T-R, Faustman, E, Haklay, MM and Meloche, M. 2020. Still in Need of Norms: The State of the Data in Citizen Science. *Citizen Science: Theory and Practice*, 5(1): 18, pp. 1–16. DOI: https://doi.org/10.5334/cstp.303

Submitted: 12 December 2019 **Accepted:** 02 July 2020 **Published:** 04 September 2020

]u[*Citizen Science: Theory and Practice* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS