BEST PRACTICES FOR DATA MANAGEMENT IN CITIZEN SCIENCE: AN INDIAN OUTLOOK

THOMAS VATTAKAVEN[1*], VIJAY BARVE[2], GEETHA RAMASWAMI[3], PRIYA SINGH[4],
SUNEHA JAGANNATHAN[5], BALASUBRAMANIAN DHANDAPANI[6]

[1]Strand Life Sciences, Ground Floor, UAS Alumni Association Building, Veterinary College Campus, Bellary Road, Bangalore, 560 024
[2]Nature Mates Nature Club, 6/7, Bijoygarh, Kolkata 700032 West Bengal, India
[3]Nature Conservation Foundation, 1311,"Amritha", 12th Main Vijayanagar 1st Stage, Mysore, 570 017.
[4]Researchers for Wildlife Conservation (RWC), National Centre for Biological Sciences, GKVK Campus, Bangalore, 560065.
[5]Dakshin Foundation, #2203, D Block, 8th Main, 16th D Cross, Sahakar Nagar, Bengaluru, 560092.
[6]French Institute of Pondicherry, 11, St. Louis Street, Pondicherry, 605001.

*Corresponding author: Thomas Vattakaven, Email: thomas.vee@gmail.com

Abstract. Citizen science has been in practice since the 1800s and is an important source of data for scientists and other applied users. It plays a vital role in democratizing science, providing equitable access to scientific participation and data, helps build the capacity of its participants, inculcates the spirit of scientific endeavor and discovery and sensitizes participants towards species and habitat conservation, creating a sense of stewardship towards nature. In recent years, citizen science, especially in biodiversity, has rapidly developed with the rising popularity of smartphones, and widespread access to the internet, leading to wider adoption globally. India has also witnessed a surge in the number of new citizen science projects being initiated and increased participation in these projects. With more proponents looking at initiating such projects, there is little documentation from an Indian perspective on setting up, collecting, managing, and maintaining biodiversity-focused citizen science projects, especially in a data-management context. We have attempted to fill this void by examining the best practices across the data life cycle of citizen science projects while keeping in mind sensitivities and scenarios in India. We hope this will prove to be an important reference for citizen science practitioners looking to better manage their data in their projects.

Key words: Biodiversity, Citizen science, Data, Data management, India, Licensing, Metadata, Protocol, Quality assurance, Standards, Data lifecycle

Citizen science has evolved as a significant field of practice, and its role in contributing to new knowledge on biodiversity is well established (e.g., Kobori et al. 2016; Schuttler et al. 2019). While originating in the West, it has spread globally, particularly in India in the last decade (Sekhsaria and Thayyil 2019). The evolution of citizen science as a field of practice and research has necessitated an inquiry towards overarching insights, standards, vocabulary, and guidelines (Vohland et al. 2021).

In 2020, India hosted its first CitSci India Conference for Biodiversity[1], bringing citizen science practitioners, researchers, educators, students, and policymakers interested in biodiversity and citizen science. This virtual meeting was hosted as part of the preparatory phase of the National Mission on Biodiversity and Human Well-being proposed by the Biodiversity Collaborative, with the National Biodiversity Authority as a nodal agency. CitSci India 2020 was a starting point to bring together the citizen science community in India under one platform to share experiences, inspire each other, and engage in discussions related to citizen science. Two prominent topics that surfaced during these discussions were the importance of ethics, diversity and inclusion, and

---

[1] https://citsci-india.org/.

data in citizen science. In this context, we focus on best practices in data management for Indian practitioners.

Following global trends, citizen science efforts involving biodiversity in India have rapidly gained pace over the past few years. With larger and smaller-scale citizen science projects increasingly launched in India each year, voluminous data is generated on various aspects of biodiversity (see list of Indian citizen science projects[2]). However, this also raises several issues regarding data, such as ownership, accessibility, attribution, storage, interoperability, and quality. The working group on Citizen Science Data was tasked with identifying significant aspects related to data on which project proponents should have clear procedures and policies. To put together this document, we surveyed existing global practices and standards and described various options that projects could adopt, with some guidance about benefits and costs associated with each option. The document is intended to form a toolkit for citizen science practitioners in India and elsewhere who seek to make informed decisions on various aspects of data.

For this document, we use a definition of 'citizen science' provided by Guerrini et al. (2019), as per which, citizen science "... generally refers to an approach to scientific inquiry in which members of the public participate in one or more steps of the research process other than, or in addition to, allowing personal data or biospecimens to be collected from them for analysis by others''. At this juncture, it is worthwhile to mention that there is an effort to replace the usage of the word "citizen" with "community" science to be more inclusive (Cooper et al. 2019), although some prefer to retain the distinction between these two terms (Dosemagen and Parker 2019). For the context of this paper, we primarily limit our reference to citizen science projects within biodiversity that at least partially utilize online participatory mediums with databases and servers that make data and its products accessible online.

Types of Citizen Science Projects

Citizen science initiatives vary extensively in their aims and objectives across disciplines and citizen engagement. An attempt to classify common citizen science projects can be undertaken based on parameters such as the research question, modes of participation, medium of participation, and mode of the survey, as discussed in detail below.

Research Question/Focus. Although citizen science has traditionally been used to address targeted research questions and hence involves specified protocols, the advent of online mediums and the ability to crowdsource content has paved the way for more open-ended platforms. Such platforms may engage citizen scientists in tasks such as gathering sightings of species or transcribing or classifying data for which the uses may be unknown or changing (Lukyanenko et al. 2016). Based on the above criteria, projects may be classified as generalist or specialist projects.

An alternate way of looking at this type of focus may be to classify projects based on the taxa of focus. There are larger generalist initiatives that have little or no restriction based on the taxonomic focus (e.g., India Biodiversity Portal[3], IBP), while targeted projects often tend to focus on selected or a single taxonomic group or species (e.g., Biodiversity Atlas - India[4], Bird Count India[5], Wild Canids-India Project[6], Marine Life of Mumbai[7]).

Modes of Participation. Citizen science initiatives can vary in terms of who initiates a project or the level and stage of involvement of volunteers or the general public in an initiative depending on the project's objectives. Project initiators play an essential role in defining the nuances of a project and, hence, determining the end goals that influence 'the political authority of science' (Kimura and Kinchy 2016). Similarly, the composition and training of citizen science initiators vary across projects. For this document, we highlight different types of public-scientist collaborations that qualify as citizen science engagements based on Veeckman et al. 2019.

Citizen science projects can be categorized based on multiple criteria: the extent of citizen participation, taxonomic focus of the project, or medium of participation. Table 1 describes different types of citizen science programs based on the extent of involvement of citizens, as described in Veeckman et al. 2019.

Medium of participation. Currently, two distinct channels allow establishing a citizen science initiative. These include:

---

2   https://citsci-india.org/citizen-science-projects/.

3   https://indiabiodiversity.org/.
4   https://www.bioatlasindia.org/.
5   https://birdcount.in/.
6   https://www.wildcanids.net/.
7   https://www.marinelifeofmumbai.in/.

Table 1. Types of citizen science programs based on the extent of participation by citizens, as described in Veeckman et al. (2019).

Type of project	Extent of citizen participation
Crowdsourcing	Volunteers remain passive while contributing time and equipment only
Distributed intelligence	Volunteers are involved with simple interpretations or categorizing material from gathered data
Participatory science	Volunteers define a problem, collect data and assist scientists in analyzing the data. Interpretation and analysis handled by scientists
Extreme citizen science	Volunteers and scientists collectively determine stages of the project, with the former handling all tasks related to the study and executing them. Scientists only act as facilitators on these projects.
Contributory project	Volunteers are invited to contribute data, while scientists decide the research focus of the study, and analyze and interpret data.
Collaborative project	Flexible projects where the scientist involved may identify the research focus of a project, while volunteers participate at different stages of the study based on their interest
Co-created project	Aimed at influencing public policy or educational agenda. Volunteers identify a set of questions, answers to which are thereafter pursued in consultation with scientists on the project.

A.	Independent platforms via web or smartphone-based methods using protocols built explicitly for the project context. Such platforms allow for flexibility in developing independent protocols tailor-made to suit the requirements of the study.

B.	Larger aggregator platforms with the ability to host independent projects within them, e.g., India Biodiversity Portal, Biodiversity Atlas - India, iNaturalist[8], CitSci.org[9]. These platforms usually host a range of projects that collectively benefit from an existing user-base of citizen science contributors, are easy to use with access to pre-vetted guidelines, instructions of usage, terms, and conditions, and other legal and technical formalities addressed. They are also equipped with measures to ensure data security and data-quality regulations. All these features allow them to be used with ease across a diversity of projects and overcome the lack of technical know-how amongst project managers.

A few initiatives are platform-independent and use social media and mobile messaging applications such as Facebook, Whatsapp, or email to gather biodiversity data. Data collected through such mediums are primarily not structured by default, nor are they controlled environments with binding data policies or licenses. Most of these are still emergent, and al-

though there is potential to crowdsource content using these increasingly popular mediums, due to their free-form nature of the interaction, much effort will be needed to extract, curate, and cleanse the content before use as meaningful, structured data.

Mode of survey. Citizen science may be carried out in a conventional scientific framework with a standardized field protocol. However, the most popular citizen science initiatives, especially those that allow for data entry through online interfaces and recruit online participation, are increasingly being done without standardized field protocols, giving rise to the term "opportunistic sampling." We discuss below some pros and cons of opportunistic versus structured data sampling from the perspective of participant motivation and data quality.

Citizen Science and Data-Life Cycle

Like most scientific data, citizen science data follow a general data life cycle. For the purpose of this paper, we have chosen to adopt the high-level Science Data Lifecycle Model (SDLM) (Faundeen et al. 2014), to illustrate how data management activities relate to citizen science data workflows and recommend actions and activities at each stage of the model. SDLM comprises primary elements that proceed sequentially and cross-cutting elements performed across stages of the data life cycle. The SDLM and its elements are summarized in Figure 1.

Keeping in mind the above data model, we have attempted to structure our paper into broad sections

---

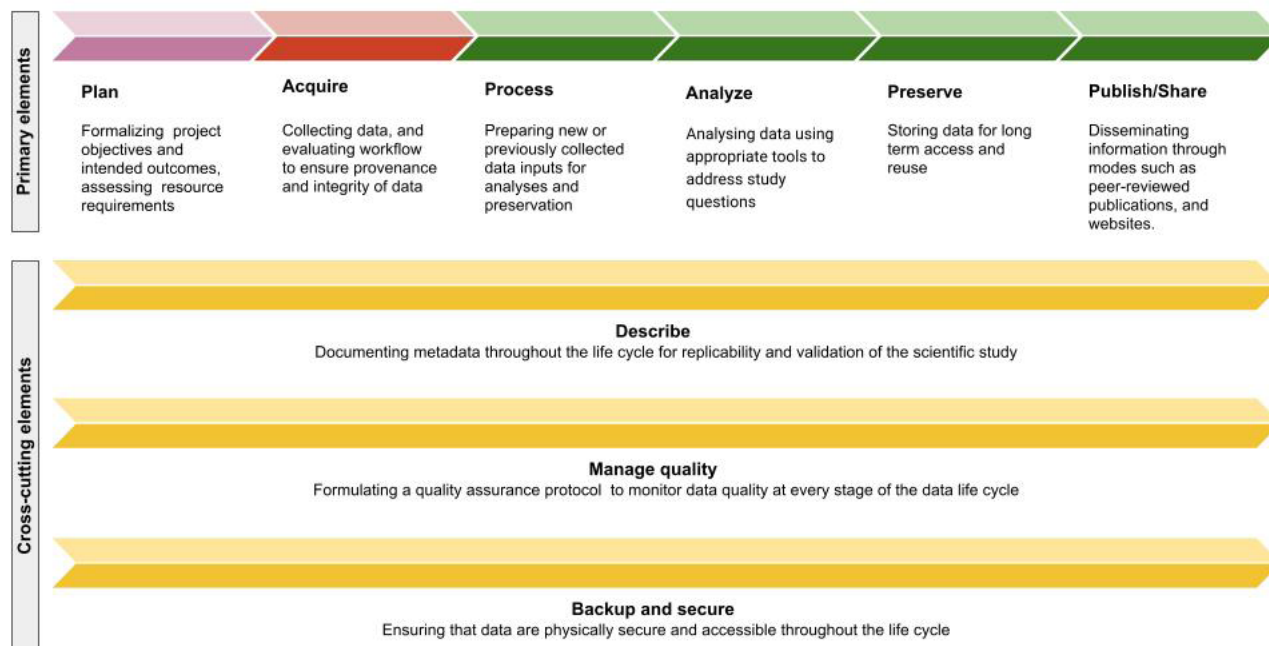[8] https://www.inaturalist.org/.
[9] https://citsci.org/.

Figure 1: The Science Data Life Cycle Model has primary and cross-cutting elements that help determine action at different stages of data collection, preservation and analyses. Primary elements include planning, acquiring, processing, analyzing, preserving, and publishing data, while the cross-cutting elements run parallelly throughout the data life cycle and involve describing metadata, managing data quality, and data security/backup (Faundeen et al 2014). The primary elements of the data life cycle can be further mapped to 'before' data acquisition (pink arrow), 'during' a project (orange arrow), and 'after' data are acquired (green arrows).

that cater to aspects of data within citizen science: before starting a project (planning), during the implementation of the project (acquiring), and after gathering data (processing, analyzing, preserving, and publishing data). Certain aspects covered below have cross-cutting implications and may be relevant at multiple stages of the data life cycle but may be covered in more detail in one section to avoid repetition.

Before Starting

In conformity with scientific practice, before initiating a citizen science project, it is essential to premeditate on a project's objectives, develop hypotheses, identify methods of acquiring, analyzing, and interpreting data, and ideal ways of disseminating results. Potential challenges need to be identified to ensure that project end goals are achieved effectively. This section summarises points to keep in mind at different stages of a citizen science project.

Identify Project Goals and Means of Implementation

Citizen science programs vary in their primary goals and the extent and role of public participation, with some projects exclusively aiming to achieve public engagement (refer to Types of Citizen Science

Projects). Projects can have a broad focus, such as creating generic biodiversity repositories or targeting focused research questions. Preliminary intensive review of the topic being pursued allows determination of the suitability of citizen science as a study technique.

Identify Target Participants/Stakeholders

Delineating target participants helps one design appropriate strategies to recruit, train and engage volunteers for a program. These can be selected based on need (not all citizen science projects require a targeted volunteer base), required skill sets (such as swimming, diving, climbing, identifying species), access to technology (such as smartphones), or age (adults or children). At times, engagement with intermediaries (such as schools, colleges, tourism ventures) may be required to enlist participation. When soliciting involvement from local communities, planning for localizing content in regional languages can help enhance participation and outreach.

Building Online and Offline Infrastructure

The backbone of a citizen science project is the infrastructure that it requires to function: to maintain registers of participants; collect, manage, curate,

and store data; and disseminate results and maintain regular communication with participants. Online infrastructure such as websites, smartphone applications, or even simple forms of communication like Whatsapp groups allows contributing data. Suitable back-end databases that store data in appropriate formats and enable interfaces to query and retrieve data efficiently need to be chosen at this stage. Establishing offline infrastructure requires high human effort, such as for collaborations, acquiring permits for access to protected areas (if needed), and initial outreach to gauge interest within target participants. It is also crucial to design data collection protocols (explained in greater detail in Data Collection Methodologies) for citizen science projects and pilot them within small focus groups to devise appropriate data collection methods.

Volunteer Recruitment, Engagement, and Outreach

Citizen science initiatives depend extensively on volunteer engagement to be successful. This engagement can be divided into three general phases, as follows: 1) **Volunteer recruitment** involves outreach to the target participants, testing out protocols with focus groups, and seeking volunteer feedback on initial processes. Typical methods employed are social media outreach, publicity articles in print media, tapping email list services, and presentations at target institutions such as nature clubs, schools, or colleges. In the case of projects targeting niche species that are uncommon or restricted in their distribution, establishing partnerships with local communities or tour operators may also be considered. 2) **Volunteer education and capacity building**: in addition to gathering data, citizen science initiatives also endeavor to increase scientific and ecological literacy among the public. In some cases, volunteers might require specific knowledge (species identification skills, basic survey skills) to participate in a program. The extent of skill training and knowledge exchange often depends on the data collection methodology. Volunteer education is a long-term exercise that needs to be carried out regularly. Practitioners need to ensure that volunteers and contributors understand the scientific problem being addressed, are trained well in collecting information, can use technology (if any) required for data contribution, and collect data in a standardized manner. Errors can be minimized by training and reiterating the collection protocol. The contribution process needs to be tested periodically to recognize new error sources and iteratively update

training and contribution processes. 3) **Volunteer retention**: citizen science efforts benefit from retaining volunteers over the long term, as their expertise and skill are likely to increase with time. However, this exercise requires innovative methods to sustain the interest of long-term volunteers. Leaderboards (to track the highest participation) or games and contests may encourage participation. Not all projects start with a captive volunteer base, and citizen participants may see turnover throughout the project duration. For long-term projects renewing interest in the project to recruit newer participants is crucial.

Data Management, Analysis, and Dissemination

Maintaining, curating, and analyzing data are critical aspects of a citizen science program. Data management involves data storage, curation, and backup techniques, ensuring that data are not lost once a project is deemed to be complete. It is important to note that citizen science is a long and evolving effort - the goals of a project might change over its lifetime. Considering this, one must follow data standards to maintain the usefulness of the data collected. Furthermore, it is essential to ensure that data are analyzed and visualized by participants, including non-experts, in an engaging manner. Mechanisms for user interaction with the data, roles, and permissions for data validators, strategies to flag erroneous data, etc., need to be thought of at this point and must be a continuous endeavor throughout the project, evolving with time.

Data Collection Methodologies

Citizen science projects vary in the rigor of sampling protocol, from simple occurrence reporting to more structured data collection techniques. This presents practitioners with a trade-off between volunteer participation and the quality of data collected. Often, programs with rigorous volunteer training and sampling protocols obtain better quality data but with reduced levels of participation. In contrast, those with simple data collection methods report higher participation, often with biased and noisy data. Techniques such as data collection forms or semi-structured surveys can reduce the need for rigorous training while still ensuring that data are collected in a prescribed format (Bonney et al. 2009; Kelling et al. 2019).

Incentivizing quality over the number of observations in leaderboards and gamification techniques can be helpful. Gamification is used to motivate par-

ticipants to contribute data to maximize contributions and enhance volunteer retention. It can range from adding a point system to ranking, creating leader-boards, giving badges or rewards, to creating an actual game that requires enhanced engagement from the participant. Prioritization of spatial and temporal scales where there are data gaps, rather than species or numbers of records, could result in more even distribution of biodiversity records, thus reducing spatial and temporal biases (Callaghan et al. 2019).

Planning for Data Quality Assurance

The credibility and quality of citizen science data is often questioned even though recent studies have challenged this view (e.g., Barve 2014). Hence, data quality assurance is an important topic to consider at each data cycle stage, i.e., data collection, upload and ingestion, storage, management, and analysis. Data should be accurate, precise, and representative to enhance credibility[10], but when the desired accuracy and precision are not achieved, it is important to document its *known quality* as data quality and usability depends on the users' questions (Chapman 2005b,a). Data needs to be thoroughly vetted for accuracy by curators or professionals pre-identified by the project and reflect reality. Consistency and replicability can ensure data precision, while spatial and temporal representativeness is necessary for any scientific exercise. Other ancillary information such as date, location, time of observation, weather conditions, etc. further aid in improving data quality and reliability. Maintaining records with data provenance allows preserving information about the evolution of data and methodologies used to acquire it, which is vital for debugging, tracking changes, auditing, and evaluating quality.

Wiggins et al. (2011) categorize potential biases in ecological data obtained using citizen science. Some of these include (a) positive spatial bias induced by areas with a higher human population, better or easier accessibility, and better accessibility to the internet in urban spaces (Geldmann et al. 2016). This may also be true for frequently visited hotspots (Boakes et al. 2016; Tiago et al. 2017); (b) Positive temporal bias induced by a more significant number of records during weekends, holidays, and contests, which is particularly of concern in phenology studies, where data seasonality is an integral part of the research (Courter et al. 2013). (c) Taxonomic bias is noticeable, with rare, cryptic, or difficult-to-identify

species often being underrepresented or remaining unidentified leading to a paucity of data for such species (Falk et al. 2019). In addition, very commonly observed species tend to be overlooked and under-reported, while highly sought after species may also be over-reported, leading to non-representative sampling (Troudet et al. 2017; Callaghan et al. 2021). Finally, (d) observer bias induced by individual perceptions and levels of experience (Gonsamo and D'Odorico 2014; Callaghan et al. 2021). Identifying sources of bias is necessary to help managers design appropriate strategies at different phases of projects to mitigate or manage them effectively.

A variety of end-users can utilize data from citizen science projects and it is important to identify the likely end-users of the data (e.g., scientists, policymakers, amateur naturalists) at the initial stages of a project. It should be noted that the onus of maintaining data quality is the shared responsibility of the project as well as data-users who must examine and put thought into the use of data and correct for biases, error rates, or quirks.

Data quality and minimization of biases can be accounted for before data collection and during the data contribution stage. Baker et al. 2021, summarize the types of data and levels of evidence at the data contribution stage which would require verification - a) Levels of evidence: reporting of sightings without evidence, and reporting sightings with evidence such as photo/video/audio/specimen, b) Types of observations: direct observation, where the taxon is observed directly, and indirect observation, wherein taxon signs (such as tracks, dung, etc.) are recorded.

The mechanisms and criteria required for data validation need to be considered at this stage and suitably incorporated into the collection procedure to provide for the availability of fields or the target precision levels to be achieved. The ability to validate or curate records may be contingent on the presence of such information fields and without which data may be unverifiable.

Data analyses should also be planned and anticipated before data collection and should be appropriate for the kind of data collected and driven by the project's goals (Wiggins et al. 2011; Balázs et al. 2021). Factors affecting data quality need to be identified. This may include improper data collection, incorrect implementation of data collection protocols, a mismatch between project goals and data collection protocols, incomprehensive protocols that do not match end-user expectations, and use of data

---

[10] https://citizenscienceguide.com/design-sample-collection.

in wrong contexts.

Balázs et al. (2021) suggest the following at the planning stage of a citizen science project to ensure data quality and to make data conducive for further analyses: (a) simple and intuitive data collection protocol supplemented by a simple user interface design that is engaging and can be applied across a diverse group of users with varied skills, (b) calibrating and standardizing devices and recognizing limitations of technology, (c) appropriate documentation, and (d) metadata to prevent misuse of data in incorrect contexts.

Conferring with experts could enhance the quality of analyses. Inferences should be cautious and consider all the caveats of data accuracy and analysis. It is also beneficial to get the analyses reviewed by experts and peer groups.

Quality assurance through following standards. It is essential to incorporate data collection methods and protocols, fitness for use, and data quality assessment as part of the metadata/documentation (Assumpção et al. 2018). Adapting and adhering to standards helps in improving data quality and usability due to breaking up data attributes into appropriate terms and following controlled vocabularies to ensure each term conveys the correct meaning. In the biodiversity realm, standards developed by the Biodiversity Information Standards, originally called the Taxonomic Databases Working Group (TDWG[11]), like Darwin Core and Audubon Core (Table 1) and tools built around them, are readily available for citizen science projects to use and adapt.

Planning for Data Infrastructure

Contemporary citizen science projects are mostly born-digital, conceived and implemented predominantly in the digital ecosystem of information technology platforms, software applications, and toolchains. There are multiple online and offline infrastructure concerns. Proponents may wish to choose between larger aggregator platforms that allow projects within them vis-a-vis building independent applications. Fast-growing mobile application technologies have made it possible to quickly deploy data collection and integration tools with little effort (Lemmens et al. 2021). On the other hand, several large biodiversity data aggregating platforms are well established and have gained a reputation across the globe (ebird[12], iNaturalist), or country-level data (India Biodiversity Portal and Biodiversity Atlas for

India, Atlas of Living Australia[13] for Australia, national GBIF nodes, etc.). Others look at specific taxa or simply to address a particular question. There are apparent advantages of using an existing platform for biodiversity data collection and aggregation, as they readily provide technological infrastructure, communities, and tested infrastructures across data lifecycle (de Sherbinin et al. 2021).

Depending on the project's larger goals, there could be challenges in fitting the needs of a citizen science project to pre-existing templates and applications provided by such platforms. Such larger-scale citizen science initiatives should allow for flexibility in engaging at different ecological levels, different aspects of ecosystem changes, and conservation issues (Devictor et al. 2010). Many large aggregator platforms already support infrastructure that allows such flexibility. For example, the IBP allows creating groups within its infrastructure for any theme of interest, such as a taxonomic group. Forms for gathering data can be extended with the capability to include custom queries and fields, as required.

Data infrastructures for citizen science projects need to be adaptable to address the unique nature of each citizen science project. The data infrastructure should generally allow for data collection, aggregation, analysis, and dissemination, thus covering the whole data life-cycle management or digital information supply chain (Brenton et al. 2018). For example, citizen science projects could use a phone or web application to collect data, a cloud server to store the data, automated code to verify data, a web portal to promote interaction among contributors, and a back-end database structure such that it can be aggregated with other types of data. This would also mean that the infrastructure enables the participation of citizen scientists in the full range of scientific methods from problem definition, research design, analysis, and action (McQuillan 2014).

Citizen science projects in biodiversity tend to collect data across a wide assortment of attributes. These include taxonomic, evolutionary, biogeographic, functional, and interspecific interaction attributes of a taxon (König et al. 2019). Data infrastructure should be flexible enough to accommodate the diversity of data types in a variety of formats such as text, tabular, geo-spatial, and varying media types, including images, audio, and video. Such capabilities will influence the scalability of storage required, particularly for long-term projects. Cloud-based storage and content delivery networks in the mainstream IT

---

[11] https://www.tdwg.org/.
[12] https://ebird.org/home.

[13] https://www.ala.org.au/.

ecosystem have matured enough to ensure scalability and high availability across geographies.

Apart from these fundamental concerns on data models and data storage, one must ensure that the platform is stable and provides continuous access to participants with minimum downtime. Platforms need to cater to data security with well-defined data access policies, user authentication systems with defined roles, transparent workflows, and user-centered design (Bowser et al. 2020). Regular backup of data with multiple copies in multiple locations and a consistent preservation policy across sites is essential for the security and integrity of citizen science data. In keeping with the spirit of open science, one can also insist on using, developing, and deploying free/open-source technology stacks to help collaboratively build, share and replicate developed technologies for wider and unrestricted use.

Data Ownership

Data ownership is crucial for a project and needs deliberation at the planning stage. Participant perception of data ownership can influence their motivation in future participation. Yet, studies have indicated ambivalence in how participants feel about data ownership. On the one hand, most participants appear far removed from thoughts of data and its ownership. Each record is more of a personal nature experience and less so as data with legal ownership. Ganzevoort et al. 2017, best summarize this as constituting "*an "imagined contract" between volunteer naturalists and nature, based on respect and wonderment...*". On the other hand, participants believed that "*...data extracted from nature should properly be used towards its preservation*" and hence "wrong" use of data can result in citizens being resentful and withholding contribution. In some instances of data sharing, moral rights may get infringed, especially if the user of such data distorts or mutilates the data contributed by volunteers through re-use or if some private/sensitive information gets accidentally disclosed. Although most participants surveyed were against the unconditional use of data generated using citizen science, most participants were undecided on ownership, with some feeling data is nobody's property and others that it could be owned by the organization conducting the study (Ganzevoort et al. 2017).

It may also be said that participants may feel strongly about data in ways that are not covered under legal ownership and may not qualify for legal protection. However, it may be possible to validate

such feelings outside of traditional law through policies that put in practice exclusive or non-exclusive access to or control over data (Guerrini et al. 2019).

Sometimes traditional knowledge belonging to communities related to bio-resources or conservation practices might be part of such data. Establishing who owns this knowledge can be very challenging. In the case of community-held knowledge, it is easy to attribute ownership to a particular community. Still, there is ambiguity when the traditional knowledge is from an unidentifiable source or shared between communities spread across large territories. The knowledge may also be based on specific practices, beliefs, and linguistic representations, which may get lost in translation. One must be mindful of specific communities' cultural sensitivities and secretiveness to divulge knowledge. The communities must have the freedom to say no to sharing their knowledge if they wish, and if they agree, they should be allowed to choose how their knowledge is used.

Data Accessibility

It is essential to have prior clarity on data accessibility regarding who can access the data, at what stages, and for what purposes. Accessibility to data generated through citizen science projects is a core aspect. Open access to data allows for democratizing science and upholding the values of universal and equitable access to scientific data, especially when gathered through public participation. Just as we strive to make citizen science accessible to a diversity of participants and make 'doing science' inclusive, the resulting data must be accessible in ways that support reproducible science and can influence policy through bridging gaps between knowledge and action. There are outlying concerns that, more often than not, a citizen scientist's contribution disappears into the closed databases within institutions, and particular emphasis needs to be paid to alleviate these concerns.

What is open data? There are variable interpretations of the term 'open data'. As stated by the Open Knowledge Foundation, "*data is open if it can be freely accessed, used, modified and shared by anyone for any purpose[14] - subject only, at most, to requirements to provide attribution and/or share-alike*". Specifically, open data is defined by the Open Definition and requires that the data be, both, a) Legally open: where it is made available under an open (data) license that allows anyone to freely access, reuse and

---

[14] https://opendefinition.org/.

redistribute the data; b) Technically open: where the data is made available freely or at a cost, no more than what is required for its reproduction and in formats that are in bulk and machine-readable.

Therefore, open data means that it is complete, preferably downloadable over the internet in a convenient and modifiable format without requiring proprietary software to process. It should also be "*provided under terms that permit reuse, redistribution, allow intermixing with other datasets and must not discriminate against fields of endeavour or against persons or groups such as against commercial use*." In this context, the data should conform to the FAIR open data principles to be Findable, Accessible, Interoperable, and Reusable.

Why is citizen science data not always open? Due to the varied nature of citizen science projects, proponents, and contexts of funding, not all projects may likely be in a position to adhere entirely to the tenets of open data. Some disagree with open data with justifications that vary in their context. With data serving as the currency for competition between scientists for limited funding and prestige through publications, the conventions of traditional academic publishing have resulted in a tendency to hoard data in closed silos (Hampton et al. 2013). Others cite the burden and expense of running massive data projects, curating data and processing, and managing people involved as a justification for exclusive access and reaping the resulting benefits (Walker et al. 2016). The data can be used as leverage to fund further project activities or, more importantly, to obtain acknowledgment, particularly as authors on publications. Some might be willing to share data but on request to keep track of how it is being used, hence not publishing them under an open-access license. Other important reasons for data not being open are projects anchored at institutions having restrictive blanket data policies, especially concerning intellectual property rights for work generated as a part of the institution. Similarly, funding bodies sometimes impose conditions on data release as a part of their terms, which may be restrictive. Finally, privacy concerns regarding those of the participants generating the data and when the data is about a species of concern may be a key consideration in limiting open access (Groom et al. 2017).

Why it is recommended that citizen science data be open-access. There are many reasons for recommending citizen science data be open-access. Groom et al. 2017, state that "*The voluntary aspect of the time invested by citizen scientists is generally interpreted as being motivated primarily by its contribution to society and that society should profit from this effort through openly accessible data*." Open-access allows participants to track their participation alongside aggregated data from other participants, learn from it, and incorporate the learning into improving their knowledge. Opening the data has also been shown to motivate participants with greater frequency and depth (Bonney et al. 2009). The availability of open data allows easy and quick access for citizens and decision-makers to use as evidence towards influencing policy without waiting for formal assessments to emerge and closing the gap between knowledge and action. Open citizen science data thus enable participants to be at the "forefront of socially relevant science" (Hampton et al. 2013). Open data also supports reproducible science.

Ethical Considerations to be Made at This Stage

Some of the key ethical considerations in the planning stage of a citizen science initiative cover the realms of recognizing contributor rights in citizen science and designing socially inclusive projects. It also includes information on making data public or open-access versus limiting access (addressed in Data Accessibility).

With the growing popularity of citizen science across geographies and academic disciplines, it is being rapidly incorporated as a methodology to identify scientific queries and find means of answering socially relevant questions with the aid of citizen contributors or collaborators. This has led to a growing recognition of contributor rights and the need to address power imbalances between project handlers and contributors. Hence, project managers must recognize the participatory nature of citizen science, where volunteers contribute data and time obligingly and ensure that projects are socially inclusive. Projects must include participants irrespective of gender, geographical location, socio-cultural, religious, linguistic, and academic backgrounds (Paleco et al. 2021). Simultaneously, participation at all stages of a project must be permitted. To maximize participation, project designers should identify means of reaching out to all potential stakeholders. These could include interested citizen participants, such as members of the public, established citizen scientists or those from the scientific fraternity, academic institutions/ organizations, policy experts, etc. (Veeckman et al. 2019). Collaborating with schools,

communities directly associated with the study subject, or government bodies also helps in increasing participation (Veeckman et al. 2019).

<div align="center">

Project Implementation
*Data and Metadata Standards
and Data Infrastructure*

</div>

Once the study design, data quality, adherence to standards, data infrastructure, and accessibility are planned for in a citizen science project, implementation is next. At this stage, data infrastructure should be appropriate, scalable, and highly available for organizing the acquired data. It is also essential to incorporate data standards at this stage. Biodiversity data standards are shared rules and conventions to describe, record, and structure biodiversity data to enable data aggregation and exchange across different organizations generating and managing different data sets. Data standards enforce unambiguous definitions of what kind of data are being collected, follow well-defined ontologies and vocabularies, and standardize the usage of established protocols. It is recommended that citizen science projects follow prevalent international standards and adopt recommended storage formats and protocols.

The use of biodiversity data standards has two key objectives, as follows. (1) Data standards provide a comprehensive set of relevant attributes for most projects and meeting individual project needs for the collection and management of data. (2) Data standards aid in identifying a subset of core biodiversity data attributes that can be used to aggregate data.

The different international standards available for citizen science data are described in Table 2. User-contributed data are typically restructured slightly to adhere to project-specific standards before storing in databases. However, this standardization and large-scale aggregation may lead to a loss of contextual richness (Turnhout and Boonman-Berson 2011; Ganzevoort et al. 2017). Apart from adhering to standards, it is good practice to ensure that data are Findable, Accessible, Interoperable and Reusable (FAIR, Wilkinson et al. 2016) and consider other principles like Collective benefit, Authority to control, Responsibility and Ethics (CARE, Carroll et al. 2021) in the context of data originating from indigenous communities. The FAIR data principles facilitate the discovery of knowledge, integration, and use by the larger scientific community. FAIR principles ensure that data are discoverable to computational agents (e.g., computers and apps) and humans through standard-

ized protocols – otherwise referred to as 'machine actionable' data (Wilkinson et al. 2016). In highly linguistically diverse contexts, such as in India, data infrastructure should be developed with support for multiple languages to enable data input and support accesses to facilitate large-scale participation.

<div align="center">

Quality Assurance and Quality Control

</div>

Citizen science projects with high data quality adhere to good practices related to standards, metadata, and documentation. Such projects also ensure that data errors, biases, uncertainty, and ethical concerns are addressed through volunteer training and validation, calibration of data collection tools, iterative evaluation, and enhancements, and flagging of erroneous records via appropriate validation methods (Kosmala et al. 2016; Ratnieks et al. 2016; Downs et al. 2021). While integrating data from different citizen science programs, incompatible design and inconsistencies in nomenclature can also affect data quality (Campbell et al. 2020).

The following fail-safes can ensure that data collection is accurate before and during data collection: profiling contributors and assessing their skill levels, piloting a citizen science project to get a sample of data and potential sources of errors and biases, following standardized methods of data collection and adopting established standards for terminology, participant training, auto-correction (e.g., erroneous geocoding), data verification, and facilitating access to data use (Balázs et al. 2021). Projects using devices should calibrate sensors, perform initial checks on devices and ascertain the ability of observers to use these devices to make accurate observations (de Sherbinin et al. 2021).

Depending on the types of observation, post-collection data verification is a crucial step to ensure data accuracy. Community/peer consensus, expert verification, automated verification, model-based verification (statistical models address random/individual variation, residual errors, and uncertainty of devices to flag erroneous observations), and linked data analysis (combine existing datasets to serve as reference data and use data mining tools to flag erroneous observations; Kelling et al. 2015; Balázs et al. 2021) are some of the existing methods that can be used for assessing data quality. In biodiversity data, species (identity, geographic co-occurrence with other species, rarity), environmental (time, date, location) and expertise (experience of the recorder) contexts should be verified through one or more methods

Table 2. International standards for citizen science data.

Standard	Description	URL
DarwinCore (DwC)	"A glossary of identifiers, labels, and definitions that facilitate the sharing of biodiversity information. DwC is based on taxa and their distribution documented through observations, specimens, samples, and related information. It is being regularly improved with the addition of terms as well as the development of extensions to map various sources of data accurately."	https://www.tdwg.org/standards/dwc/
Audubon Core Multimedia Resources Metadata Schema (AC):	"A set of vocabularies designed to represent metadata for biodiversity multimedia resources and collections, with the aim of determining the suitability of the media for specific biodiversity science applications. Among others, the vocabularies address such concerns as the management of the media and collections, descriptions of their content, their taxonomic, geographic, and temporal coverage, and the appropriate ways to retrieve, attribute and reproduce them."	https://www.tdwg.org/standards/ac/
The Access to Biological Collections Data (ABCD)	"An evolving comprehensive standard for the access to and exchange of primary biodiversity data (i.e. specimens and observations)"	https://www.tdwg.org/standards/abcd/
Ecological Metadata Language (EML)	"Defines a comprehensive vocabulary and a readable XML markup syntax for documenting research data. EML includes modules for identifying and citing data packages, for describing the spatial, temporal, taxonomic, and thematic extent of data, for describing research methods and protocols, for describing the structure and content of data within sometimes complex packages of data, and for precisely annotating data with semantic vocabularies."	https://eml.ecoinformatics.org/
Taxonomic Concept Transfer Schema (TCS)	"A schema to allow the representation of taxonomic concepts as defined in published taxonomic classifications, revisions and databases. It specifies the structure for XML documents to be used for the transfer of defined concepts. Currently, this standard is not followed widely."	https://www.tdwg.org/standards/tcs/

(Baker et al. 2021). Observer biases can also be accounted for in the post-collection stage through data filters and models that account for levels of contributor expertise and using AI-based techniques to reduce biases in model training data (Johnston et al. 2018; Chen and Gomes 2019; Steen et al. 2019).

Data that pass through the validation stages need to be curated. This involves processing raw data in terms of end-user requirements, ensuring that data meet reproducibility standards (for analyses), and lending themselves well to being combined with other standardized datasets. If the end-use of the data includes re-use or integration, data credibility can be increased by doing analyses on sampling approaches and quality and triangulating against other data sources. It is ideal to store citizen science data in the most disaggregated form with minimal privacy concerns and documented data quality assurance protocols (de Sherbinin et al. 2021; Downs et al. 2021).

Licensing

Licensing is necessary to ensure that media contributed by users to a citizen science project is used appropriately and data is appropriately cited. Copyright is a state-guaranteed right covering 'work' that includes intellectual creations, such as text, photographs, diagrams, maps, movies, etc. Ideas, knowledge, information, or data are traditionally not copyright-protected, and scientists have traditionally been content with being cited for their original work (Hagedorn et al. 2011), to facilitate public access and dissemination of knowledge. Although it is commonly assumed that data with no license applied is free for unrestricted use, this is not the case. The lack of a license poses ambiguity in its reuse, especially where the data usage terms have to be made explicit, especially for commercial usage (Groom et al. 2017), and may lead to unwitting copyright violations. Data must be made available under carefully crafted licenses where the terms and conditions for its reuse are made clear.

Adopting open, machine-readable licenses are recommended to meet the FAIR data principles discussed earlier (de Sherbinin et al. 2021). The most common license employed in citizen science data is the Creative Commons License (CC[15]). This license seeks to find a 'balance between public and private interests, and between the free flow of expressions of ideas and knowledge and state-guaranteed control and monopolies' (Hagedorn et al. 2011). Creative Commons licenses are not an alternative to copyright

and work alongside copyright, enabling one to modify copyright terms to best suit their needs. A violation of a CC license is a copyright violation. The CC licenses provide standardized terms-of-use definitions that have been adapted for various jurisdictions and upheld in court in several countries (Hagedorn et al. 2011). The license has been adapted for India under the aegis of Wikimedia India, Centre for Internet and Society, and Acharya Narendra Dev College[16].

CC licenses by default allow people to reuse, remix and adapt original works while still providing attribution to the original author. However, it understands that no single license can cover all use cases and instead offers a set of licenses to cover a wide range of use cases (Table 3). The CC licenses are accordingly adopted in whole or part by large data repositories such as the Global Biodiversity Information Facility (GBIF[17]), Wikipedia, and Wikimedia Commons, among others. CC0, CC-BY, and CC-BY-NC are the only CC license options recommended by GBIF. As scientific data is primarily facts and is not copyrightable, CC0 is the recommended license for data[18]. If any media are contributed as part of the data, the terms of use of the platform gathering the data should be clear on the applicability of the CC license to such media as well.

Another such relevant license is the Open Data Commons (ODC) maintained by the Open Knowledge Foundation[19]. Although Open Data Commons licenses are more suitable for data licensing, they are more specific to databases and apply only to database frameworks and structures, not to the particular content within a database. It allows for the "distinction between the data (base) and material (content) generated from it ("produced works")". ODC provides three types of licenses (Table 4).

India's open government data initiative started with the notification of the National Data Sharing and Accessibility Policy (NDSAP), by the Department of Science and Technology to the Union Cabinet in 2012 and the subsequent launch of the Open Government Data Platform India. The recommended licenses to be used for datasets published under NDSAP through the OGD platform remained unspecified until the release of the Government Open Data

---

[15] https://creativecommons.org/.
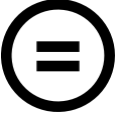
[16] https://wiki.creativecommons.org/wiki/india.
[17] https://www.gbif.org/.
[18] https://wiki.creativecommons.org/wiki/CC0_use_for_data.
[19] https://opendatacommons.org/.

Table 3: Creative Commons rights and licenses for data. The Creative Commons logo and icons used are from Wikipedia[21] and are under the public domain.

Icon	Right	Description
	Attribution ("BY")	Is a part of all CC licenses and requires users to give appropriate attribution to the creators of a work
	Share-Alike ("SA")	Allows the distribution of derivative works, but requires that all such works must also be shared under the same conditions ensuring that more restrictive licenses are not applied to derivatives.
	No Derivative Works ("ND")	States that the user "may not alter, transform, or build upon this work"
	Non-Commercial ("NC")	States that one "may not use this work for commercial purposes"

Types of CC licenses

Icon	Name	Applicable rights
	CC BY[22]	Attribution
	CC BY-SA[23]	Attribution-ShareAlike
	CC BY-ND[24]	Attribution-NoDerivatives
	CC BY-NC[25]	Attribution-NonCommercial
	CC BY-NC-SA[26]	Attribution-NonCommercial-ShareAlike
	CC BY-NC-ND[27]	Attribution-NonCommercial-NoDerivatives

Table 4: Open Data Commons licenses.

Name	Applicable rights
Open Data Commons Open Database License[28] (ODbL)	Attribution Share-Alike
Open Data Commons Attribution License[29] (ODC-By)	Attribution
Open Data Commons Public Domain Dedication and License[30] (PDDL)	Public Domain (All rights waived)

---

[20]https://en.wikipedia.org/wiki/Creative_Commons_license.
[21]https://creativecommons.org/licenses/by-sa/4.0.
[22]https://creativecommons.org/licenses/by-nd/4.0.
[23]https://creativecommons.org/licenses/by-nc/4.0.
[24]https://creativecommons.org/licenses/by-nc-nd/4.0.
[25]https://opendatacommons.org/licenses/odbl/.
[26] https://opendatacommons.org/licenses/by/.
[27] https://opendatacommons.org/licenses/pddl/.

License - India[28], which is governed by Indian law. It allows end-users to "*use, adapt, publish (either in original, or in adapted and/or derivative forms), translate, display, add value, and create derivative works (including products and services), for all lawful commercial and non-commercial purposes.*" However, the terms of the license remain ambiguous and have been criticized for being incomplete in many aspects, such as privacy and accountability of data providers (Kodali 2017).

In addition, it is possible to set up custom bespoke licenses for a citizen science project. However, this is not a trivial endeavor and will almost certainly have to include legal offices and organizational research departments (Ball 2011). Such cases are usually unnecessary considering the availability of standard licenses as documented above, except when exceptional circumstances require the same. Creating additional bespoke licenses adds to the burden on end-users of the data in ensuring compliance and adhering to multiple license requirements.

Once a suitable license has been decided upon, one must attach that license to the data. This mainly involves putting out a statement that the data is released under the chosen license or public domain and a mechanism for retrieving the full text of the license itself. The rights statement must be displayed prominently to avoid ambiguity and confusion. Adding the rights statement within downloaded zip files in an RDF/XML format for machine recognition is highly recommended (Ball 2011).

Ethical Considerations at This Stage

Some of the key ethical considerations in the stage of data acquisition include clear, prior communication with potential participants before collecting data, information on data licenses, encouraging participants to contribute data collected following fair practices, and legal and social conformity to data being incorporated from indigenous communities. Prior communication regarding objectives of a project, terms of data usage, methods for data storage, recognition of participant roles, etc., should be communicated to participants at one or more stages during project implementation. Before participation, participant consent should be sought to ensure that contributed data are not misused, and participants are aware of data licenses. While participant data-usage trends can be used to communicate about project developments, unauthorized usage of participant personal in-

formation should be prevented (Sullivan et al. 2014).

Project managers need to be aware of regional laws and other legal components that govern the usage of information pertaining to indigenous communities. Traditional knowledge must be handled sensitively and collected only after receiving consent from indigenous groups. National laws related to copyright or protection of imagery and text narratives must be well understood before accessing such data and complied with. Efforts must be made to ensure project participants abide by government and community laws and regulations while accessing contributed data. In the case of biodiversity projects, the safety of biodiversity and participants should be a priority over data collection.

Gaming elements are often used in citizen science projects to positively influence participant engagement by creating an environment of fun, competition or both (Bowser et al. 2013; Iacovides et al. 2013). Gamification may reward participants for attaining high scores and can enhance user participation. However, it may also demotivate participants who do not achieve competitive targets, distract participants from scientific data collection, trigger unfair practices to inflate competitive scores, and withhold 'winning' information, conflicting with citizen science principles of open data. Skills and resources required to participate in games may create inequity by putting some participants at an advantage over others. Ponti et al. (2018), note that game design and its influence on participant strategies, contributor values and motivations, and acknowledgment of participation outside of the gaming context are essential aspects of citizen science gamification.

At the data-validation stage, artificial intelligence techniques such as deep learning and convolutional neural networks are now used in citizen science projects to classify images, especially for species identification. AI applications may be used to automatically classify visual, acoustic, and spatial information through learning algorithms that utilize vast datasets and extract and organize images from social media (Lamba et al. 2019; August et al. 2020). Here, ethical challenges arise when black-boxed artificial intelligence systems are trained with citizen science contributed open data but exclude citizens from understanding how their data contributions are used. Transparency in such AI systems is essential and can help detect biases in training datasets, thus improving the efficacy of these systems. For example, eBird's human/computer learning network

---

[28] https://data.gov.in/government-open-data-license-india.

(Kelling et al. 2012) is cited as an example of such a transparent system (McClure et al. 2020).

What To Do With the Data

Once data begins accumulating, considerations related to data storage, processing, analysis, and dissemination in standardized and verified forms and ensuring its longevity have to be implemented.

Data Infrastructure

Managing physical risks associated with data storage, ensuring security, and ensuring access to data across its lifecycle is critical. The USGS data lifecycle model recommends that such security measures cover "raw and processed research data, original science plan, data management plan, data acquisition strategy, processing procedures, versioning, analysis methods, published products, and associated metadata" (Faundeen et al. 2014).

The Bouchout Declaration for Open Biodiversity Knowledge Management[29], which aims "to promote free and open access to data and information about biodiversity by people and computers and to bring about an inclusive and shared knowledge management infrastructure," lists some core principles that are vital towards data perseverance:

- An agreed infrastructure, standards, and protocols to improve access to and use of open data;
- Persistent identifiers for data objects and physical objects such as specimens, images, and taxonomic treatments with standard mechanisms to take users directly to content and data;
- Tracking the use of identifiers in links and citations to ensure that sources and suppliers of data are assigned credit for their contributions;
- Registers for content and services to allow discovery, access, and use of open data;
- Linking data using agreed vocabularies, both within and beyond biodiversity, that enable participation in the Linked Open Data Cloud.

These approaches to data lifecycle management point to implementing various strategies pertaining to aggregation and processing of data for analysis, transformative action, and tracking usage. Multiple techniques in deploying persistent identifiers and URLs, Digital Object Identifiers, LifeScienceID, and Personally Identifiable Information are adopted across platforms to ensure that data in its various

types and stages are traceable. Implementation of such persistent identifiers will become a norm soon and will help ensure data quality, access and accreditation.

Trustworthy data repositories like Zenodo/Dryad, Mendeley Data, among others, could be considered for storing citizen science data that has been curated for research quality. Developed as part of efforts from Research Data Alliance, a set of harmonized common requirements for certification of research data repositories certifies that these remain trustworthy (CoreTrustSeal Standards And Certification Board 2019). For occurrence data, global repositories like GBIF, eBird, and IBP in India could act as apt data repositories to ensure the perpetuity of data. While many such data repositories are evolving with Long Term Ecological Observatories[30] and other state-sponsored initiatives, it is pertinent to note the significance of archiving citizen science initiatives with their raw data and the context within which they are conducted (Williams et al. 2018). This will ensure the dual goals of securing the perpetuity of citizen science data and maximizing re-use. Such public data archiving for citizen science initiatives are required but a challenge to build (Pearce-Higgins et al. 2018).

Data Standards

Data standards play an important role in biodiversity data publishing. Following data standards makes publishing either through aggregators like GBIF or in the form of data papers simple. It saves effort in describing metadata and makes published data readily usable for the intended user base. Data papers and data repositories often require the metadata to be marked up in standardized formats such as EML. Independent projects may use software such as R or Morpho[31] to markup the metadata from their datasets. Many larger platforms like iNaturalist or IBP serve as an archive and a publishing platform. They already have some standardization inbuilt within their structure, allowing data downloads to be served under such standards. Such platforms also have arrangements on publishing the data to global biodiversity repositories such as GBIF through common standards.

Data Accessibility

As stated earlier, Open Access data means that "data must be freely available for download online."

---

[29] http://www.bouchoutdeclaration.org/declaration/.

[30] https://lteo.iisc.ac.in/.

[31] https://old.dataone.org/software-tools/morpho.

This also implies that the data is accessible in formats that do not need proprietary software to open and must have an open license for reuse. The CSV format is generally used for tabular data download and ensures compatibility for machine-reading of the data in a machine-readable format.

Many sites require prior registration or serve download requests via a user's registered email. Imposing registration for data downloads is an accepted means of tracking data usage, ensuring compliance with the project's policies and the site's data licensing.

From the accessibility perspective, there is a need to involve citizens beyond the mere act of data collection and provide them with opportunities and incentives to interact with the data they have generated. Participants are rarely given opportunities beyond data collection, such as data analysis or interpretation (Kennett et al. 2015; Lukyanenko et al. 2016). Activities such as data consumption influence learning and conservation outcomes and may lead to better user retention in the project (Cooper et al. 2017). Such interaction can be achieved through participatory data analysis and visualization that can be user-generated as per their needs and variables of interest. Many projects are increasingly gravitating towards developing such interactive visualizations for participant engagement. However, since data analysis is usually an end-user's specific perspective, generic visualizations and analyses inbuilt into portals may be limited as they are usually set up to predefined criteria. Such limitations can be overcome through developing and offering APIs and client packages for popular data analysis software such as R or Python. Some examples of such packages are the 'rgbif'[32] and 'pygbif'[33] clients for interfacing with GBIF and the 'galah' R package[34] for acquiring data from the Atlas of Living Australia. This capability would allow users to fetch data flexibly, do further analysis and generate custom visualizations as per their needs.

Dissemination of Knowledge Gathered Through Citizen Science

Citizens should not be viewed only as data contributors in the scientific endeavor; they are also the end-users in many situations. While the purpose of a citizen science project may vary (publishing a scientific paper, data repositories, outreach to the public, etc.), knowledge generated through citizen science must find its way back to its contributors.

Citizen science participation can be enhanced by incorporating clear channels of communication and data dissemination (Vohland et al. 2021). This allows access to a wide audience, makes people aware of the project, and keeps them in continued engagement with the project. The traditional means of disseminating scientific knowledge through publication in peer-reviewed journals can often be too technical for the lay public to understand. Involving the public in science is one of the core principles in citizen science, and hence knowledge should also reach the public in a digestible manner. This can be done through activities such as creating data visualizations to communicate results attractively e.g., eBird Status and Trends abundance animations that reveal migratory pathways of birds[35].; writing articles in popular media sources like newspapers, magazines, and online magazines; visual communication of knowledge through art, videos, and graphic design and using social media to disseminate results.

It is worth noting that disseminating knowledge to the public is crucial to ensure long-term participation and collaboration in any citizen science program through various means, targeting multiple stakeholder communities. However, excessive emails or other means of contacting participants can adversely affect and discourage participation.

Data Attribution

Attribution is the act of giving credit to data providers during publication. Author attribution has historically been a tricky issue across disciplines and this has only been accentuated with the advent of big data and data papers with proper guidelines on giving authorship not being stabilized even today (Venkatraman 2010; Escribano et al. 2018). While protocols such as the Science Commons advocate publishing data openly, there is no mention of providing attribution. Authors typically negotiate their order within the author list, assuming that the first author is the most coveted and has led the publication idea. The last typically is the head of the lab and the point of contact (Venkatraman 2010). As the contributor list grows, especially in large collaborative projects, the contribution order becomes less understandable and meaningless. Some journals provide a separate text or list stating individuals' roles and contributions instead of authorship.

There is much ambiguity in citizen science as to who should get attributed and how and whether indi-

---

[32] https://cran.r-project.org/web/packages/rgbif/index.html.
[33] https://github.com/gbif/pygbif.
[34] https://atlasoflivingaustralia.github.io/galah/index.html.
[35] https://ebird.org/science/status-and-trends/abundance-animations.

vidual citizens will be acknowledged in publications. The Joint Declaration of Data Citation Principles (Crosas 2013), states that when cited, there should be '*legal attribution to all contributors to the data, but recognizes that a single style or mechanism of attribution may not be applicable to all data*'.

However, large datasets or data involving many contributors, such as citizen science data, are prone to the issue of 'attribution stacking' where citing every person involved in the generation of the dataset may become unwieldy and difficult to manage. This issue is further magnified when citizen science and other data from multiple projects are combined for further use. Ensuring the correct citation formats are maintained manually or by machines itself becomes challenging. To tackle this, it becomes necessary to allow for 'lightweight attribution mechanisms' (Ball 2011).

In this context, it is also worth considering that citizens may be less likely to be motivated by citation in academic journals as against acknowledgment of their contribution that is visible to their local peers and that projects should support attribution in a way that matters to the citizen scientists. Some sites, such as ebird, provide the option to hide user names and anonymize them. However, in such cases, attribution for the data is not provided to the contributor for apparent reasons. Attribution and user privacy are interlinked, and setting conditions on one of these usually has inverse effects on the other.

Data Policy

Having clear and robust data policies is a means of ensuring that the data collected through citizen science projects are stored, shared, attributed, and utilized ethically. Citizen science project proponents should be mindful of different stakeholders, from contributors to end-users of data and data policies, of the differing rights and responsibilities that each party may possess. While the definition of citizen science is still evolving, it generally encompasses participation from individuals without specific scientific training who participate as volunteers in activities. Such activities may cover the breadth of the data life, including study design, data collection and analysis, and dissemination of results (Guerrini et al. 2018). This information is then used in ways that may or may not be fully understood by volunteers, and so informed consent must be obtained from volunteers on how the data will be used and what credit they will receive for it. Informed consent and refusal are some of the essential components of research ethics that the volunteer willingly gives themselves up for use as a resource (Reiheld and Gay 2019).

Informed consent can be ensured by using easy-to-understand documents with minimal text and ensuring participants have agreed to the project terms. It is advisable to place the documents in a conspicuous place on the portal. These documents are a collection of guidelines that constitute the project's policies that determine how a citizen science project and the users, a website, or a citizen science volunteer may interact or transact. Such documents are usually presented as different types of formalized policy documents (Bowser et al. 2013). These include:

Terms of use - These form the conditions that a user is expected to know and accept before they begin using the portal. It also encompasses guidelines for acceptable behavior between the user and the portal. Terms and conditions may be explicit, requiring the user to accept and consent to the site's terms before proceeding with registration and usage (clickwrap), or it may be implicit, assuming that the user agrees to the terms simply by continued use of the portal (browsewrap). The terms and conditions set out the conditions of usage of the portal, covering aspects along the lifecycle of the data. It indicates the portal's stand on data ownership, data access, reuse, and providing attribution to users or recommended citation policies. Clarity on aspects of data ownership, including any media uploaded by the user, is imperative. Further, the terms need to specify how owners of the site will use the data. It would also need to indicate terms of being contacted for communication regarding outreach or marketing purposes, acceptance of terms and conditions of any third party website linked to the portal (such as YouTube or Google Maps), liability clauses that protect the owner of the portal from any inappropriate content posted on the website by a third party and indemnity clauses against harm caused to any third party from the content of the portal. It would be beneficial to list all the activities that are prohibited on the portal, similar to what is observed in the European Citizen Science portal[36]. Additional terms of use may allow the portal to block a user in case they violate the terms of use. In some countries, the project may need to clarify if it is merely an intermediary where the adminis-

---

[36] https://eu-citizen.science/terms/.

trator does not initiate the transmission by posting the information or select who will be able to view the information or make changes to the information, thereby claiming exemption from liability arising out of the conduct of its users. Otherwise, the portals should safeguard themselves from potential legal liability through clear terms of use for all classes of users with clear contracts.

Legal policies - This would cover information on how the site deals with the legal aspects such as its obligations to national or local laws, liabilities of the project, disclaimers, and waivers. It is best practice to include or link to texts containing specific legal or non-legal documentation.

Privacy policies - This covers information on how and what kind of information the project gathers from participants, including information collected during registration, data upload, and how such information is saved, used, and kept confidential. It would also need to disclose the usage of cookies, whether for functionality within the portal such as for login and role-based permissions or through the usage of features provided by third-party sites such as social media networks or advertising providers.

Privacy Concerns in Citizen Science

Much attention has been paid to privacy concerns about citizen science data involving medical and genetic information participants. However, data obtained as part of biodiversity inventories or ecological phenomena may also require close perusal for violations of the privacy rights of participants and federal laws that prevent sharing of sensitive information that could jeopardize the safety of endangered species.

When collecting biodiversity-related information, privacy breaches can occur at two levels:

● Personal information of the observer
● Georeferenced data associated with a species record being contributed

Most projects collect basic personal information of participants, such as names, email IDs, and addresses to keep them informed of the progress of the project. Through these mediums citizen science projects wittingly or unwittingly end up with personally identifiable information (PII) of participants in their projects. Additionally, smartphones equipped with tools that utilize cameras, audio-recorders, and location-capturing applications to capture biodiversity-related information often end up revealing PII (Cartwright 2016), that may reveal near real-time information about their locations, patterns of daily or weekend travel, types of phones used, etc.

Geo-locations of species, commonly required by biodiversity inventories, may reveal sensitive information related to endangered species. Information on the location of species could lead to poaching, unethical collection, or disturbance through excessive attention from nature enthusiasts and photographers. This is particularly important when dealing with range-restricted, endangered, frequently traded, or breeding populations of uncommon species.

Although participants are generally aware of these issues while contributing data (Bowser et al. 2013), it is still imperative to get informed consent and brief them on the terms of service employed by the project. A recent study showed that 51% of projects that did not focus exclusively on people data often overlooked the fact that they were still collecting PII (Cooper et al. 2019). The Personal Genome Project[37] (PGP) has been globally acclaimed for its approach to informed consent that transcends traditional boundaries. The project proponents ensure that all participants pass an examination that tests their knowledge of genomic science and privacy issues. After that, they sign access to their personal and genomic data for the project (Angrist 2009).

The US and the EU have implemented legal provisions to safeguard the privacy of citizen science contributors. Under the US privacy laws, citizen science project managers are mandated to make users aware of their rights and are provided with the Privacy Act Statement. Under the Children's Online Privacy Protection Rule, collection of personal information of children below the age of 13 is illegal; and the Freedom of Information and the Privacy Acts require cleansing all personal information of participants from data collected by projects supported by the federal government before such databases are made public. In the EU, the General Data Protection Regulation (GDPR) seeks the right to be informed, the right of access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object and rights around automated decision making and profiling. Under GDPR, project managers are mandated to get fully informed consent from contributors and inform them of how data contributed by them would be used.

---

[37] https://www.personalgenomes.org/.

Such existing and upcoming legal provisions have potential implications for the privacy of participants in Citizen Science portals (Ganzevoort et al. 2017).

Conclusions

Including the above considerations, every project has to consider its unique situation in terms of biodiversity such as between the explored and unexplored, the documented and undocumented, conservation threats, along with specific challenges to discover, document, and disseminate. Each design is significant on its own, reflecting the needs of the socio-ecological system that information technology has to integrate into and co-evolve.

Although citizen science is rapidly gaining popularity, data generated through it still deals with a perceived "image problem" regarding data quality. While the debate around this issue rages, several studies have indicated that with the appropriate data quality checks in place, citizen science data is no less reliable than data gathered by experts (Jordan et al. 2012; Ganzevoort et al. 2017). The sole objective of a citizen science project is not necessarily data. Through the duration of the project, it builds the capacity of its participants and inculcates the spirit of scientific endeavor and discovery while also sensitizing participants towards species and habitat conservation, creating a sense of stewardship towards nature.

Another challenge with citizen science is ensuring sustained participation both from citizens and scientists to help validate the data (Irwin 2018). From this perspective, imposing too much rigor in data collection and quality can reduce inclusivity and lead to reduced participation. As one of citizen science's objectives involves broader participation, holding participants to unrealistic scientific standards could mean missing out on opportunities to "fully engage with people in the core objective of discovery" (Lukyanenko et al. 2016).

Multiple competing citizen science initiatives operating within the same region and data sharing between various sources often result in duplication of data contributed in multiple places. This issue will need attention and effort to identify and de-duplicate. Global aggregators such as GBIF are already investing effort in algorithms to identify potentially related records and cluster them. Identifying individual contributors across portals such as through an ORCID id can also help in these efforts, although this is still not widely used beyond the academic community yet.

To conform to the expectations of its varied user bases, citizen science has to meet the dual objectives of providing high-quality summarized data to the general public as well as spatially, temporally, and taxonomically explicit data to the research community. These have to be achieved while protecting sensitive information and providing privacy protection. Achieving these objectives requires significant investment in technology solutions, clear data policies, and transparency. Anhalt-Depies et al. (2019), give a set of recommendations that may be apt to cater to data quality, privacy, transparency, and trust in citizen science. These include constant communication and consultation with stakeholders, addressing volunteer needs on aspects such as data sharing and user privacy through clear policy documents that evolve through iterative evaluation based on user feedback. Among other resources, we refer readers to the 10 Principles of Citizen Science developed by the European Citizen Science Association, which set out the key principles that underlie good practice in citizen science[38].

In the Indian context, it would be ideal for envisaging a directory of citizen science projects and a repository for citizen science projects, which could allow design, host, store, and archive initiatives. This is necessitated by the nature of present-day data infrastructures, which are stretched to provide the full set of features for citizen science practitioners to engage through all the stages of the data lifecycle. Many act as platforms for data collection, organization, and aggregation but for various reasons focus less on providing tools to analyze collected data by citizen science practitioners. Given the immense potential to contribute to biodiversity monitoring at different scales, a culture of integration covering various tenets of biodiversity information, technical design, and stakeholder networks needs to be promoted (Kühl et al. 2020). This is truer for small, focused, and independent citizen science projects for which there is a dire need in a mega-diverse country like India. Technology and data infrastructures need to evolve in a direction where modular, decentralized, and federated architectures are imagined and attempted. Such architectures will help address the spatial, temporal, and taxon bias and empower communities in sensitive socio-ecological systems to participate in biodiversity conservation effectively. Such infrastructure could help transform data infrastructure into knowledge infrastructures, helping enhance

---

[38] https://eu-citizen.science/about/.

the biodiversity knowledge commons and shape policy and practice.

Acknowledgments

Competing Interests

The authors have declared that no competing interests exist.

Literature Cited

Angrist, M. 2009. Eyes wide open: the personal genome project, citizen science and veracity in informed consent. Pers. Med. 6:691–699.

Anhalt-Depies, C., J. L. Stenglein, B. Zuckerberg, P. A. Townsend, and A. R. Rissman. 2019. Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. Biol. Conserv. 238:108195.

Assumpção, T. H., I. Popescu, A. Jonoski, and D. P. Solomatine. 2018. Citizen observations contributing to flood modelling: opportunities and challenges. Hydrol. Earth Syst. Sci. 22:1473–1489.

August, T. A., O. L. Pescott, A. Joly, and P. Bonnet. 2020. AI naturalists might hold the key to unlocking biodiversity data in social media imagery. Patterns 1:100116.

Baker, E., J. P. Drury, J. Judge, D. B. Roy, G. C. Smith, and P. A. Stephens. 2021. The verification of ecological citizen science data: current approaches and future possibilities. Citiz. Sci. Theory Pract. 6:12. Ubiquity Press.

Balázs, B., P. Mooney, E. Nováková, L. Bastin, and J. Jokar Arsanjani. 2021. Data quality in citizen science. Pp. 139–157 in K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, and K. Wagenknecht, eds. The Science of Citizen Science. Springer International Publishing, Cham.

Ball, A. 2011. How to License Research Data. Digital Curation Centre, Edinburgh.

Barve, V. 2014. Discovering and developing primary biodiversity data from social networking sites: A novel approach. Ecol. Inform. 24:194–199.

Boakes, E. H., G. Gliozzo, V. Seymour, M. Harvey, C. Smith, D. B. Roy, and M. Haklay. 2016. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. Sci. Rep. 6:33051.

Bonney, R., C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk. 2009. Citizen science: A developing tool for expanding science knowledge and scientific literacy. BioScience 59:977–984.

Bowser, A., C. Cooper, A. de Sherbinin, A. Wiggins, P. Brenton, T.-R. Chuang, E. Faustman, M. (Muki) Haklay, and M. Meloche. 2020. Still in need of norms: The state of the data in citizen science. Citiz. Sci. Theory Pract. 5:18.

Bowser, A., A. Wiggins, and R. D. Stevenson. 2013. Data Policies for Public Participation in Scientific Research: A Primer. DataONE Public Participation in Scientific Research Working Group.

Brenton, P., S. von Gavel, E. Vogel, and M.-E. Lecoq. 2018. Technology infrastructure for citizen science. Pp. 63–80 in Citizen Science: Innovation in Open Science, Society and Policy. UCL Press.

Callaghan, C. T., A. G. B. Poore, M. Hofmann, C. J. Roberts, and H. M. Pereira. 2021. Large-bodied birds are over-represented in unstructured citizen science data. Sci. Rep. 11:19073.

Callaghan, C. T., J. J. L. Rowley, W. K. Cornwell, A. G. B. Poore, and R. E. Major. 2019. Improving big citizen science data: Moving beyond haphazard sampling. PLOS Biol. 17:e3000357. Public Library of Science.

Campbell, D. L., A. E. Thessen, and L. Ries. 2020. A novel curation system to facilitate data integration across regional citizen science survey programs. PeerJ 8:e9219. PeerJ Inc.

Carroll, S. R., E. Herczog, M. Hudson, K. Russell, and S. Stall. 2021. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci. Data 8:108.

Cartwright, J. 2016. Technology: Smartphone science. Nature 531:669–671.

Chapman, A. 2005a. Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility.

Chapman, A. 2005b. Principles of Data Quality. Global Biodiversity Information Facility.

Chen, D., and C. P. Gomes. 2019. Bias reduction via end-to-end shift learning: Application to citizen science. Proc. AAAI Conf. Artif. Intell. 33:493–500.

Cooper, C., L. Larson, K. K. Holland, R. Gibson, D. Farnham, D. Hsueh, P. Culligan, and W. McGillis. 2017. Contrasting the views and actions of data collectors and data consumers in a volunteer water quality monitoring project: Implications for project design and management. Citiz. Sci. Theory Pract. 2:8. Ubiquity Press.

Cooper, C., L. Shanley, T. Scassa, and E. Vayena. 2019. Project categories to guide institutional oversight of responsible conduct of scientists leading citizen science in the United States. Citiz. Sci. Theory Pract. 4:7.

CoreTrustSeal Standards And Certification Board. 2019. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022. doi: 10.5281/ZENODO.3638211. Zenodo.

Courter, J. R., R. J. Johnson, C. M. Stuyck, B. A. Lang, and E. W. Kaiser. 2013. Weekend bias in Citizen Science data reporting: implications for phenology studies. Int. J. Biometeorol. 57:715–720.

Crosas, M. 2013. Joint Declaration of Data Citation Principles - FINAL. FORCE11.

de Sherbinin, A., A. Bowser, T.-R. Chuang, C. Cooper, F. Danielsen, R. Edmunds, P. Elias, E. Faustman, C. Hultquist, R. Mondardini, I. Popescu, A. Shonowo, and K. Sivakumar. 2021. The critical importance of citizen science data. Front. Clim. 3. Frontiers.

Devictor, V., R. J. Whittaker, and C. Beltrame. 2010. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography: Citizen science and conservation biogeography. Divers. Distrib. 16:354–362.

Dosemagen, S., and A. J. Parker. 2019. Citizen science across a spectrum: Broadening the impact of citizen science and community science. Sci. Technol. Stud. 32.

Downs, R. R., H. K. Ramapriyan, G. Peng, and Y. Wei. 2021. Perspectives on citizen science data quality. Front. Clim. 3. Frontiers.

Escribano, N., D. Galicia, and A. H. Ariño. 2018. The tragedy of the biodiversity data commons: a data impediment creeping nigher? Database J. Biol. Databases Curation 2018:bay033.

Falk, S., G. Foster, R. Comont, J. Conroy, H. Bostock, A. Salisbury, D. Kilbey, J. Bennett, and B. Smith. 2019. Evaluating the ability of citizen scientists to identify bumblebee *Bombus* species. PLOS ONE 14:e0218614. Public Library of Science.

Faundeen, J., T. E. Burley, J. A. Carlino, D. L. Govoni, H. S. Henkel, S. L. Holl, V. B. Hutchison, E. Martín, E. T. Montgomery, C. Ladino, S. Tessler, and L. S. Zolly. 2014. The United States Geological Survey Science Data Lifecycle Model. U.S. Geological Survey, Reston, VA.

Ganzevoort, W., R. J. G. van den Born, W. Halffman, and S. Turnhout. 2017. Sharing biodiversity data: citizen scientists' concerns and motivations. Biodivers. Conserv. 26:2821–2837.

Geldmann, J., J. Heilmann-Clausen, T. E. Holm, I. Levinsky, B. Markussen, K. Olsen, C. Rahbek, and A. P. Tøttrup. 2016. What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Divers. Distrib. 22:1139–1149.

Gonsamo, A., and P. D'Odorico. 2014. Citizen science: best practices to remove observer bias in trend analysis. Int. J. Biometeorol. 58:2159–2163.

Groom, Q., L. Weatherdon, and I. R. Geijzendorffer. 2017. Is citizen science an open science in the case of biodiversity observations? J. Appl. Ecol. 54:612–617.

Guerrini, C. J., M. Lewellyn, M. A. Majumder, M. Trejo, I. Canfield, and A. L. McGuire. 2019. Donors, authors, and owners: how is genomic citizen science addressing interests in research outputs? BMC Med. Ethics 20:84.

Guerrini, C. J., M. A. Majumder, M. J. Lewellyn, and A. L. McGuire. 2018. Policy for citizen science. Science 361:134–136.

Hagedorn, G., D. Mietchen, R. Morris, D. Agosti, L. Penev, W. Berendsohn, and D. Hobern. 2011. Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. ZooKeys 150:127–149. Pensoft Publishers.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2013. Big data and the future of ecology. Front. Ecol. Environ. 11:156–162.

Iacovides, I., C. Jennett, C. Cornish-Trestrail, and A. L. Cox. 2013. Do games attract or sustain engagement in citizen science? a study of volunteer motivations. Pp. 1101–1106 *in* CHI '13 Extended Abstracts on Human Factors in Computing Systems. Association for Computing Machinery, Paris, France.

Irwin, A. 2018. No PhDs needed: how citizen science is transforming research. Nature 562:480–482.

Johnston, A., D. Fink, W. M. Hochachka, and S. Kelling. 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol. Evol. 9:88–97.

Jordan, R. C., W. R. Brooks, D. V. Howe, and J. G. Ehrenfeld. 2012. Evaluating the performance of volunteers in mapping invasive plants in public conservation lands. Environ. Manage. 49:425–434.

Kelling, S., D. Fink, F. A. La Sorte, A. Johnston, N. E. Bruns, and W. M. Hochachka. 2015. Taking a 'Big Data' approach to data quality in a citizen science project. Ambio 44:601–611.

Kelling, S., J. Gerbracht, D. Fink, C. Lagoze, W.-K. Wong, J. Yu, T. Damoulas, and C. Gomes. 2012. ebird: A human/computer learning network for biodiversity conservation and research. P. *in* Twenty-Fourth IAAI Conference.

Kelling, S., A. Johnston, A. Bonn, D. Fink, V. Ruiz-Gutierrez, R. Bonney, M. Fernandez, W. M. Hochachka, R. Julliard, R. Kraemer, and R. Guralnick. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. BioScience 69:170–179.

Kennett, R., F. Danielsen, and K. M. Silvius. 2015. Citizen science is not enough on its own. Nature 521:161–161.

Kimura, A. H., and A. Kinchy. 2016. Citizen science: probing the virtues and contexts of participatory research. Engag. Sci. Technol. Soc. 2:331–361.

Kobori, H., J. L. Dickinson, I. Washitani, R. Sakurai, T. Amano, N. Komatsu, W. Kitamura, S. Takagawa, K. Koyama, T.

Ogawara, and A. J. Miller-Rushing. 2016. Citizen science: a new approach to advance ecology, education, and conservation. Ecol. Res. 31:1–19.

Kodali, S. 2017. Not Open or Accountable: The Government Open Data Use License Is Flawed.[39]

König, C., P. Weigelt, J. Schrader, A. Taylor, J. Kattge, and H. Kreft. 2019. Biodiversity data integration—the significance of data resolution and domain. PLOS Biol. 17:e3000183.

Kosmala, M., A. Wiggins, A. Swanson, and B. Simmons. 2016. Assessing data quality in citizen science. Front. Ecol. Environ. 14:551–560.

Kühl, H. S., D. E. Bowler, L. Bösch, H. Bruelheide, J. Dauber, David. Eichenberg, N. Eisenhauer, N. Fernández, C. A. Guerra, K. Henle, I. Herbinger, N. J. B. Isaac, F. Jansen, B. König-Ries, I. Kühn, E. B. Nilsen, G. Pe'er, A. Richter, R. Schulte, J. Settele, N. M. van Dam, M. Voigt, W. J. Wägele, C. Wirth, and A. Bonn. 2020. Effective biodiversity monitoring needs a culture of integration. One Earth 3:462–474.

Lamba, A., P. Cassey, R. R. Segaran, and L. P. Koh. 2019. Deep learning for environmental conservation. Curr. Biol. 29:R977–R982.

Lemmens, R., V. Antoniou, P. Hummer, and C. Potsiou. 2021. Citizen science in the digital world of apps. Pp. 461–474 *in* K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, and K. Wagenknecht, eds. The Science of Citizen Science. Springer International Publishing, Cham.

Lukyanenko, R., J. Parsons, and Y. F. Wiersma. 2016. Emerging problems of data quality in citizen science. Conserv. Biol. 30:447–449.

McClure, E. C., M. Sievers, C. J. Brown, C. A. Buelow, E. M. Ditria, M. A. Hayes, R. M. Pearson, V. J. D. Tulloch, R. K. F. Unsworth, and R. M. Connolly. 2020. Artificial intelligence meets citizen science to supercharge ecological monitoring. Patterns 1:100109.

McQuillan, D. 2014. The countercultural potential of citizen science. MC J. 17.

Paleco, C., S. G. Peter, N. S. Seoane, J. Kaufmann, and P. Argyri. 2021. Inclusiveness and diversity in citizen science. P. 529 *in* K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, and K. Wagenknecht, eds. The Science of Citizen Science. Springer.

Pearce-Higgins, J. W., S. R. Baillie, K. Boughey, N. A. D. Bourn, R. P. B. Foppen, S. Gillings, R. D. Gregory, T. Hunt, F. Jiguet, A. Lehikoinen, A. J. Musgrove, R. A. Robinson, D. B. Roy, G. M. Siriwardena, K. J. Walker, and J. D. Wilson. 2018. Overcoming the challenges of public data archiving for citizen science biodiversity recording and monitoring schemes. J. Appl. Ecol. 55:2544–2551.

Ponti, M., T. Hillman, C. Kullenberg, and D. Kasperowski. 2018. Getting it right or being top rank: Games in citizen science. Citiz. Sci. Theory Pract. 3:1.

Ratnieks, F. L. W., F. Schrell, R. C. Sheppard, E. Brown, O. E. Bristow, and M. Garbuzov. 2016. Data reliability in citizen science: learning curve and the effects of training method, volunteer background and experience on identification accuracy of insects visiting ivy flowers. Methods Ecol. Evol. 7:1226–1235.

Reiheld, A., and P. L. Gay. 2019. Coercion, consent, and participation in citizen science. ArXiv190713061 Phys.

Schuttler, S. G., R. S. Sears, I. Orendain, R. Khot, D. Rubenstein, N. Rubenstein, R. R. Dunn, E. Baird, K. Kandros, T. O'Brien, and R. Kays. 2019. Citizen science in schools: Students collect valuable mammal data for science, conservation, and community engagement. BioScience 69:69–79.

Sekhsaria, P., and N. Thayyil. 2019. Citizen Science in ecology in India - an initial mapping and analysis. DST Centre for Policy Research, Indian Institute of Technology Delhi.

Steen, V. A., C. S. Elphick, and M. W. Tingley. 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. Divers. Distrib. 25:1857–1869.

Sullivan, B. L., J. L. Aycrigg, J. H. Barry, R. E. Bonney, N. Bruns, C. B. Cooper, T. Damoulas, A. A. Dhondt, T. Dietterich, A. Farnsworth, D. Fink, J. W. Fitzpatrick, T. Fredericks, J. Gerbracht, C. Gomes, W. M. Hochachka, M. J. Iliff, C. Lagoze, F. A. La Sorte, M. Merrifield, W. Morris, T. B. Phillips, M. Reynolds, A. D. Rodewald, K. V. Rosenberg, N. M. Trautmann, A. Wiggins, D. W. Winkler, W.-K. Wong, C. L. Wood, J. Yu, and S. Kelling. 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biol. Conserv. 169:31–40.

Tiago, P., A. Ceia-Hasse, T. A. Marques, C. Capinha, and H. M. Pereira. 2017. Spatial distribution of citizen science casuistic observations for different taxonomic groups. Sci. Rep. 7:12832.

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7.

Turnhout, E., and S. Boonman-Berson. 2011. Databases, Scaling Practices, and the Globalization of Biodiversity. Ecol. Soc. 16. The Resilience Alliance.

Veeckman, C., Talboom, S., Gijsel, L., Devoghel, H., and Duerinckx, A. 2019. Communication in Citizen Science. A practical guide to communication and engagement in citizen science. SCIVIL; Leuven, Belgium.

Venkatraman, V. 2010. Conventions of Scientific Authorship. Sci. AAAS, doi: https://www.science.org/careers/2010/04/conventions-scientific-authorship.

Vohland, K., A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, and K. Wagenknecht (eds). 2021. The Science of Citizen Science. Springer International Publishing, Cham.

Walker, D., C. McCord, N. Stradiotto, M. Zhou, and D. Singh. 2016. Citizen's Guide to Open Data.

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

---

[39] https://thewire.in/102905/open-data-licensegovernment/.

Wilkinson, M. D., M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3:160018.

Williams, J., C. Chapman, D. G. Leibovici, G. Loïs, A. Matheus, A. Oggioni, S. Schade, L. See, and P. P. L. van Genuchten. 2018. Maximising the impact and reuse of citizen science data. Pp. 321–336 *in* Citizen Science: Innovation in Open Science, Society and Policy. UCL Press.