DATA AND CITIZEN SCIENCE IN INDIA

A PRACTITIONER'S TOOLKIT

with support from

Office of the Principal Scientific Adviser to the Government of India and Rohini Nilekani Philanthropies

Design & editing credits:

Aditi Sajwan: Toolkit Designing
Digangana & Dhruv Gangadharan: Editing

Cover photo credits:

Thomas Vattakaven

How to cite the toolkit:

Vattakaven, Thomas, Vijay Barve, Geetha Ramaswami, Priya Singh, Suneha Jagannathan, and Balasubramanian Dhandapani. 2022. "Data and Citizen Science in India: A Practioner's Toolkit". CitSci India Conference. https://doi.org/10.5281/zenodo.7328043

CONTENTS

Preface

The Biodiversity Collaborative, a consortium of leading Indian biodiversity science and conservation organisations*, proposed a National Mission on Biodiversity and Human Well-Being to the Prime Minister's Science, Technology and Innovation Advisory Council (PM-STIAC) in 2018. In August 2019 the Ministry of Environment, Forest and Climate Change (MoEFCC) designated the National Biodiversity Authority (NBA) as the nodal agency to work with the Biodiversity Collaborative to develop a Detailed Project Report for the Mission.

As per the Mission Statement, "The Mission aims to strengthen the science of restoring, conserving, and sustainably utilising India's natural heritage; embed biodiversity as a key consideration in all developmental planning, particularly in agriculture, ecosystem services, health, bio-economy, and climate-change mitigation; establish a citizen and policy oriented biodiversity information system; and enhance capacity across all sectors for realisation of India's national biodiversity targets, United Nations Sustainable Development Goals (https://sdgs.un.org/goals) and the post-2020 Global Biodiversity Framework" (UNEP, 2021; Bawa et al., 2020).

Programme 7 of the National Mission focuses on biodiversity, capacity building and outreach, seeking to,

1. build capacity in biodiversity science through training programmes,
2. provide resources to new and existing citizen engagement projects, and
3. mainstream the understanding of biodiversity into India's consciousness through communication, and outreach to government, private entrepreneurs, school and college students, media professionals and the public at large.

The conference on Citizen Science for Biodiversity in India (https://citsci-india.org/) is hosted as part of the Preparatory Phase Project of the National Mission. The conference is a virtual meeting of various stakeholders such as practitioners of citizen science, researchers, educators, students, policy makers, and individual contributors, who actively engage in citizen science. 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 in India. Two prominent topics that surfaced during these discussions were the importance of diversity and inclusion, and that of citizen science data. As a result, two working groups were formed to catalyse the development of a toolkit for practitioners, project proponents and the larger community.

Following global trends, citizen science efforts involving biodiversity in India have rapidly been gaining pace over the past few years. With more and more large and small-scale citizen science projects being launched in India each year, voluminous data are being generated on various aspects of biodiversity. However, this also raises a number of issues related to data such as ownership, accessibility, attribution, storage, interoperability,

quality, and others.

This working group on Citizen Science Data was tasked with the aim of identifying major aspects related to data on which project proponents should have clear procedures and policies, while simultaneously remaining mindful of the tenets of diversity and inclusion, which remain critical at every stage of data generation in a citizen science initiative. To put together this document we have surveyed existing global practises and standards, and described various options that projects could adopt, with some guidance about benefits and costs to each option. This document is intended to form a toolkit for citizen science practitioners in India, who seek to make informed decisions on various aspects of data.

* The current members of the Biodiversity Collaborative are: 1. Ashoka Trust for Research in Ecology and the Environment (ATREE), 2. Echo Network, 3. Indian Institute of Science (IISc), 4. Metastring Foundation, 5. National Centre for Biological Sciences – Tata Institute of Fundamental Research (NCBS-TIFR), 6. Nature Conservation Foundation (NCF), 7. Srishti Manipal Institute of Art, Design and Technology, 8. TERI School of Advanced Studies (TERI-SAS), 9. The University of Trans-Disciplinary Health Sciences and Technology (TDU), and 10. University of Agricultural Sciences - Bangalore (UAS-B).

Acknowledgments

We are grateful to Suhel Quader, Pankaj Sekhsaria, Farida Tampal, Shannon Olsson and Prabhakar Rajagopal, all members of the Organising Committee of CitSci India, for conceptualising the idea of this working group and providing guidance and feedback at various stages of developing this toolkit. We thank Akshata Pradhan for facilitating the functioning of the working group. Mridula Vijairaghavan, Sushmitha Viswanathan, and Shyama Kuriakose from Wildlife Conservation Society-India vetted the legal components of this document for accuracy and provided additional inputs. We thank them for their time, and for enhancing the quality of this document. The need for this toolkit was drawn from discussions that were held during the first CitSci India 2020 conference; hence, we acknowledge the role of participants at the conference and for sharing their thoughts. We are most grateful to Townsend Peterson, Naveen Thayyil and Shannon Olsson for reviewing this document and providing useful feedback, most of which has been incorporated here.

About the Authors - Working Group on Citizen Science Data

The Working Group on Citizen Science Data was put together after the CitSci India 2020 conference in order to develop a toolkit on data for the wider citizen science community. Conceived by the Organising Committee, the purpose of the working group is to generate a practical toolkit for citizen science practitioners across the country on best practises to manage and handle data. Members of the working group are:

Thomas Vattakaven

Thomas is a Senior Application Scientist at Strand Life Sciences, where he coordinates the development and activities of the India Biodiversity Portal. He obtained his PhD in Microbiology from the University of the West of England, Bristol, UK in 2010. He has a keen interest in biodiversity, biodiversity informatics and engaging the public in citizen science and its role in aggregating open access biodiversity information. He is a wildlife photography enthusiast, with a special interest in birds and macro photography.

Priya Singh

Priya works as an independent researcher. She is interested in carnivore biology and conservation, with a focus on wild felids in north-eastern India, and striped hyenas in north-western India. In 2008, she graduated with a degree in Wildlife Biology and Conservation from the Post-graduate programme offered by WCS-India and the National Centre for Biological Sciences (TIFR), Bangalore. Currently, she works on carnivore communities in the Indo-Myanmar Biodiversity Hotspot, and is a member of the Wild Canids-India Project, a citizen science initiative aimed at making ecological and conservation assessments of canids and the hyena in India.

Balasubramanian Dhandapani

Balu is a research engineer at the French Institute of Pondicherry where the mainstay of his research is critically situating technology practises and communities with technology as a lens in exploring its complex relationship with society and environment. He has been coordinating the biodiversity informatics program which strives to build open access, collaborative information systems for biodiversity.

Geetha Ramaswami

Geetha has been Programme Manager for SeasonWatch, a citizen science project looking at the effects of seasons on tree phenology, since March 2018. Children and adults across India contribute data on the leaf-out, flowering and fruiting of 135 common tree species to this project via pen-paper, an Android app and the project website. In the past, she has been a post-doctoral research associate with the Nature Conservation Foundation, and has a PhD in the ecology of an invasive plant – Lantana camara – from the Centre for Ecological Sciences at the Indian Institute of Science.

Vijay Barve

Vijay has been a nature enthusiast from childhood, interested in birds and insects. He has a Master's degree in computer science and PhD in Geography. He has continued his interest in biodiversity and now works as researcher in biodiversity informatics and citizen science. In 2001, he initiated DiversityIndia, a citizen science group contributing to biodiversity documentation in India and beyond.

Suneha Jagannathan

Suneha is an independent marine biologist working on a range of themes – marine habitat restoration, marine education and citizen science along the East Coast of India. She regularly consults with tourism companies to incorporate conservation ethos in sustainable tourism and adventure sport education. As a Research Affiliate with Dakshin Foundation, she coordinates Reeflog, a marine citizen science program aimed at recreational SCUBA divers. With a Master's in Tropical Biodiversity from the Erasmus Mundus program, she works in the intersection of conservation, tourism and coastal communities.

1. The Scope of This Toolkit

The concept of 'citizen science' has been in a steady state of evolution, with no single unified definition (Kreitmair & Magnus, 2019). The scope and means of participation in citizen science initiatives have also evolved and adapted over time with advances in technology, participatory mediums and the extent of volunteer involvement. For the purpose of 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 bio-specimens to be collected from them for analysis by others". At this juncture it is notable that there is a move, especially in some regions, to replace the usage of the word "citizen", with "community", and many organisations rebranding their programs as community science, to be more inclusive (Cooper et al., 2021).

Citizen science is widely employed across disciplines to engage volunteers in a variety of tasks. Biological citizen science projects and particularly biodiversity-related citizen science remain the most dominant as well as the most rapidly growing theme within citizen science (Follett & Strezov, 2015). For the context of this paper, we largely limit our reference to citizen science projects associated with biodiversity, that at least partially utilise online participatory mediums with databases and servers that make data and their products accessible online.

Just as in citizen science, the definition of ''data'' varies widely depending upon subjective or objective interpretations and varying with the domain and approach such as from human-centric or computational perspectives (Zins, 2007). For this toolkit, we adopt a broad definition that "data are facts that are the result of observation or measurement" (Landry & Rusk, 1970). In this context, we limit our domain within the field of biodiversity.

1.1 Types of Citizen Science Projects

Citizen science initiatives vary extensively in their aims and objectives – and across disciplines, and in citizen engagement. An attempt to classify common citizen science projects in India can be undertaken based on the following broad parameters:

Research Question/Focus

Although citizen science has traditionally been used to address targeted research questions and hence involve specified protocols, the advent of online mediums and the ability to crowdsource content has paved the way for more open-ended platforms which

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 mostly have little or no restriction based on the taxonomic focus (e.g. India Biodiversity Portal), while targeted projects often tend to focus on selected or a single taxonomic group or species (e.g. Biodiversity Atlas - India, Bird Count India, Wild Canids-India Project, Marine Life of Mumbai).

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 objectives of a project. Project initiators play an important role in defining nuances of a project, and hence determining the end goals which in-turn influence 'the political authority of science' (Kimura & Kinchy, 2016). Similarly, the composition and training of citizen science initiators varies across projects. For the purpose of this document, we highlight different types of public-scientist collaborations that qualify as citizen science engagements based on Veeckman et al. (2019):

A. Crowdsourcing - Volunteers remain passive while contributing time and device only, e.g. providing access to personal computers for data processing
B. Distributed intelligence - Volunteers are involved with simple interpretations or categorising material from gathered data
C. Participatory science - Volunteers play an important role by defining a problem, collecting data and assisting scientists in analysing the data. However, the interpretation and analytical sections of the project are primarily handled by scientists.
D. 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.
E. Contributory project - Projects of this nature are commonly encountered in the field of ecology, where volunteers are invited to contribute data, while scientists decide the research focus of the study, and analyse and interpret data. Projects of this nature may be led by trained scientists working in mainstream scientific institutions aiming to address a scientific research question that they may have framed.
F. Collaborative project - These are 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.
G. Co-created project - These projects are primarily aimed at influencing public policy or have an educational agenda. Citizen participants identify a set of questions, answers to which are thereafter pursued in consultation with scientists on the project.

In ecology, contributory models of citizen science are most common, while the potential for collaborative and co-created projects remains poorly tapped, even though they are predicted to have great potential in influencing policy decisions.

Medium of Participation

Currently two distinct channels allow establishing a citizen science initiative. These are:
A. Independent platforms via web or smart phone-app based methods using protocols specifically built 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, CitSci.org. 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 issues of 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 gathered through such mediums are largely not structured by default. Nor are they controlled environments with binding data policies or licences. Most of these are still emergent and although there is potential to crowdsource content using these increasingly popular mediums, due to their free-form nature of interaction, much effort will need to be put in to extract, curate and cleanse the content before it can be used as meaningful, structured data.

Mode of Survey

Citizen science may be carried out in a conventional scientific framework with a standardised 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 standardised 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.

1.2 Motivations of Participants and Benefits of Citizen Science

Participants volunteer time and effort for citizen science projects and can have various motivations to do so. For practitioners intending to set up a new citizen science project it is crucial to understand some key motivations of potential participants. Veeckman et al. (2019) list several reasons that drive participant motivation. These include:

A. Collective motivation - participant identifies with the objectives of the project.
B. Reward based motivation - participant aims to build a reputation or make friends in the process of engaging with the project. Incentivisation such as prizes offered to top contributors or competing to be placed higher on leaderboards also count under this.

C. Norm-related motivation - hoping to generate a positive response from friends and colleagues.
D. Collective identification - identifies with the group and its objectives.
E. Hedonistic/intrinsic motivation - the project makes a participant feel good about contributing.

Citizen science is a two-way street, where there is benefit both for the participants as well as the practitioners. In general, some broad but salient benefits of citizen science projects include:
A. Involving members of the public as contributors or co-creators in the scientific process of identifying questions and finding solutions pertaining to society and the environment, and in the process enhancing the democratisation of science, and expanding its reach.
B. Raising public awareness and enhancing public understanding of science, and the relationship between science and society.
C. Enabling means to collect data at large spatial and temporal scales using limited resources via access to digital infrastructure such as phones, internet, camera-traps, etc.
D. Allowing the detection of rare events across large spatial and temporal scales, which would normally be difficult to survey.
E. Promoting the concept of open data without compromising privacy rights of participants.
F. Generating high-quality, easy to comprehend, visualisations.
G. Building stakeholder capacity.
H. Helping generate scientific information on environmental issues such as climate change, biodiversity loss, unsustainable environmental practises, etc.
I. Generating high quality data which can feed into critical policy advice.

Motivations of contributors and general benefits notwithstanding, citizen science projects are eventually recognised by the data contributed to them, and the uses that these data can be put to. Hence, data emerges as a central facet to citizen science endeavours necessitating a comprehensive understanding of all aspects related to it. In the following sections, we describe the life cycle of this data and the considerations that need to be made by citizen science projects before, during and after data collection.

1.3 Citizen Science Data and Life Cycle

Like most other scientific data, citizen science data also follows the general data life cycle. There are multiple models with minor variations that have been proposed to capture the pathway of data generation within a citizen science project. However, for the purpose of this paper we have chosen to adopt the high-level Science Data Lifecycle Model (SDLM) (Faundeen et al., 2014), which was developed by the U.S. Geological Survey, to illustrate how data management activities relate to citizen science data workflows, and to recommend actions and activities at each stage of the model.

The SDLM consists of primary model elements that proceed sequentially, and cross-cutting elements that are performed continuously across all stages of the life cycle.

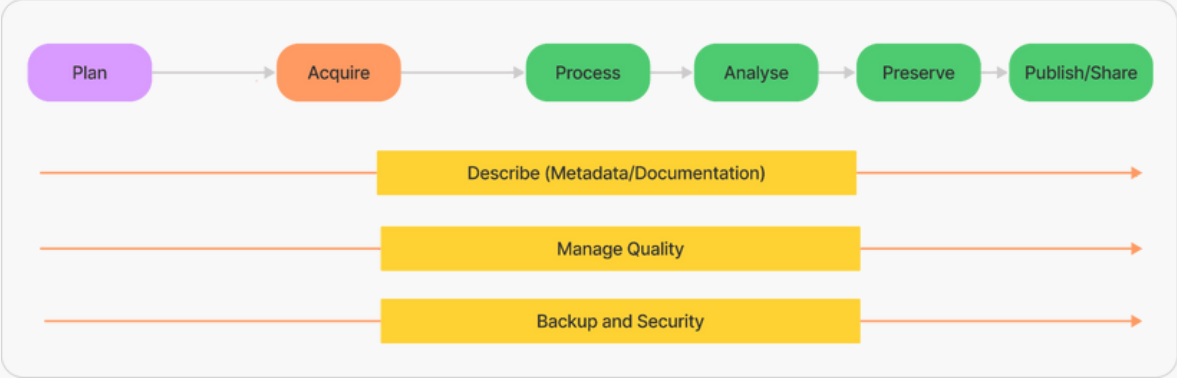

Figure 1. Representation of the high-level Science Data Lifecycle Model (SDLM) (Faundeen et al., 2014) which was developed by the U.S. Geological Survey, used to structure this toolkit.

Primary Model Elements

Primary Model Elements	Description
Plan	The planning stage involves making considerations for handling all data generated by the project. This involves an assessment of objectives, resources required, and intended project outcomes, at each stage of the data life cycle.
Acquire	The acquiring stage represents tasks involved in collection of new data or reuse of existing data. In this phase, the project needs to evaluate workflows that ensure provenance and integrity of data.
Process	Activities associated with preparation of new or previously collected data inputs
Preserve	This phase comprises all activities regarding data storage for long term accessibility to enable reuse of data in the future.
Publish/Share	Peer-reviewed publications allow for traditional means of knowledge dissemination. Popular mediums of publication can include websites, data catalogues, data products and social media.

Table 1.1. Description of primary model elements as defined by Faundeen et al., 2014.

Cross Cutting Model Element	Description
Describe	Documenting data through the description of metadata throughout the life cycle allows other researchers to understand the data collected, thereby allowing for replication to test the validity of scientific principles. This enables the data to be useful for research in the future.
Manage quality	A quality assurance protocol needs to be formulated, to monitor data quality at every stage of the data life cycle. The protocol may need to be modified or adjusted at various stages to ensure that the protocols perform as planned.
Backup and secure	Data at every stage needs to be kept physically secure, while maintaining accessibility. Data backup protocols need to be enforced on raw and processed research data, original science plan, data management plan, data acquisition strategy, processing procedures, versioning, analyses methods, published products, and associated metadata

Table 2: Description of cross-cutting model elements in the data life cycle, as defined by Faundeen et al., 2014.

Keeping in mind the above data model, we have attempted to structure our toolkit into broad sections that cater to aspects of data within citizen science:

- Before starting a project (planning)
- During the implementation of the project (acquiring)
- After gathering data (processing, analysing, publishing and preserving 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.

2. Data Considerations Before Starting a Citizen Science Project

2.1 Project Planning and Design

As with any scientific endeavour, questions and hypotheses, methods of acquiring data, analyses, and sharing of results are all stages of a citizen science project that need to be planned in advance. As a citizen science practitioner, one needs to be able to foresee the challenges and outcomes well in advance to ensure that project end-goals are achieved. This section of the toolkit provides a summary of points to keep in mind while planning a citizen science project.

2.1.1 How to Make a Citizen Science Project

When starting a citizen science programme, it is essential to plan various stages of the project to ensure efficiency and reach the required goals of the project. Planning a citizen science programme, as with any other project based on biodiversity conservation, has the following main phases:

1.Identify Project Goals

Citizen science programmes vary widely in the primary goals of setting up an initiative. Citizen science programmes can function as citizen contributed repositories of biodiversity data that have a broad focus (e.g. IBP, iNaturalist, Biodiversity Atlas - India: spatial and temporal distribution, taxonomic, trait characters), or have clearly focussed research questions that need to be addressed (e.g. SeasonWatch: phenology). The objective of a citizen science programme is crucial to design various aspects of the program (refer to Section 1.2). Once program goals have been clearly defined, one needs to thoroughly scope out existing research in the chosen theme to identify research gaps that can be filled using citizen science as a medium. This helps identify the type of data that needs to be collected through the citizen science programme, and reduce redundancy. It must be noted that public engagement itself can be a goal in citizen science, and the project may not have other research agendas.

2. Identify Modes of Implementation

Citizen science varies widely in its nature of public participation, which has been outlined in Section 1.3. Defining the roles one wants citizens to play in a project helps in identifying the model a project aims to follow, and the project outcomes that the project proponents aim to achieve.

Plan

3. Identify Target Participants/Stakeholders

Delineating target participants helps design appropriate strategies to recruit, train and engage volunteers for the 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) might have to be identified as well.

When soliciting participation from local communities, localising content in regional languages helps towards greater participation and more effective outreach. The planning stage should evaluate the extent of localisation, efforts required, and whether both data and metadata will be translated.

4. Building Online and Offline Infrastructure

The backbone of a citizen science project is the infrastructure that it needs to function – to maintain registers of participants, to collect, manage and curate data, and maintain regular communication with participants. Online infrastructure includes development of an interface to contribute data. These can be websites or smartphone applications, or even simple forms of communication such as Whatsapp groups. Suitable back-end databases that will store data in appropriate formats and allow interfaces to query and retrieve data quickly and efficiently need to be chosen at this stage. Developing a framework to curate, store and backup data is a key component of citizen science platforms, which needs to be addressed at the planning stage of the project.

Offline infrastructure primarily relates to assistance required in terms of human resources. This involves establishing collaborations with required bodies, acquiring permits for access to protected areas (if needed), and initial outreach to gauge interest amongst the target participants. It is also crucial to design data collection protocols (explained in greater detail in Section 2.1.2) and pilot them within small focus groups, to devise appropriate data collection methods. Finally, framing a data policy is key to ensuring long term participation in a citizen science project (detailed in Section 5). This step can involve expert consultations from experienced citizen science practitioners if required.

5. Volunteer Recruitment, Engagement and Outreach

Engagement with volunteers at every phase of a project is absolutely essential to see a citizen science program evolve and grow. Volunteer engagement can be divided into three general phases:

1. Volunteer recruitment: this involves outreach to the target participants, testing out protocols with focus groups and seeking volunteer feedback on initial processes. This stage is essential to build traction and start a citizen science programme. Social media outreach, publicity articles in print media, tapping email list services and physical presentations at target institutions such as nature clubs, schools or colleges are typical strategies employed. In the case of

projects targeting niche species that are uncommon or restricted in their distribution, partnerships with local communities or tour operators may also be considered.

2. Volunteer education and capacity building: any citizen science project is also an exercise in increasing scientific and ecological literacy among the public, in addition to gathering data. In some cases, volunteers might require specific knowledge (species ID skills, basic survey skills) to participate in a citizen science program. The extent of skill training and knowledge exchange often depends on the data collection methodology. Volunteer education is a long-term exercise, which needs to be done on a regular basis. Practitioners need to ensure that volunteers and contributors understand the scientific problem being addressed, are trained well in collecting information and can use technology (if any) that is required for data contribution, and collect data in a standardised manner. Errors can be minimised by training and reiterating the collection protocol. The contribution process needs to be tested periodically to recognise new sources of error, and build it into training and contribution processes. This is another "quality assurance" step to ensure quality of data is maintained throughout the lifespan of the project.

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

6. Data Management, Analysis and Dissemination

Maintaining, curating and analysing data are the heart of a citizen science program. Data management involves data storage, curation and backup techniques – this ensures that data is not lost once a project is deemed to be complete. One must remember 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 data collected through citizen science. Furthermore, it is important to ensure that data is analysed using scientifically acceptable methodologies and presented via easy to comprehend graphical means to participants, that can be consumed by non-experts. 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.

Recommendations for planning a project

- Citizen science practitioners should determine the end use of project data and accordingly choose the larger research question or broad goals of education, awareness, etc.

Plan

- Practitioners should determine modes of implementation and required infrastructure.
- Volunteer demographics should be well scoped out, including access and ability to use technology, requirements of special skills, and plans put in place to retain volunteer interest over the duration of the project.

2.1.2 Data Collection Methodologies

Citizen science projects vary in the rigour of sampling protocols, from simple occurrence reporting, to more structured data collection techniques. This presents practitioners with a trade-off between volunteer participation and quality of data collected. Often, programmes with rigorous volunteer training and sampling protocols obtain data of better quality, but this minimises the levels of participation within the project. However, programmes with very simple data collection methods report much higher rates of participation, but data is often biased and noisy.

The use of simple techniques such as data collection forms or semi-structured surveys reduces the need for rigorous training, at the same time ensuring that data is collected in a prescribed format. (Bonney et al., 2009, Kelling et al., 2019).

Incentivising quality of observations, rather than number of observations/records in leaderboards and gamification techniques can be incorporated. Deterding et al. (2011) described Gamification as "the use of game design elements in non-game context", and is used to motivate participants to contribute data with the aim of collecting as many records as possible in a specified time period, and enhancing 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 participants.

Prioritisation of spatial and temporal areas where data is required, rather than species and/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

A common challenge in biodiversity citizen science programs is that credibility and quality of data are often questioned, considering that data is not collected by trained professionals. However, repositories of citizen-contributed information, such as geocoded photographs, have been found to accurately and reliably supplement ecological information on species distributions and ranges, e.g. Barve (2015). Citizen science data can be considered of good quality if they are accurate, follow standards, and can be used in reproducible analyses across a variety of stakeholders.

Ensuring data quality is a continuous process, and needs to be carefully kept in mind at each stage of the data life cycle: during data collection, data upload, data storage and

management, infrastructure and finally, data analysis. In order to ensure that data collected through citizen science is credible, data needs to meet the following criteria (https://citizenscienceguide.com/design-sample-collection):

- Accuracy: Data collected through a citizen science project needs to reflect reality. Accuracy of data can be checked by data verification (by experienced members of citizen science community or professionals)

- Precision: Data needs to have a degree of similarity between entries, i.e. data needs to have a degree of consistency and replicability.

- Representativeness: Data collected via citizen science needs to be representative of spatial and temporal scales. Collecting data on date, location, time of observation, weather conditions, etc aids in this effort.

In order to ensure that data from citizen science is usable, stringent methods to ensure data quality are required. It is also important to keep track of the provenance of data. Data provenance is the documentation of the origin of the data and the processes and methodology by which it was produced or by which it evolved over its life cycle. This information is vital towards debugging, tracking changes, auditing, and evaluating quality.

Sources of Bias in Citizen Science Data

In order to effectively mitigate or manage bias in citizen science data, understanding the source of biases for biodiversity and ecology data is key. This helps managers design appropriate strategies at various phases of a project. It is important to ensure good study design, recognise sources of bias and error at every stage of the study depending on the nature of contributions, and extent of contributor involvement in collecting, processing and publishing data; develop the programme iteratively as and when new sources of errors or biases are recognised, and foresee how validated data will be stored, preserved, analysed and made available to an end user, and document this. According to Wiggins et al. (2011), this stage can be categorised under “quality assurance”.

The sources of bias in citizen science data vary according to project design, and can be largely categorised as below.

1.Spatial bias: due to human infrastructure (such as presence of roads, agricultural fields, etc.) and population density (higher in urban areas), a higher number of records are reported from more populated and easily accessible spaces and hotspots (Tiago et al., 2017, Boakes et al., 2016). This results in a bias in citizen science databases that record biodiversity (Geldmann et al., 2016). Easy and better access to technology and internet from urban spaces also contribute towards a better availability of records from such regions.

2. Temporal bias: Citizen science projects report higher rates of records during weekends, holidays and contests. This is particularly of concern in phenology studies, where temporal occurrences form a central part of the research question (Courter et al., 2013). Temporal bias is also noticeable in online citizen science competitions and campaigns that are held annually over limited time periods.

3. Taxonomic bias: Rare species, or species which are difficult to observe or identify, often go under-reported or unidentified in citizen science projects, leading to a paucity of data for such species (Falk et al., 2019). In addition, very commonly observed species tend to be skipped, under reported or even over reported, leading to non-representative sampling (Troudet et al., 2017, Callaghan et al., 2021).

4. Observer bias: Individual perceptions and levels of experience often bias the quality of data collected. E.g. the same event may be interpreted and reported differently by observers having experience or training versus first time observers (Callaghan et al., 2021; Gonsamo & D'Odorico, 2014).

The data collected from citizen science projects can be utilised by a variety of end-users from academic researchers to policymakers. It is important to determine at the planning stage itself, who the likely end-users of the data would be. One or more end-user communities (e.g. scientists, policymakers, amateur naturalists) need to be identified along with the quality of data that is most suited to their purposes. Data quality thus becomes an important consideration keeping end-use in mind. It should also be borne in mind that the onus for maintaining data quality is not only on the programme. Shared responsibility to examine and put adequate thought into their own use of the data to cater for biases, error rates or foibles also rests with the users of data.

Data quality and minimisation of biases can be accounted for before data collection as well as at the data contribution stage. Baker et al. (2021) have summarised the types of data and levels of evidence at the data contribution stage, which would require a suitable verification process:

Levels of evidence -

1. Simple reporting of sightings without other evidence.
2. Photo/video/audio/specimen, where evidence is added.

Types of observations -

1. Direct observation, where taxon is observed directly.
2. Indirect observation, wherein taxon signs (such as tracks, dung, etc.) are recorded.

The mechanisms and criteria that would be used for validation of data need to be thought about at this point and suitably incorporated into the collection procedure to provision 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. Some of these include improper data collection because a given protocol was not followed, incorrect implementation of data collection protocols, mismatch between project goals and data collection protocols, incomprehensive protocols that do not match end-user expectations and inappropriate use of data 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:

Plan

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 standardising devices and recognising limitations of technology

c) appropriate documentation, and

d) metadata to prevent misuse of data in incorrect contexts.

Conferring with experts (such as statisticians, computer scientists, etc.) could enhance the quality of analyses. Inferences should be cautious and take into account all the caveats of data accuracy and analysis. It is also beneficial to get the analyses reviewed by experts and peer groups.

Recommendations for data collection and maintenance of data quality during data collection phase

- To ensure high data quality, appropriate standards should be adopted, and infrastructure to collate, analyse and preserve data should be in place before starting a project.
- Recognise sources of errors and biases in data collection at the outset, and plan data collection and collation processes that minimise these.

2.1.3 Quality Assurance Through Following Standards

It is important to incorporate data collection methods and protocols, fitness of use and assessment of data quality as part of the metadata/documentation itself (Assumpção et al., 2018). Adapting and adhering to the standards inherently helps in improved quality due to breaking up data attributes into appropriate terms and following standard controlled vocabulary to make sure each term conveys the right meaning. In the biodiversity realm, standards developed and continually improved by the Biodiversity Information Standards or originally called the Taxonomic Databases Working Group (TDWG), like Darwin Core and Audubon Core and tools built around them are readily available for Citizen Science projects to use and adapt.

2.1.4 Planning for Data Infrastructure

Data contributed by citizen science participants needs to be collected, stored and processed using technology infrastructure, and this is an important consideration for first-time citizen science practitioners. Contemporary citizen science projects are mostly born-digital, being conceived and implemented predominantly in the digital ecosystem of

information technology platforms, software applications and tool chains. As discussed earlier in the introduction section, there are multiple concerns in terms of online and offline infrastructure – proponents may wish to choose between larger aggregator platforms that allow projects within them vis-à-vis building independent applications. Fast growing mobile application technologies have made it possible to deploy tools for data collection and integration, quickly and with little effort (Lemmens et al., 2021). On the other hand, several large biodiversity data aggregating platforms are well established and have gained reputation across the globe (e.g. ebird, iNaturalist), or for country-level data (India Biodiversity Portal (IBP) and Biodiversity Atlas for India, Atlas of Living Australia for Australia, national GBIF nodes, etc.). Others look at specific taxa or geography (biodiversity hotspots such as Western Ghats and Eastern Himalayas) or simply to address a specific question. There are obvious advantages of using an existing platform like iNaturalist or CitSci.org 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 larger goals of the project, 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 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 allow such flexibility. E.g. the IBP allows creating groups within its infrastructure for any theme of interest such as a taxonomic group, e.g. Shieldtails (Uropeltidae). Forms for gathering data can be extended to include custom queries and fields.

Data infrastructures for citizen science projects need to be adaptive to address the unique nature of each citizen science project. In general, the data infrastructure should 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). E.g. citizen science projects could use a phone or web application to collect data, a cloud server to store the data, an automated code to verify data, a web portal to promote interaction among contributors, and a backend database structure such that it can be aggregated with other types of data. This would also mean that the infrastructure enables participation of citizen scientists in the full range of scientific methods from problem definition, research design, analysis and action (McQuillan, 2014).

As citizen science projects in biodiversity tend to collect data across 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 such as text, tabular, geo-spatial and varying media types including images, audio and video. Such capabilities will have impacts on the scalability of storage required, particularly if the project is of a long duration. Cloud-based storage and content delivery networks in the mainstream IT ecosystem have matured enough to ensure such scalability and high availability across geographies.

Apart from these fundamental concerns on data models and data storage, one also needs to ensure that the platform is stable and ensures continuous access to participants with minimum downtime. Platforms need to cater to the overall security of data with well-defined data access policies, user authentication systems with defined roles, transparent workflows and user-centred design (Bowser et al., 2020). Regular backup of data with

multiple copies in multiple locations and a preservation policy with consistency managed across sites are important for both 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.

Recommendations for planning data infrastructures

- Depending on the nature of the project, choose between large ready-to-go platforms or custom-built data infrastructures.
- Data platforms should be able to accommodate diverse types of data (text, table, geo-spatial and media such as images, audio and video) and remain scalable, highly available and secure.
- Emphasis on free and open-source technologies will help in replicability and sustainability of infrastructure.

2.2 Data Ownership

Who owns the data that comes into a citizen science project is a crucial consideration one must make at the planning stage of the project. The manner in which participants perceive ownership of data that they help generate may guide their motivation in participation in citizen projects. Yet, studies have indicated that there is much ambivalence in how participants feel about data ownership. On one hand, it can be said that most participants are far-removed from thoughts of data and its ownership with each record being more of a personal nature experience that is recorded and less so as data with legal ownership. Ganzevoort et al. (2017) best summarise this as constituting "an "imagined contract" between volunteer naturalists and nature, based on respect and wonderment…".

On the other hand, in general, participants feel that "...data extracted from nature should properly be used towards its preservation", and hence "wrong" use of data can result in citizens being upset and withholding contribution (Ganzevoort et al., 2017). 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 felt that data generated from citizen science projects should not be unconditionally usable, most participants are undecided on ownership with some feeling data is nobody's property and some others that it could be owned by the organisation 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 such as through policies that put in practice exclusive or non-exclusive access to or control over data (Guerrini et al., 2019). A lot of times, traditional knowledge belonging to communities related to bio-resources or conservation practises 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, but 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 certain practises, beliefs and linguistic representations of the same, which may get lost in translation. One has to also be mindful of the cultural sensitivities and secretiveness shown by certain communities to divulge their knowledge. The communities must have the freedom to say no to sharing of their knowledge if they wish and if they do agree, then they should be allowed to choose the manner in which their knowledge is being used within a citizen science project.

Recommendations on data ownership during planning

- Organisations that manage data should avoid ambiguity and misunderstanding and make their data sharing policies clearly known to participants through best practises.
- The manner in which contributed data will be used or transformed and what parts of processed and raw data will be accessible for edit and download from the project should be made explicit to participants.

2.3 Data Accessibility

Another consideration of importance at the planning stage is to have clarity on who can access the data contributed to citizen science projects, at what stages and for what purposes. Accessibility of data generated through citizen science projects is a core aspect to consider for proponents when designing the project. Open access to data is important towards democratising science as well as upholding the values of universal and equitable access to scientific data, especially when it is gathered through public participation. Just as we strive to make citizen science accessible to a diversity of participants and make 'doing science' as inclusive and participative as possible, it also needs to keep the resulting data accessible in a manner that would support reproducible science, and the public at large by influencing policy by 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 allay these concerns by institutions conducting citizen science projects.

What is Open Data?

There are variable interpretations of the term 'open data' and it is best to clarify at this stage, what open data is. 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 – 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,

1.Legally open: where it is made available under an open (data) licence that allows anyone to freely access, reuse and redistribute the data.

2. 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.

Open data, therefore, means that it is complete, preferably downloadable over the internet in a format that is convenient, and modifiable 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, it is likely that not all projects may be in a position to adhere completely to the tenets of open data. Some do not always agree 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). Some also cite the burden and expense of running massive data projects, curating data and processing as well as 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 activities in a project, or more importantly to obtain acknowledgement particularly as authors on publications. Some might be readily willing to share data but on request to keep a track on how the data is being used, hence not publishing them under an open access licence. Other important reasons for data not being open are due to projects being 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 both, regarding those of the participants generating the data as well as 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.

The rationale behind recommending that citizen science data be made openly accessible are many. 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". It further 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 better 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.

Plan

2.4 Ethical Considerations to be Made at This Stage

The toolkit being prepared by the 'Diversity and Inclusion Working Group' aims to address ethical considerations in citizen science in detail, while for the purpose of this document, we briefly highlight the most pertinent ethical considerations to be kept in mind at each stage of data processing.

Here we focus on ethical considerations with respect to the planning stage of a citizen science initiative. These cover the realms of recognising contributor rights in citizen science and designing projects that are socially inclusive. It also includes information on making data public or open-access versus limiting access to it. The latter component has already been addressed in detail above (Section 2.3).

With 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 that may exist between project handlers and contributors. Here we recognise some of the more apparent forms of ethical considerations that project managers must oblige to.

It is imperative for project managers to recognise the participatory nature of citizen science, where volunteers contribute data and time obligingly. It is also important to ensure that projects are socially inclusive. This approach of data collection is recognised for its abilities to democratise science and hence an important facet of it involves being inclusive of participants irrespective of gender, geographical location, socio-cultural, religious, linguistic and academic backgrounds (Paleco et al., 2021). Simultaneously, a project must aspire to be open to participation at all stages of a project.

Project designers should consider identifying means of reaching out to potential stakeholders to ensure maximum participation. These could include interested citizen participants, such as interested members of the public, established citizen scientists or those from the scientific fraternity, academic institutions/ organisations, 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).

Ethical considerations at the planning stage of a Citizen Science Project: a summary

- Recognise contributor rights
- Design socially inclusive projects
- Encourage open-access data
- Design projects open to participation at all stages
- Maximise participation by creating means to reach out to all potential participants

Acquire

3. Data Considerations During Implementation of a Project

Once the study design, data quality, adherence to standards, data infrastructure and accessibility are planned for in a citizen science project, the next step is to implement the same. In this section, we summarise data considerations to be made at the stage of implementing a citizen science project.

3.1 Data Infrastructure

In the course of the implementation of a citizen science initiative, it is imperative to provide appropriate, scalable, highly available, secure data infrastructures for acquiring and organising incoming data streams. For an effective organisation of data with transparent workflows and well-defined roles, adherence to standards become essential. Discipline-specific standards have evolved over the last decades and are still being shaped in tandem with the spurt of diverse data types in life sciences. Some of these are discussed in the next section. Adherence to such standards not only make the ensuing process of analysis easier and effective but also enables sharing and aggregation of data across different repositories.

Apart from adhering to the relevant standards and guiding principles such as open data, data infrastructure also needs to ensure compliance with FAIR data principles. Findable, Accessible, Interoperable and Reusable (FAIR) organisation of data facilitates, in the long term, the discovery of knowledge, allowing integration and reuse by the larger scientific community. As the role of digital data increases everyday, it becomes important to ensure data organisation that is easily accessible to humans and also to their computational agents. FAIR data requires that computational agents such as computers, smartphone applications, servers, algorithms and toolchains, can autonomously discover data with established protocols. The proponents of FAIR data principles terms this as 'machine-actionable' (Wilkinson et al., 2016). Other notions such as CARE data have emerged in the recent past in the context of data originating from indigenous communities where CARE stands for Collective benefit, Authority to control, Responsibility and Ethics around data (Carroll et al., 2021). Such evolving debates over the nature of data generation, processes and purposes of intended use in the larger socio-political context will also have ramifications for citizen science data and related infrastructures, which are discussed in the 'Ethics' sections.

Infrastructure and content for local language support, if required, will need to be sourced, developed and incorporated at this stage. Given the linguistic diversity in India, providing user interfaces in relevant vernacular languages might help in large scale participation. This localisation effort includes structuring databases capable of supporting multilingual data or coming up with other on-the-fly frontend mediated mechanisms to enable translations where only metadata support may be required.

Acquire

Recommendations for data infrastructure during implementation

- Adherence to standards is crucial, along with transparent workflows and well-defined roles.
- Ensure compliance with FAIR data principles.
- Internationalisation and localisation capability should be adopted where possible to facilitate vernacular language support.

Acquire

3.2 Data and Metadata Standards

Biodiversity data standards are shared rules and conventions to describe, record and structure biodiversity data to enable data aggregation and exchange across different organisations generating and managing different data sets. Standards like Darwin Core are already being widely used across various data aggregators. Data standards enforce unambiguous definitions of what kind of data is being collected, follow well defined ontologies and vocabularies and standardise the usage of established protocols. It is recommended that each project follow existing international standards, and adopt recommended storage formats and protocols. Adherence to the standards makes it easier to seamlessly share data with project partners, projects or global data aggregators.
The use of Biodiversity Data Standards addresses two key objectives:

1. It provides a comprehensive set of attributes that are relevant for most projects and meet individual project needs for collection and management of data.
2. Projects may identify a subset of the core biodiversity data attributes that can be used to aggregate data.

Given below (Table 2.1) are some of the important international biodiversity data standards.

Standard	Description	URL/source
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/

Table 2.1. A non-exhaustive list of commonly-used biodiversity data standards

Data submitted by users is typically restructured slightly to adhere to the standards followed by the project to store in the database (Turnhout & Boonman-Berson, 2011). To avoid friction, either this change needs to be communicated to users or the data input itself is accepted in a structured manner. During such standardisation and large-scale aggregations, it is important to note that all the contextual richness may not be preserved (Ganzevoort et al., 2017), i.e. the original submission having vivid description of courtship process may just get restructured to presence of male and female organism and a tag for courtship ticked.

Acquire

Recommendations on Data and Metadata Standards

- It is important to select the most appropriate standard for the kind of data being compiled and shared.
- Documenting metadata in standardised formats is equally important

3.3 Quality Assurance and Quality Control

Citizen science projects with good data quality rely on multiple methods to ensure data accuracy while accounting for biases and iteratively developing the project at the stage of data-acquisition. Not only do such projects adhere to good practises related to standards, metadata and documentation, they also ensure that errors and biases are addressed at the level of volunteer training and testing, and any erroneous observations are flagged via validation methods (Kosmala et al., 2016). The method engaged to train volunteers can influence data quality, with studies showing direct training, or remote but repeated training, being the most effective in ensuring robust data (e.g. Ratneiks et al., 2016). If a citizen science project relies heavily on technology for contributions, care must be taken to ensure that the usage of the tool does not affect the accuracy of the information (Downs et al., 2021 and references therein). Assessing citizen science data quality can be extremely difficult due to heterogeneous observers and methods, and lack of information about such methods. In particular, data bias, errors, uncertainty, and ethical issues pose challenges that should be assessed regularly as part of citizen science research projects (Downs et al., 2021). Most considerations on data quality are from the point of view of integrating data from different sources. Data standards thus become an important component of data quality. Incompatible design of citizen science studies and inconsistencies in nomenclature can affect data quality, resulting in challenges for integrating data from different citizen science programs (Campbell et al., 2020).

The following fail-safes can be used to 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, using standardised methods of data collection and following established standards of terminology, participant training, auto correcting entries (e.g. erroneous geocoding), data verification, facilitating access to data use (Balázs et al., 2021). In projects using devices, sensors should be calibrated and initial checks on devices and ability of observers to use these devices should be made (de Sherbinin et al., 2021).

Depending on the types of observations, post collection data verification is a very important step to ensure data accuracy. Baker et al. (2021) compiled information on 259 citizen science studies, and found that nearly 45% did not have a verification process, and therefore the accuracy of these data could not be assessed. According to Wiggins et al. (2013), this step would fall under "quality control" of data. One or a combination of the below processes should be implemented for citizen science data verification:

Acquire

1. Community consensus/Peer verification - two or more members of the community agree on the accuracy of the data.
2. Automated verification - data passes through an automated filter to quickly tag potentially erroneous entries, which can then be validated through experts or community consensus.
3. Expert verification - get the data assessed by one or more experts for accuracy. This step is typically implemented when data are flagged as potentially erroneous through the community or automated methods.
4. Model-based verification - using statistical models to address random variation and residual errors in phenomena of interest to flag potentially erroneous observations (Balázs et al., 2021), building in uncertainty of devices and individual measurements into the data quality check process (e.g. Kelling et al., 2015)
5. Linked data analysis: combines freely available data and helps establish data quality through techniques such as data mining (Balázs et al., 2021)

Verification by experts and community may involve observers being asked for additional documentation (such as photographs) to help reduce ambiguity and confirming the accuracy of the observations.

Baker et al. (2021) also recognise three main contexts that are key to the data verification post-collection:

1. Species (ID, geographic co-occurrence with other species, rarity) - A majority of ecology citizen science projects in India and elsewhere require accurate taxonomic information. Taxonomic verification thus becomes a key aspect of quality assurance. It is good practice to identify the taxonomic classification system that would be followed while verifying contributed data as scientific names may differ based on this, e.g. the Pongamia tree may be referred as Pongamia pinnata or Millettia pinnata or Derris indica depending on the classification adopted.
2. Environmental (time, date, location): Information collected in this context can be used to identify data that is potentially incorrect by comparison against known phenology, range or species activity.
3. Expertise (experience of recorder): The reliability of individual contributors in making accurate observations.

One recommendation emergent across multiple studies is to implement iterative evaluation and development – which involves seeking feedback and assessing performance of participants iteratively and implementing these learnings in making the project more robust and thus ensuring data quality (e.g. Kosmala et al., 2016).

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 meets standards of reproducibility (for analyses) and lend themselves well to being combined with other standardised datasets. Citizen science projects may benefit from explicitly stating the mechanisms they use to ensure data quality, and follow data standards. Information about data quality helps potential data users to determine whether and how data can be used and enables the analysis and interpretation of such data. Providing data quality information improves opportunities for data reuse by increasing the trustworthiness of the data (Downs et al., 2021). 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 (such as in the linked data analysis described above). de Sherbinin et al., 2021, suggest the storage of data in its most disaggregated form in citizen science

projects maximising privacy along with explicit documentation of biases and other assessments of quality.

Considering that there is a likelihood of bias in citizen science data, these biases can be addressed during the data analysis stage as well to ensure data quality. Some examples are listed below.

- Incorporating an "observer expertise score" as a covariate while modelling citizen science data (Johnston et al., 2018).
- Many citizen science projects today incorporate AI in data analysis. Considering the existing bias already in citizen science data, care must be taken to reduce bias in training data. An example of such an effort is the shift compensation network, which learns shifts between scientifically objective data, and biased data; and incorporates this into the training model (Chen & Gomes, 2019).
- Using data filters based on sampling effort or observer expertise are also used to reduce noise in citizen science datasets (Steen et al., 2019).

Recommendations for maintaining data quality during project implementation

- Ensure that data accuracy is maintained during data capture.
- Assure data quality through review – manual, automated or a combination of both.

3.4 Licensing

How can one ensure that media such as images or videos contributed to a citizen science project are used appropriately? Data ownership and accessibility at the citizen science implementation stage can be facilitated by providing opportunities for licensing the data contributed by participants. Copyright is a state-guaranteed right covering 'work' including intellectual creations, such as text, photographs, diagrams, maps, movies, etc. that are eligible by being 'original, individual, singular and new'. Ideas, knowledge, information, or data are traditionally not copyright-protected and scientists have been content with being cited for their original work (Hagedorn et al., 2011) with the intention of public access and dissemination of knowledge. Although it is commonly assumed that data with no licence applied is free for open use, this is not the case. The lack of a licence poses ambiguity in its reuse which is risky especially where the terms of usage of the data have to be made explicit, especially for commercial usage (Groom et al., 2017) and may lead to unwitting copyright violations. In this context it is essential that data be made available under carefully crafted licences where the terms and conditions for its reuse are made clear.

What Licences to use for Citizen Science Data

In addition to making data open, additional mechanisms are required to make data findable, accessible, interoperable, and reusable (FAIR; Wilkinson et al., 2016). The adoption of open, machine-readable licences is recommended to achieve this objective (de Sherbinin et al., 2021).

Acquire

The most common licence employed in citizen science data is the Creative Commons Licence (CC, https://creativecommons.org/). This licence 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). As stated on their website, "Creative Commons licences are not an alternative to copyright. They work alongside copyright and enable you to modify your copyright terms to best suit your needs." A violation of a CC licence is a copyright violation. The CC licences provide standardised terms-of-use definitions that have been adapted for various jurisdictions and upheld in court in several countries (Hagedorn, 2011). The licence has been adapted for India under the aegis of Wikimedia India, Centre for Internet and Society, and Acharya Narendra Dev College (https://wiki.creativecommons.org/wiki/india).

CC licences 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 licence can cover all use cases and instead provides a set of licences to cover a wide range of use cases. These range from combinations of four conditions:

1. The "Attribution" condition (abbreviated "BY"), which is a part of all CC licences and requires users to give appropriate attribution to the creators of a work.
2. The "Share Alike" condition (abbreviated "SA") allows the distribution of derivative works, but requires that all such works must also be shared under the same conditions.
3. The "No Derivative Works" condition (abbreviated "ND"), which states that the user "may not alter, transform, or build upon this work".
4. The "Non-Commercial" condition (abbreviated "NC"), which states that one "may not use this work for commercial purposes".

The CC Zero or CC0 Public Domain Dedication Licence, in which "No Rights Reserved" or "all rights granted", is the most liberal licence. CC BY, CC BY-SA, CC BY-ND, CC BY-NC, CC BY-NC-SA, CC BY-NC-ND are the other available licence combinations with CC BY-NC-ND being most restricted. SA, ND and NC limit the ability to derive and reuse freely and often in unexpected ways (Hagedorn et al., 2011), hindering value-added data and services based on raw data. The former is not recommended for datasets and the latter two are also not recommended for scholarly or scientific use. The CC licences are accordingly adopted in whole or part by large data repositories such as the Global Biodiversity Information Facility (GBIF, https://www.gbif.org/), Wikipedia, Wikimedia Commons among others. As of now, CC0, CC-BY and CC-BY-NC are the only CC licence options recommended by GBIF. As scientific data is mostly facts and is not copyrightable, CC0 is the recommended licence for data (https://wiki.creativecommons.org/wiki/CC0_use_for_data). In the event that any images are contributed as part of the data, the terms of use of the platform gathering data should be clear on the applicability of the CC0 licence to such images as well.

Another such licence that is relevant is the Open Data Commons (ODC) maintained by the Open Knowledge Foundation (https://opendatacommons.org/). Although Open Data Commons licences are more suitable for data licensing, they are more specific to databases, and apply only to database frameworks and structures, not to the 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 licences:

- Open Data Commons Open Database Licence (ODbL), providing "Attribution Share-Alike for data/databases"
- Open Data Commons Attribution Licence (ODC-By) providing "Attribution for data/databases"

- Open Data Commons Public Domain Dedication and Licence (PDDL) providing "Public Domain for data/databases"

When these licences are used, one would still require a standard licence to be used in combination with it to protect copyrighted content within.

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 licences to be used for datasets published under NDSAP through the OGD platform remained unspecified until the release of the Government Open Data License - India, that is governed by Indian law (Government Open Data License - India, 2017). 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 ". The terms of licence however, remains ambiguous and it has been criticised 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 licences for a citizen science project. However, this is not a trivial endeavour and will almost certainly have to include the participation of legal offices and organisational research departments (Ball, 2011). Such cases are usually unnecessary considering the availability of standard licences as documented above, except when there are exceptional circumstances requiring the same. This is provided that adequate and standard safeguards are already in place. Creating additional bespoke licences adds to the burden on end users of the data in ensuring compliance and adhering to multiple licence requirements.

Once a suitable licence has been decided upon, one needs to attach that licence to the data. This mostly involves a statement that the data is released under the chosen licence or public domain and a mechanism for retrieving the full text of the licence itself. It is important that the rights statement 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 also highly recommended (Ball, 2011).

Recommendations on setting project licence

- Ensure that all citizen science data be made openly and freely accessible through open licences.
- Data should not be left unlicenced.
- Custom generated licences should be avoided unless absolutely necessary
- Open standardised, machine-readable licences should be adopted to the extent possible.
- Data and creative media may need to be separately licensed.

3.5 Ethical Considerations to be Made 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-licences, encouraging participants to contribute data collected following fair practises, and legal and social conformity to data being incorporated from indigenous communities.
In recent times, much emphasis is being laid on prior communication regarding project components such as objectives of a project, terms of data usage, methods for data storage, recognition of role of participants, amongst others. This is particularly true for contributory projects which are common in biodiversity related studies.

Detailed informed consent forms are strongly recommended for many reasons, some of which as listed out by Sullivan et al. (2014) include the following:

1. ensuring that data contributed by citizens is not misused,
2. making participants aware of the project's data licence, and mandates the project to respect its objectives set-out at the onset of the project and communicated to its participants,
3. helping project designers analyse data-usage trends by registering participants, which in-turn makes it possible to communicate with them and keep them informed of developments in the project, and
4. preventing unauthorised use of personal information contributed by participants. It is equally important to ensure that consent forms are available in languages understood by the participants, especially in multi-linguistic nations such as India.

Clear communication related to data-licences (Section 3.4) is also important, with respect to both, overall project and individual data contributed by participants such as photographs, audio-recordings, personal memoirs or other similar data.
Project managers need to be aware of legal components that govern usage of information pertaining indigenous communities or other regional laws that may be applicable. Traditional knowledge pertaining to indigenous communities must be handled sensitively and collected only after taking prior consent from such groups. National laws related to copyright or protection of imagery and text narratives, must be well understood before accessing such data and complied with.

Effort must be made to ensure project participants abide by government and community laws and regulations while accessing data which they aim to contribute to the project. In the case of biodiversity projects, safety of biodiversity and participants should come as a priority over data collection.

If gamification or incentivisation is used to encourage data contribution, pros and cons of the strategy must be carefully looked into. Gaming elements can positively influence participant engagement by creating an environment of fun, competition or both (Iacovides et al., 2013; Bowser et al., 2013B). They may also reward certain participants for achieving the highest scores or winning competitions. This in-turn results in enhanced user participation. However, it may also compromise the quality of data or demotivate participants for not reaching the highest target. Citizen science initiatives are broadly developed with scientific objectives in mind. However, gamification may distract the participants from the actual pursuit of science and digress their objective to the solitary aim of winning a game. This pursuit for victory may encourage them to indulge in unfair practises in order to increase their chances of winning or focus exclusively on the fun

elements of the game. Similarly, they may guard their methods of participation closely or information pertaining locations of species, with the aim of not sharing their secret of winning, in-turn contradicting the fundamental principle of citizen science, i.e. openness. Methods or skills required or employed in gamification may also interfere with the principle of equality – an essential component of citizen science – by putting certain participants at an advantage of winning over others, due to access to resources, or other means.

As per Ponti et al. (2018), below is a list of implications of gamification that citizen science projects may want to look into when designing projects:

A. Important to give thought to the game design element of a project, (e.g. how to score a game) because this could influence strategies used by participants, and hence their performance.
B. Contributors may develop different values of science and citizen science in particular, and hence project designers need to be sensitive towards such value changes that their project has the potential to trigger.
C. Games may instantly recognise the highest score; however, it is equally important to give recognition to volunteer contributors in non-gaming context in a fair and objective way. E.g. how does one recognise the role of participants who contribute data outside of the competition period?
D. Competition may be rewarding for some and demotivating for others.

Another fast-growing area is the application of artificial intelligence techniques such as deep learning and convolutional neural networks to classify images amassed through citizen science projects, especially for species identification. Such AI applications are also found to be useful to extract and classify species images from social media, thus helping in biodiversity monitoring (August et al., 2020). With the inherent ability of deep learning algorithms to self-learn from vast datasets, particularly multimedia data such as image, video and audio, in the context of environmental conservation, they offer a promising approach to automatically classify visual, spatial and acoustic information (Lamba et al., 2019). As most citizen science projects generate multimedia information with geo-location, the complementarity of citizen science and artificial intelligence for ecological monitoring is acknowledged, particularly in rapid data analysis. But ethical challenges exist, where black boxed artificial intelligence systems with their closed algorithms are trained with citizen science contributed open data thus excluding citizens from understanding how their data contributions are used. Transparency in such AI systems then becomes essential, which will also help detect biases in training datasets thus improving the efficacy of these systems where eBird's human/computer learning network (Kelling et al., 2012) is cited as an example of such a transparent system (McClure et al., 2020).

Ethical considerations at the data-acquisition stage: a summary

- Clear prior communication with potential participants explicitly mentioning project objectives, terms of data usage, methods of data storage, means to recognise contributions by participants and data-license details of the project.
- Obtain consent forms from participants before accepting data from them. A multi-linguistic consent form is highly desired.
- Indigenous communities and areas governed by them may be protected under special legal regulations. Project proponents need to acquaint themselves with these laws and work within the prescribed legal framework.
- Ensure participants are respectful of and abide by national laws governing biodiversity and those on copyright or protection of imagery and text narratives.
- The safety of biodiversity and participants is paramount while collecting data, and respect for government and community regulations is necessary.
- Explore pros and cons of gamification, and if it must be used, ensure fair practice, and equal opportunity.
- Despite the advantages of artificial intelligence systems, keep in mind issues pertaining to transparency that may emerge as a result of such systems.

4. What to do With Data That Comes into a Citizen Science Project

In this section, we outline considerations related to data storage, processing, analysis and dissemination, once standardised, verified data have come into a citizen science project.

4.1 Data Infrastructure

Data acquired into a citizen science project needs further processing. Given the dynamism of participation, study design and funding in citizen science initiatives, prevention of loss of data and ensuring security of data is of paramount concern. It begins with managing the physical risks for data storage and ensuring access to collected data through its entire lifecycle. 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 (http://www.bouchoutdeclaration.org/declaration/), with its stated mission "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", specifically lists among its ten fundamental principles;

- Using identifiers in links and citations to ensure that sources and suppliers of data are assigned credit for their contributions;
- An agreed infrastructure, standards and protocols to improve access to and use of open data;
- Registers for content and services to allow discovery, access 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;
- Linking data using agreed vocabularies, both within and beyond biodiversity, that enable participation in the Linked Open Data Cloud

Table 4.1. Important fundamental principles for free and open access data, as listed under the Bouchout Declaration for Open Biodiversity Knowledge Management

Approaches to data lifecycle management and principles espoused in the Bouchout Declaration point to implementing various strategies pertaining to aggregation and processing of data thereby facilitating analysis and transformative action. Various

techniques in deploying persistent identifiers such as Persistent URLS, Digital Object Identifiers, LifeScienceID and Personally Identifiable Information are being adopted by many platforms to ensure that data in its various types and stages are traceable with identifiers. Implementation of such persistent identifiers will become a bottomline in the near future and will help in ensuring data quality, access and accreditation.

For storing citizen science data that has been curated for research quality, trustworthy data repositories such as Zenodo/Dryad and Mendeley Data could be considered. These repositories were developed as part of the efforts of the Research Data Alliance, which instituted a set of harmonised common requirements for certification of research data repositories, ensuring that they 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 perpetuity of data. While many such data repositories are evolving with Long Term Ecological Observatories 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 perpetuity of citizen science data and maximise its re-use. Such public data archiving for citizen science initiatives, although required, is also a challenge to build (Pearce-Higgins et al., 2018).

Recommendations for data infrastructures in processing and sharing phases

- Ensure free and open access data for both people and machines by implementing strategies like persistent identifiers.
- Ensure data perpetuity by making data available on relevant public trustworthy data repositories for effective reuse.

4.2 Data Standards

Data standards play an important role in biodiversity data publishing. Following data standards makes data publishing, either through aggregators like GBIF or in the form of data papers, simple. It saves effort involved in describing metadata and makes the published data readily usable for the intended user base. Data papers and data repositories often require the metadata to be marked up in standardised formats such as in EML. Independent projects may use software such as R or Morpho to markup the metadata from their datasets (https://old.dataone.org/software-tools/morpho). Many larger platforms such as iNaturalist or IBP serve as archive as well as a publishing platform and already have some standardisation inbuilt within their structure allowing data downloads to be served under such standards. Such platforms also have arrangements regarding publishing the data to global biodiversity repositories such as GBIF through common standards.

Recommendations on data standards while publishing

- Ensure published data is structured, and conforms to standards.
- Utilise third party services and software to map data into standardised formats if it is not compliant already.
- Adequate metadata should be generated during publishing.

4.3 Data Accesibility

As stated earlier Open Access data means that "data must be freely available for download online". This also implies that the data is accessible in formats that do not need proprietary software to open and as discussed above, must have an open licence for reuse. The CSV format is generally used for tabular data download and this also ensures compatibility for machine reading of the data in a machine-readable format.

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

From an accessibility perspective, there is a need to involve citizens beyond the act of data collection. In general, participants are rarely given opportunities beyond data collection such as those involving data analysis or interpretation (Kennett et al., 2015; Lukyanenko et al., 2016). However, it is important to provide them with opportunities and incentives to interact with data that they have played a part in generating. When users are able to also reuse the data that they generate, it has the potential to influence learning and conservation outcomes, and may also lead to better user retention in a project (Cooper et al., 2017). Such interaction can be achieved through participatory data analysis and visualisation that can be user generated as per needs and variables of interest. Many projects are increasingly gravitating towards developing such interactive visualisations for participant engagement to involve citizen participants in data-analysis. However, since data analysis is usually from an end-user's specific perspective, generic visualisations and analyses inbuilt into portals may be limited in what they offer as they are usually set up to predefined criteria. Such limitations can be overcome through developing and offering application programming interfaces (APIs) and client packages for popular data analyses software such as R or Python. Some examples of such packages are the 'rgbif' (https://cran.r-project.org/web/packages/rgbif/index.html) and 'pygbif' (https://github.com/gbif/pygbif) clients for interfacing with GBIF and the 'galah' R package (https://atlasoflivingaustralia.github.io/galah/index.html) for acquiring data from the Atlas of Living Australia. This kind of capability would allow users to fetch data flexibly, do further analysis and to generate custom visualisations as per their needs.

Recommendations on data accessibility

- Ensure data is available in accessible and open formats that do not require proprietary software for processing
- Allow opportunity for participatory analysis of generated data.
- Offer accessibility to data for developing visualisation and analytics in a dynamic manner through APIs and software packages.

4.4 Dissemination of Knowledge Gathered Through Citizen Science

Citizens cannot only be viewed as contributors and participants in a scientific endeavour; they are also the end users. 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.

One way of encouraging participation in citizen science projects is by incorporating clear channels of communication and data dissemination (Vohland et al., 2021). This allows access to a larger audience, makes people aware of the project, and keeps them in touch with the project, which is essential to retain participants. Dissemination is a one-way communication mainly at the end of a project. The traditional means of disseminating scientific knowledge is through publications in peer-reviewed journals. However, information presented in academic journals can often be technical for non-scientists 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:

- Creating data visualisations on the project website that communicates results in an attractive manner - spatial data can be displayed through heat maps, e.g., eBird status and trends abundance animations that reveal migratory pathways of birds.
- Writing articles in popular media sources like newspapers, online magazines, etc.
- Visual communication of knowledge through art, videos and graphic design.
- Social media can be leveraged to disseminate results

It is worthwhile to note that disseminating knowledge to the public is crucial to ensure long-term participation and collaboration in any citizen science program. However, as a best practice, it is important to ensure that communication with participants is done in moderation. Frequent emails or other means of contacting participants can have adverse effects and can in fact discourage participation. It is also important to acknowledge the advantages of varied forms of data dissemination means. E.g. While popular means of communication allow reaching non-scientists, reports and academic publications are of significant value to the scientific community and policy makers.

4.5 Data Attribution

Attribution is the act of giving credit to the data providers in publications. Author attribution has historically been a tricky issue across disciplines which has further accentuated with

the advent of big data, and in the absence of consensus on proper guidelines for authorship even today (Escribano et al., 2018; Venkatraman, 2010). While protocols such as the Science Commons advocate publishing data openly, there is no mention of how to provide attribution. Authors typically negotiate their order within the author list with the assumption being that the first author being the most coveted and having led the publication idea and the last typically being the head of the lab and also the point of contact (Venkatraman, 2010). As the contributor list grows, especially in large collaborative projects, the contribution order becomes obscure and meaningless. Some journals provide the provision for a separate text or list stating individual roles and contributions instead of authorship.

In citizen science too, there is much ambiguity regarding who should get attributed and how, and whether individual citizens should 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 in the case of 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, where ensuring the correct citation formats are maintained either manually or by machines itself becomes a challenge. To tackle this, it becomes necessary to allow for 'lightweight attribution mechanisms' (Ball, 2011).

Recommendations on data attribution

- Publish and make clear attribution requirements on websites or apps from where data is consumed, ideally as a short simple statement.
- Allow for flexibility in attribution by the end-users, who often consume data from multiple platforms and sources.
- Provide clear persistent URLs or Digital Object Identifiers or other means to allow for automated tracking of citation/usage of provided data.
- Wherever possible, provide a list of citizen contributors who have contributed each data record, to enable attribution in derived works.
- Clear terms should be set out within the data policy informing citizens how their data would be cited and attributed.

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

5. Data Policy

One way to ensure that the data collected through citizen science projects are stored, shared, attributed and utilised ethically, is to have a clear and robust data policy for all stages of the data cycle. Citizen science project proponents should be mindful of different stakeholders – for instance there are volunteers who provide the data or those who use the data or any other related activity pertaining to the project. Data policies created for such projects must be mindful of the differing rights and responsibilities that each party may possess. While the definitions of citizen science are still evolving, it generally encompasses participation from individuals without specific scientific training who participate as volunteers in activities, allowing either personal specimens or observational data to be collected from them. 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. It is vital that informed consent be obtained from such volunteers on how the data collected from them will be used with specifics of how they will be credited. Informed consent and refusal is one of the essential components of research ethics that the volunteer willingly gives themselves up for use as a resource (Reiheld & Gay, 2019).

Informed consent can be ensured by the use of 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 project portal. These documents should constitute the policies of the project that are a collection of guidelines that determine how a citizen science project and the users or a website or a citizen science volunteers may interact or transact. Such documents are usually presented as different types of formalised policy documents (Bowser et al., 2013A). These include:

a) Terms of use - These form the conditions that a user is expected to be aware of and accept before they begin using the portal. It also encompasses guidelines for acceptable behaviour between the user and the portal. Terms and conditions may be explicit, requiring the user to accept and consent to the terms of the site before proceeding with registration and usage (clickwrap) or it may be implicit, assuming that the user consents to the terms simply by continued use of the portal (browsewrap). The terms of use sets the conditions of usage of the portal, covering aspects along the lifecycle of the data generation, curation and output stages. 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 clearly specify how the data will be used by owners of the site. 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 protects 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

the portal, similar to what is observed in the European Citizen Science portal (https://eu-citizen.science/terms/). Additional terms of use may allow the portal to block a user in case they violate the terms of use.

b) 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 texts containing specific legal or nonlegal documentation.

c) Privacy policies - This covers information on how and what kind of information the project gathers from participants, including information gathered 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.

Rights and Responsibilities of Project Proponents

Citizen science project proponents also need to be clear on what their rights and responsibilities are, before hosting the portal. In the Indian context, as per the Information Technology Act, 2000 (ITA) (Section 79), if portals satisfy the following criteria, they will be not be liable for information posted by its users:

- the transmission of the information was not initiated by the portal or its administrator
- the recipient of the information was not chosen by the portal or its administrator
- the information was not modified or altered by the portal or its administrator

An illustration of an intermediary: A user posts information taken by them in relation to a citizen science project on a citizen science portal. The user has the option of selecting whether their pictures can be viewed by specific users, in addition to the administrator or by the public. Once the information is posted by the user, the administrator of the portal cannot make changes to the same. In this case, the users have complete control of the information on the portal. Since the administrator does not initiate the transmission by posting the information, select who will be able to view the information or make changes to the information, the portal can claim exemption as an intermediary from liability arising out of the conduct of its users. Social media platforms and other platforms which allow individuals to post content are examples of intermediaries since they act as channels through which information is communicated.

If it is clear that the portal is an intermediary and all the conditions under Section 79 are fulfilled, then the portal will not be liable for any information posted by its users or any other third party. At the same time, the portal will be required to abide by the duties under the Act and the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.

In all other instances, it is better for the portals to safeguard themselves from potential legal liability, through clear terms of use for all classes of users and clear contracts with entities providing technical support that is required to run the portal. This can be effected through ITA (Section 10A) which recognises the legal validity of electronic contracts.

Recommendations on implementing data policy

- Include clear text stating a project's terms of use, privacy policies and legal policy in easily accessible, understandable formats.
- Ensure informed consent is obtained from all users of the project.
- Project managers may need to include a section on terms and conditions that safeguard them from potential legal liability, in case data is required to be altered such as for standardisation or adherence to taxonomic nomenclatures.

Privacy Concerns in Citizen Science

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

App-enabled smart phones equipped with tools such as cameras, audio-recorders, location-capturing applications are usually preferred means to capture biodiversity related information, which then is uploaded to citizen science platforms (Cartwright, 2016). Additionally, 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.
To compound matters, geo-locations of species are commonly required by biodiversity inventories or projects focusing on co-occurrence data, which in-turn may reveal sensitive information related to endangered species. Information on location of species could lead to poaching, unethical collection or disturbance caused by 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.

Similarly, when observers upload ecological data, they may be required to share geolocations of species of interest, and in-turn end up sharing real-time information pertaining their personal locations, patterns of daily or weekend travel, types of phones used, etc.

Although participants are generally aware of these issues while contributing data (Bowser et al., 2013A), it is still imperative to get informed consent and brief them on the terms of service employed by the project. In a study by Cooper et al., (2019), 51% of projects that did not focus exclusively on people data, often overlooked the fact that they were still collecting PII. The Personal Genome Project (PGP) has been globally acclaimed for its

.

approach to informed consent that transcends traditional boundaries. In the case of this project, the project proponents ensure that all participants pass an examination that tests their knowledge of genomic science and privacy issues, and thereafter sign access to their personal and genomic data for the project (Angrist, 2009).

Recommendations on privacy

- A project's privacy policies should explicitly disclose the potential risks arising from sharing such information.
- Platforms can institute automated or optionally triggered mechanisms such as anonymising records, masking/ hiding sensitive locations, obscuring locations etc. to reduce privacy risks.
- Individual privacy preferences can be offered where possible.

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 a fully informed consent from contributors, and inform them of the exact ways that data contributed by them would be used. Such existing and upcoming legal provisions have potential implications for the privacy of participants in Citizen Science portals (Ganzevoort et al., 2017).

Privacy Policies and the Indian Legal Context

Lately, internet privacy has been gaining much attention in India too, especially with the rising instances of its mention in current affairs in the political context. The internet appears to be driving an increased debate on privacy and awareness in India. There is a blurred line between public and private information, especially in the case of social media posts, with Indian courts not yet having declared whether social media content is public or private information, at the time of writing this (Internet Privacy in India — The Centre for Internet and Society, n.d.). An individual's data is subjected to different levels of protection depending on the jurisdiction it is residing in and therefore accessible to the law enforcement agencies of that jurisdiction. This implies that data residing in a country that is foreign to the Indian contributor is often beyond the jurisdiction of Indian laws and courts. In this context, there have been calls from many in the government and industry asking for the establishment of 'domestic servers' to host the data of Indian users on international platforms.

Currently, the Indian Information Technology Act (ITA), 2000 contains some provisions that only cater toward defining data protection standards for corporations and providing increased access and monitoring to law enforcement agencies without directly addressing the privacy policy concerns of users. The ITA states that any person who has obtained access to material containing personal information of another person, must not disclose such information to any other person and that such disclosure is punishable with imprisonment or fine. The duty of Citizen Science portals to handle the personal information of their users with care flows from this provision.

Since the ITA, a Report of the Group of Experts on Privacy in 2012 defined nine National Privacy Principles that would apply to all data controllers both in the private sector and the public sector, holding them accountable and allowing individuals to seek redress. The principles include

1. Giving clear notice before and on what personal information is collected, including what it will be used for and with whom it will be shared. Notice is also to be provided on data breaches.
2. The ability to explicitly opt-in or opt-out of providing their personal information and to withdraw the information at a later stage.
3. Collect only the necessary information that is required for the purposes.
4. Use the data only for the purpose which was stated at the time of collection and inform users of any change in purpose.
5. Provide individuals access to the data on them and allow correction, amendments, or deletion.
6. Only disclose information to third parties after obtaining consent from users.
7. Provide adequate measures to secure personal information that they have collected, including against loss or unauthorised access.
8. Provide information on policies in clear and plain language in an easily accessible manner to all individuals without discrimination.
9. Be accountable for complying with measures stated in the privacy policy.

The Indian Supreme Court has recognised that privacy in relation to informational privacy is included within the ambit of the Right to Privacy, which in turn is a facet of the Right to Life and Personal Liberty. The principles prescribed by the above Group of Experts were also given recognition by the Supreme Court (Justice K Puttaswamy (Retd) & Anr. v. Union of India (2017) 10 SCC 1). Following this judgement, the Personal Data Protection Bill, 2019 was tabled in the Indian Parliament by the Ministry of Electronics and Information Technology in 2019. This bill is inspired by the GDPR and seeks to overhaul India's current data protection regime. However, the bill has not yet been passed by parliament and is not enforceable as of date.

6. Conclusion

Including all above considerations, every project has to consider its unique situation in terms of biodiversity, explored/unexplored, documented/undocumented, challenges to discover, document, disseminate and threats. Each design is significant on its own, reflecting a socio-ecological-system that technology has to integrate into.

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). We discuss above, various aspects of data quality, various biases that could affect data quality and recommendations for overcoming them.

The sole objective of a citizen science project is not necessarily data and through the duration of the project, it builds the capacity of its participants and inculcates the spirit of scientific endeavour and discovery, while also sensitising them towards species and habitat conservation, creating a sense of stewardship towards nature.

Another challenge with citizen science is in ensuring sustained participation both from citizens as well as from scientists and experts to help validate the data (Irwin, 2018). From this perspective, imposing too much rigour in data collection and quality can reduce inclusivity and lead to reduced participation. As the main objective of citizen science is in involving wider 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 multiple sources often results in duplication of data contributed in multiple places. This is an issue that eventually may 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 ORCID ids can also help in these efforts, although as of now this is not widely used beyond the academic community.

To conform to the expectations of its varied user bases, citizen science has to meet the dual objectives of providing high quality summarised 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 require significant investment in technology solutions, clear data policies and transparency. Anhalt-Depies et al. (2019) provide 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 evaluations 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 which underlie good practice in citizen science (https://eu-citizen.science/about/).

In the Indian context, it would be ideal to envisage 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 data lifecycle. Many act as platforms for data collection, organisation and aggregation but for various reasons, focus less on providing tools to analyse 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, decentralised and federated architectures are imagined and attempted. Such architectures will help address the spatial, temporal and taxon bias and also empower communities in sensitive socio-ecological systems to participate in conservation efforts effectively. Such infrastructures as socio-technical systems could help transform data infrastructures to knowledge infrastructures enhancing biodiversity knowledge commons thereby shaping policies and practises.

7. Bibliography

Angrist, M. (2009). Eyes wide open: the personal genome project, citizen science and veracity in informed consent. Personalized Medicine, 6(6), 691–699. https://doi.org/10.2217/pme.09.48

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

Assumpção, T. H., Popescu, I., Jonoski, A., & Solomatine, D. P. (2018). Citizen observations contributing to flood modelling: opportunities and challenges. Hydrology and Earth System Sciences, 22(2), 1473–1489. https://doi.org/10.5194/hess-22-1473-2018

August, T. A., Pescott, O. L., Joly, A., & Bonnet, P. (2020). AI Naturalists Might Hold the Key to Unlocking Biodiversity Data in Social Media Imagery. Patterns, 1(7), 100116. https://doi.org/10.1016/j.patter.2020.100116

Baker, E., Drury, J. P., Judge, J., Roy, D. B., Smith, G. C., & Stephens, P. A. (2021). The Verification of Ecological Citizen Science Data: Current Approaches and Future Possibilities. Citizen Science: Theory and Practice, 6(1), 12. https://doi.org/10.5334/cstp.351

Balázs, B., Mooney, P., Nováková, E., Bastin, L., & Jokar Arsanjani, J. (2021). Data Quality in Citizen Science. In: Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (Eds.), The Science of Citizen Science. Springer, Cham, pp. 139–157. https://doi.org/10.1007/978-3-030-58278-4_8

Ball, A. (2011). How to License Research Data. DCC How-to Guides. Digital Curation Centre, Edinburgh. Available from: https://www.dcc.ac.uk/guidance/how-guides/license-research-data. [Accessed 15-10- 2022].

Barve, V. V. (2015). Discovering and developing primary biodiversity data from social networking sites. Doctoral dissertation. University of Kansas. Available from: https://kuscholarworks.ku.edu/handle/1808/19011. [Accessed 15-10-2022].

Bowser, A., Cooper, C., de Sherbinin, A., Wiggins, A., Brenton, P., Chuang, T.-R., Faustman, E., Haklay, M. (Muki), & Meloche, M. (2020). Still in Need of Norms: The State of the Data in Citizen Science. Citizen Science: Theory and Practice, 5(1), 18. https://doi.org/10.5334/cstp.303

Bawa, K., Sengupta, A., Chavan, V., Chellam, R., Ganesan, R., Krishnaswamy, J., Mathur, V. B., Nawn, N., Olsson, S., Pandit, N., Quader, S., Rajagopal, P., Ramakrishnan, U., Ravikanth, G., Sankaran, M., Shankar, D., Seidler, R., Shaanker, R., & Vanak, A. (2020). Securing biodiversity, securing our future: A national Mission on biodiversity and human well-being for India. Biological Conservation, 253(1), 1–15. https://doi.org/10.1016/j.biocon.2020.108867.

Boakes, E. H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D. B., & Haklay, M. (2016). Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. Scientific Reports, 6(1), 33051. https://doi.org/10.1038/srep33051

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. BioScience, 59(11), 977–984. https://doi.org/10.1525/bio.2009.59.11.9

Bowser, A., Wiggins, A., & Stevenson, R. D. (2013A). Data Policies for Public Participation in Scientific Research: A Primer. DataONE Public Participation in Scientific Research Working Group. Available from: https://old.dataone.org/sites/all/documents/DataPolicyGuide.pdf. [Accessed 15-10-2022].

Bowser, A., Hansen, D., He, Y., Boston, C., Reid, M., Gunnell, L., & Preece, J. (2013B). Using gamification to inspire new citizen science volunteers. Gamification '13: Proceedings of the First International Conference on Gameful Design, Research, and Applications, 18–25. https://doi.org/10.1145/2583008.2583011

Brenton, P., Gavel, S. von, Vogel, E., & Lecoq, M.-E. (2018). Technology infrastructure for citizen science. In: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 63–80. Available from: http://www.jstor.org/stable/j.ctv550cf2.12. [Accessed 15-10-2022].

Callaghan, C. T., Rowley, J. J. L., Cornwell, W. K., Poore, A. G. B., & Major, R. E. (2019). Improving big citizen science data: Moving beyond haphazard sampling. PLOS Biology, 17(6), e3000357. https://doi.org/10.1371/journal.pbio.3000357

Callaghan, C. T., Poore, A. G. B., Hofmann, M., Roberts, C. J., & Pereira, H. M. (2021). Large-bodied birds are over-represented in unstructured citizen science data. Scientific Reports, 11(1), 19073. https://doi.org/10.1038/s41598-021-98584-7

Campbell, D. L., Thessen, A. E., & Ries, L. (2020). A novel curation system to facilitate data integration across regional citizen science survey programs. PeerJ, 8, e9219. https://doi.org/10.7717/peerj.9219

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. Scientific Data, 8(1), 108. https://doi.org/10.1038/s41597-021-00892-0

Cartwright, J. (2016). Technology: Smartphone science. Nature, 531(7596), 669–671. https://doi.org/10.1038/nj7596-669a

Chen, D., & Gomes, C. P. (2019). Bias Reduction via End-to-End Shift Learning: Application to Citizen Science. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 493–500. https://doi.org/10.1609/aaai.v33i01.3301493

Cooper, C., Larson, L., Holland, K. K., Gibson, R., Farnham, D., Hsueh, D., Culligan, P., & McGillis, W. (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. Citizen Science: Theory and Practice, 2(1), 8. https://doi.org/10.5334/cstp.82

Cooper, C., Shanley, L., Scassa, T., & Vayena, E. (2019). Project Categories to Guide Institutional Oversight of Responsible Conduct of Scientists Leading Citizen Science in the United States. Citizen Science: Theory and Practice, 4(1), 7. https://doi.org/10.5334/cstp.202

Cooper, C. B., Hawn, C. L., Larson, L. R., Parrish, J. K., Bowser, G., Cavalier, D., Dunn, R. R., Haklay, M. (Muki), Gupta, K. K., Jelks, N. O., Johnson, V. A., Katti, M., Leggett, Z., Wilson, O. R., & Wilson, S. (2021). Inclusion in citizen science: The conundrum of rebranding. Science, 372(6549), 1386–1388. https://doi.org/10.1126/science.abi6487

CoreTrustSeal Standards and Certification Board. (2019). CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (v02.00-2020-2022). Zenodo. https://doi.org/10.5281/zenodo.3638211

Courter, J. R., Johnson, R. J., Stuyck, C. M., Lang, B. A., & Kaiser, E. W. (2013). Weekend bias in Citizen Science data reporting: Implications for phenology studies. International Journal of Biometeorology, 57(5), 715–720. https://doi.org/10.1007/s00484-012-0598-7

Crosas, M. (2013, October 30). Joint Declaration of Data Citation Principles—FINAL. FORCE11. https://www.force11.org/datacitationprinciples

de Sherbinin, A., Bowser, A., Chuang, T. R., Cooper, C., Danielsen, F., Edmunds, R., Elias, P., Faustman, E., Hultquist, C., Mondardini, R., Popescu, I., Shonowo, A., & Sivakumar, K. (2021). The Critical Importance of Citizen Science Data. Frontiers in Climate, 3. https://doi.org/10.3389/fclim.2021.650760

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". MindTrek'11: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, 9-15. https://doi.org/10.1145/2181037.2181040.

Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: Citizen science programmes as useful tools for conservation biogeography: Citizen science and conservation biogeography. Diversity and Distributions, 16(3), 354–362. https://doi.org/10.1111/j.1472-4642.2009.00615.x

Downs, R. R., Ramapriyan, H. K., Peng, G., & Wei, Y. (2021). Perspectives on Citizen Science Data Quality. Frontiers in Climate, 3. https://doi.org/10.3389/fclim.2021.615032

Escribano, N., Galicia, D., & Ariño, A. H. (2018). The tragedy of the biodiversity data commons: A data impediment creeping nigher? Database: The Journal of Biological Databases and Curation, 2018, 1–6. https://doi.org/10.1093/database/bay033

Falk, S., Foster, G., Comont, R., Conroy, J., Bostock, H., Salisbury, A., Kilbey, D., Bennett, J., & Smith, B. (2019). Evaluating the ability of citizen scientists to identify bumblebee (Bombus) species. PLOS ONE, 14(6), e0218614. https://doi.org/10.1371/journal.pone.0218614

Faundeen, J., Burley, T. E., Carlino, J. A., Govoni, D. L., Henkel, H. S., Holl, S. L., Hutchison, V. B., Martín, E., Montgomery, E. T., Ladino, C., Tessler, S., & Zolly, L. S. (2014). The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open File Report 2013–1265. 1–4. http://dx.doi.org/10.3133/ofr20131265

Follett, R., & Strezov, V. (2015). An Analysis of Citizen Science Based Research: Usage and Publication Patterns. PLOS ONE, 10(11), e0143687. https://doi.org/10.1371/journal.pone.0143687

Ganzevoort, W., van den Born, R. J. G., Halffman, W., & Turnhout, S. (2017). Sharing biodiversity data: Citizen scientists' concerns and motivations. Biodiversity and Conservation, 26(12), 2821–2837. https://doi.org/10.1007/s10531-017-1391-z

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., & Tøttrup, A. P. (2016). What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Diversity and Distributions, 22(11), 1139–1149. https://doi.org/10.1111/ddi.12477

Gonsamo, A., & D'Odorico, P. (2014). Citizen science: Best practices to remove observer bias in trend analysis. International Journal of Biometeorology, 58(10), 2159–2163. https://doi.org/10.1007/s00484-014-0806-8

Groom, Q., Weatherdon, L., & Geijzendorffer, I. R. (2017). Is citizen science an open science in the case of biodiversity observations? Journal of Applied Ecology, 54(2), 612–617. https://doi.org/10.1111/1365-2664.12767

Guerrini, C. J., Majumder, M. A., Lewellyn, M. J., & McGuire, A. L. (2018). Policy for citizen science. Science, 361(6398), 134–136. https://doi.org/10.1126/science.aar8379

Guerrini, C. J., Lewellyn, M., Majumder, M. A., Trejo, M., Canfield, I., & McGuire, A. L. (2019). Donors, authors, and owners: How is genomic citizen science addressing interests in research outputs? BMC Medical Ethics, 20(1), 84. https://doi.org/10.1186/s12910-019-0419-1

Hagedorn, G., Mietchen, D., Morris, R., Agosti, D., Penev, L., Berendsohn, W., & Hobern, D. (2011). Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. ZooKeys, 150, 127–149. https://doi.org/10.3897/zookeys.150.2189

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., & Porter, J. H. (2013). Big data and the future of ecology. Frontiers in Ecology and the Environment, 11(3), 156–162. https://doi.org/10.1890/120103

Iacovides, I., Jennett, C., Cornish-Trestrail, C., & Cox, A. L. (2013). Do games attract or sustain engagement in citizen science? A study of volunteer motivations. CHI EA '13: CHI '13 Extended Abstracts on Human Factors in Computing Systems, 1101–1106. https://doi.org/10.1145/2468356.2468553

Internet Privacy in India—The Centre for Internet and Society. (n.d.). Available from: https://cis-india.org/telecom/knowledge-repository-on-internet-access/internet-privacy-in-india [Accessed 14 July 2021]

Irwin, A. (2018). No PhDs needed: How citizen science is transforming research. Nature, 562(7728), 480–482. https://doi.org/10.1038/d41586-018-07106-5

Johnston, A., Fink, D., Hochachka, W. M., & Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. Methods in Ecology and Evolution, 9(1), 88–97. https://doi.org/10.1111/2041-210X.12838

Jordan, R. C., Brooks, W. R., Howe, D. V., & Ehrenfeld, J. G. (2012). Evaluating the Performance of Volunteers in Mapping Invasive Plants in Public Conservation Lands. Environmental Management, 49(2), 425–434. https://doi.org/10.1007/s00267-011-9789-y

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

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

Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W. M., Julliard, R., Kraemer, R., & Guralnick, R. (2019). Using Semistructured Surveys to Improve Citizen Science Data for Monitoring Biodiversity. BioScience, 69(3), 170–179. https://doi.org/10.1093/biosci/biz010

Kennett, R., Danielsen, F., & Silvius, K. M. (2015). Citizen science is not enough on its own. Nature, 521(7551), 161–161. https://doi.org/10.1038/521161d

Kimura, A. H., & Kinchy, A. (2016). Citizen Science: Probing the Virtues and Contexts of Participatory Research. Engaging Science, Technology, and Society, 2, 331–361. https://doi.org/10.17351/ests2016.99

Kodali, S. (2017). Not Open or Accountable: The Government Open Data Use License Is Flawed. The Wire [Online]. Available from: https://thewire.in/business/open-data-license-government. [Accessed xx yyyy 2022]

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J., & Kreft, H. (2019). Biodiversity data integration—The significance of data resolution and domain. PLOS Biology, 17(3), e3000183. https://doi.org/10.1371/journal.pbio.3000183

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

Kreitmair, K. V., & Magnus, D. C. (2019). Citizen Science and Gamification. Hastings Center Report, 49(2), 40–46. https://doi.org/10.1002/hast.992

Kühl, H. S., Bowler, D. E., Bösch, L., Bruelheide, H., Dauber, J., Eichenberg, David., Eisenhauer, N., Fernández, N., Guerra, C. A., Henle, K., Herbinger, I., Isaac, N. J. B., Jansen, F., König-Ries, B., Kühn, I., Nilsen, E. B., Pe'er, G., Richter, A., Schulte, R., … Bonn, A. (2020). Effective Biodiversity Monitoring Needs a Culture of Integration. One Earth, 3(4), 462–474. https://doi.org/10.1016/j.oneear.2020.09.010

Lamba, A., Cassey, P., Segaran, R. R., & Koh, L. P. (2019). Deep learning for environmental conservation. Current Biology, 29(19), R977–R982. https://doi.org/10.1016/j.cub.2019.08.016

Landry, B. C., & Rusk, J. E. (1970). Toward a Theory of Indexing—II. Opinion Paper. Journal of the American Society for Information Science, 21(5), 358.

Lemmens, R., Antoniou, V., Hummer, P., & Potsiou, C. (2021). Citizen Science in the Digital World of Apps. In: Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (Eds.), The Science of Citizen Science, Springer, Cham, pp. 461–474. https://doi.org/10.1007/978-3-030-58278-4_23

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

McClure, E. C., Sievers, M., Brown, C. J., Buelow, C. A., Ditria, E. M., Hayes, M. A., Pearson, R. M., Tulloch, V. J. D., Unsworth, R. K. F., & Connolly, R. M. (2020). Artificial Intelligence Meets Citizen Science to Supercharge Ecological Monitoring. Patterns, 1(7), 100109. https://doi.org/10.1016/j.patter.2020.100109

McQuillan, D. (2014). The Countercultural Potential of Citizen Science. M/C Journal, 17(6). https://doi.org/10.5204/mcj.919

Paleco, C., Peter, S. G., Seoane, N. S., Kaufmann, J., & Argyri, P. (2021). Inclusiveness and Diversity in Citizen Science. In Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (Eds.), The Science of Citizen Science, Springer, Cham, pp. 261–281. https://doi.org/10.1007/978-3-030-58278-4_14

Pearce-Higgins, J. W., Baillie, S. R., Boughey, K., Bourn, N. A. D., Foppen, R. P. B., Gillings, S., Gregory, R. D., Hunt, T., Jiguet, F., Lehikoinen, A., Musgrove, A. J., Robinson, R. A., Roy, D. B., Siriwardena, G. M., Walker, K. J., & Wilson, J. D. (2018). Overcoming the challenges of public data archiving for citizen science biodiversity recording and monitoring schemes. Journal of Applied Ecology, 55(6), 2544–2551. https://doi.org/10.1111/1365-2664.13180

Ponti, M., Hillman, T., Kullenberg, C., & Kasperowski, D. (2018). Getting it Right or Being Top Rank: Games in Citizen Science. Citizen Science: Theory and Practice, 3(1), 1. https://doi.org/10.5334/cstp.101

Ratnieks, F. L. W., Schrell, F., Sheppard, R. C., Brown, E., Bristow, O. E., & Garbuzov, M. (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 in Ecology and Evolution, 7(10), 1226–1235. https://doi.org/10.1111/2041-210X.12581

Reiheld, A., & Gay, P. L. (2019). Coercion, Consent, and Participation in Citizen Science. ArXiv:1907.13061 [Physics]. http://arxiv.org/abs/1907.13061

Steen, V. A., Elphick, C. S., & Tingley, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. Diversity and Distributions, 25(12), 1857–1869. https://doi.org/10.1111/ddi.12985

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., Merrifield, M., Morris, W., Phillips, T. B., Reynolds, M., Rodewald, A. D., Rosenberg, K. V., Trautmann, N. M., Wiggins, A., Winkler, D. W., Wong, W.-K., Wood, C. L., Yu, J., & Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. Biological Conservation, 169, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Tiago, P., Ceia-Hasse, A., Marques, T. A., Capinha, C., & Pereira, H. M. (2017). Spatial distribution of citizen science casuistic observations for different taxonomic groups. Scientific Reports, 7(1), 12832. https://doi.org/10.1038/s41598-017-13130-8

The Personal Data Protection Bill, 2019. (n.d.). PRS Legislative Research. Retrieved August 2, 2021, from https://prsindia.org/billtrack/the-personal-data-protection-bill-2019

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. Scientific Reports, 7(1), 9132. https://doi.org/10.1038/s41598-017-09084-6

Turnhout, E., & Boonman-Berson, S. (2011). Databases, Scaling Practices, and the Globalization of Biodiversity. Ecology and Society, 16(1). https://doi.org/10.5751/ES-03981-160135

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. Scientific Reports, 7(1), 9132. https://doi.org/10.1038/s41598-017-09084-6

UNEP, U. N. (2021). 1st Draft of The Post-2020 Global Biodiversity Framework. UNEP - UN Environment Programme. Available from: http://www.unep.org/resources/publication/1st-draft-post-2020-global-biodiversity-framework. [Accessed 15-10-2022].

Veeckman, C., Talboom, S., Gijsel, L., Devoghel, H., & Duerinckx, A. (2019). Communication in Citizen Science: A practical guide to communication and engagement in citizen science. SCIVIL; Leuven, Belgium. Available from: https://eu-citizen.science/resource/52. [Accessed 15-10-2022].

Venkatraman, V. (2010). Conventions of Scientific Authorship. Science [Online]. Available from: https://www.science.org/careers/2010/04/conventions-scientific-authorship. [Accessed15-10-2022].

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (Eds.). (2021). The science of citizen science. Springer, Cham.
Walker, D., McCord, C., Stradiotto, N., Zhou, M. & Singh, D. (2016). Citizen's Guide to Open Data. Available from: https://citizens-guide-open-data.github.io/. [Accessed 15-10-2022].

Walker, D., McCord, C., Stradiotto, N., Zhou, M. & Singh, D. (2016). Citizen's Guide to Open Data. Available from: https://citizens-guide-open-data.github.io/. [Accessed 15-10-2022].

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011). Mechanisms for Data Quality and Validation in Citizen Science. 2011 IEEE Seventh International Conference on E-Science Workshops, 14–19. https://doi.org/10.1109/eScienceW.2011.27

Wiggins, A., Bonney, R., Graham, E., Henderson, S., Kelling, S., Littauer, R., LeBuhn, G., Lotts, K., Michener, W., Greg, N., Russell, E., Stevenson, R., & Weltzin, J. (2013). Data managment guide for public participation in scientific research. DataONE Public Participation in Scientific Research Working Group. Available from: https://osf.io/nfmp4/download. [Accessed 15-10-2022].

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Williams, J., Chapman, C., Leibovici, D. G., Loïs, G., Matheus, A., Oggioni, A., Schade, S., See, L., & Genuchten, P. P. L. van. (2018). Maximising the impact and reuse of citizen science data. In: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 321–336. Available from: http://www.jstor.org/stable/j.ctv550cf2.29 [Accessed 15-10-2022].

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology, 58(4), 479–493. https://doi.org/10.1002/asi.20508