Chapter 8
Data Quality in Citizen Science

Bálint Balázs, Peter Mooney, Eva Nováková, Lucy Bastin, and Jamal Jokar Arsanjani

Abstract This chapter discusses the broad and complex topic of data quality in citizen science – a contested arena because different projects and stakeholders aspire to different levels of data accuracy. In this chapter, we consider how we ensure the validity and reliability of data generated by citizen scientists and citizen science projects. We show that this is an essential methodological question that has emerged within a highly contested field in recent years. Data quality means different things to different stakeholders. This is no surprise as quality is always a broad spectrum, and nearly 200 terms are in use to describe it, regardless of the approach. We seek to deliver a high-level overview of the main themes and issues in data quality in citizen science, mechanisms to ensure and improve quality, and some conclusions on best practice and ways forwards. We encourage citizen science projects to share insights on their data practice failures. Finally, we show how data quality assurance gives credibility, reputation, and sustainability to citizen science projects.

Keywords Peer verification · Expert verification · Quality assessment

B. Balázs (✉)
Environmental Social Science Research Group (ESSRG), Budapest, Hungary
e-mail: balazs.balint@essrg.hu

P. Mooney
Department of Computer Science, Maynooth University, Maynooth, Ireland

E. Nováková
Department of Environmental Geography, Institute of Geonics of the Czech Academy of Sciences, Ostrava-Poruba, Czech Republic

L. Bastin
European Commission Joint Research Centre (JRC), Ispra, Italy

Department of Computer Science, Aston University, Birmingham, UK

J. Jokar Arsanjani
Geoinformatics Research Group, Department of Planning, Aalborg University, Copenhagen, Denmark

K. Vohland et al. (eds.), *The Science of Citizen Science*,
https://doi.org/10.1007/978-3-030-58278-4_8

Introduction

Imagine that a group of city-level stakeholders (a researcher, a citizen, a policymaker, and a business consultant) would like to create a new citizen science project. How can they conceptualise *data accuracy* and design *data quality* protocols? During their planning, they would need to think through a range of issues about the arrangements of their city-level project with unforeseeable knowledge difficulties and reach a collective understanding. However, from the outset of any citizen science project, there are contrasting data needs and motivations. A researcher might look for a level of scientific accuracy to achieve their analytical objective and therefore set thresholds for unreliable data and implement training protocols for volunteers. In contrast, a policymaker may rank avoiding bias in the data of the highest importance, whereas a citizen may require easy to understand data which is relevant to their perceived problem.

How then, even in this hypothetical example, can these different stakeholders create a minimum standard for data quality practices in a citizen science project? It is not an easy task – thousands of citizen science projects have produced extensive data sets that would otherwise be prohibitively expensive to collect. Many citizen science projects produce high-quality data (i.e. accurate, complete, relevant), but some projects are plagued with deficits in data practices: lack of accuracy, no standardised sampling protocol, poor spatial or temporal representation, and insufficient sample size (Anhalt-Depies et al. 2019). This is not unique to citizen science: a 2016 poll by *Nature* of 1500 scientists showed that more than two-thirds had failed to reproduce at least one other scientist's experiment and half of them had even failed to reproduce one of their own results (Baker 2016).

In this chapter, we show that data quality in citizen science is multifaceted and often disputed, with no 'one-size-fits-all' approach. In fact, data quality is the most valued normative claim by citizen science project stakeholders, anchored in multiple levels of expectation. Our focus is on the most typical data quality problems and the generally accepted mechanisms for assessing and verifying the quality of data generated by citizen science. We propose that citizen science project owners can always seek to improve data quality if necessary.

Furthermore, citizen science can learn a lot from purely academic research (basic, applied, or frontier research), for example, from the replication crisis that hits the classic results of social psychology and medicine. Data quality improvements create trade-offs between project resources (time, skills, technology, participants), but there are also protocols, training, and automated solutions to maintain minimum standards of data quality. Moreover, citizen science projects can do more to facilitate the learning among projects by sharing their insights and data quality reports on failures and pitfalls in their data practices.

Coming from various countries in Europe to join the community of practice created by COST Action CA15212 *Citizen Science to Promote Creativity, Scientific Literacy, and Innovation throughout Europe*, the chapter authors have gained their professional experience at the intersections of ecological and social sciences and are

now engaged academics in fields including systems analysis, environmental sociology, land change modelling, geoinformatics, and environmental justice. Citizen science projects have been formative experiences in our lives as researchers. We recognise that academic researchers are now more privileged than ever due to the abundant funding available for professional scientists. In contrast, *volunteer-based citizen science* does not enjoy the same investment. Participants most often find that their greatest challenge is not enough training resources (Turrini et al. 2018; Larson et al. 2020). We identify that, despite the lack of resources, data quality issues are the Achilles heel of citizen science projects. Here we deliver a critical understanding of the positionality of data quality in citizen science and promote an approach to improve citizen science projects.

Science wars and the replication crisis have led to considerable distrust in science, and analysts remind us that we need to face the challenges of the post-truth science era (Saltelli 2018). It has been clear since the inception of citizen science that building up trust with volunteers is difficult due to the structural contradictions of modern science. Data quality and funding (sustainability) of citizen science projects are still the most critical concerns of citizen science practitioners (Hecker et al. 2018). The literature in this field tends to be mostly project specific and provides no framework on how to transform multiple approaches on data quality to more general guidance. Even within a specific domain (e.g. invasive species monitoring), a wide range of approaches and protocols exist. The quality of the collected data may be adequate according to the standards of each project. However, if, using aggregation or meta-analysis, citizen science data from different initiatives are reduced to their minimum common facets or generalised to the lowest common granularity, the resulting data set may no longer meet the original quality thresholds.

Several factors combine to make structuring and forming the focus of data quality discussions in citizen science challenging. Firstly, the growth and popularity of citizen science present citizens, civic society, and governments with multiple challenges and opportunities. New citizen science projects appear daily (Larson et al. 2020). The proliferation of literature in this area is hard to digest: a Google Scholar search using the search terms ‘citizen science’ and ‘data quality’ identifies more than 200 articles published in January–February 2020. However, if existing citizen science projects all have different and potentially incompatible ways of dealing with data quality and sharing data, then the future reuse of project data is significantly impacted. In turn, this has the knock-on effect of making developing ‘follow-on’ citizen science projects from previous projects problematic.

Secondly, the majority of citizen science projects are *contributory* in approach, with three major stages: *data gathering*, *data manipulation*, and *data classification* (Haklay 2013). Some projects are solely quantitative data projects, while others are solely qualitative. Mixed-method citizen science projects also exist which include both quantitative and qualitative data collection, generation, and manipulation. To ensure a minimum standard of data quality, a plan or protocol of data collection (methods) must be set out at the start of a project (Freitag et al. 2016). We consider a dimensionality of data quality needs in both practical and philosophical terms. For

example, in some projects, geographical positional accuracy may not be relevant; in other projects, quality may not relate to data at all (Wiggins et al. 2011).

Thirdly, most citizen science projects have multiple goals, and all must deal with the various legitimacy problems around citizen science. Scientists, funders, authorities, policymakers, and citizens often have different and not always complementary requirements from citizen science data. Veiga et al. (2017) convincingly argued for prioritising data quality needs from the data user's perspective. All citizen science project stakeholders should be invited to co-develop standards for data quality and explicitly state the data quality levels they expect in order to form an agreed approach to data quality.

In summary, the data quality challenge exists at multiple levels. Data quality approaches developed for projects are usually reported when successful, but problems with these approaches are rarely shared or published. Variation in methods of data generation and capture has developed; and, similarly, the potential spectrum of end users, end user applications, and purposes for citizen science data can vary significantly. This leads to a broad range of expectations of data quality (accuracy, temporality, etc.) from varied stakeholders.

In this chapter, we deliver commentaries on five interconnected components of data quality. We begin by asking why is data a critical factor in citizen science projects? Given the wide variation in projects in citizen science and the types of problem domains, we then attempt to set out a definition of data quality in citizen science. Successful examples of high-quality, high-impact data generated by citizen science are plentiful, but what about the hidden cases that are not publicised? Our third commentary discusses the factors which can cause data quality problems in citizen science projects. Validation and verification of all scientific data are important, but how is this performed in citizen science projects? Finally, we discuss how to assure and control data quality in citizen science projects in a flexible, robust, and sustainable manner.

Data as a Risk Factor in Citizen Science Projects

Data from citizen science is unparalleled as it represents evidence that is otherwise difficult for professional science to generate or obtain. Awareness of data quality is growing in citizen science, but it is only one relevant aspect of data accuracy (see Fig. 8.1). Another significant aspect is *data contextualisation*, that is, how citizen science communicates the context in which a particular – often high-volume – data set has been created. Metadata, attribution, and curation are the most prominent examples of data contextualisation. More extensive metadata is helpful to communicate the 'known quality' of the data (Bowser et al. 2015), while *data reuse* is enabled by extensive metadata descriptions of data set purposes and methods of creation. Moreover, data reuse needs to clarify data ownership and future accessibility through open data, open standards, et cetera. This contextualisation is fundamental to understand why data quality is imperative in terms of the goals and

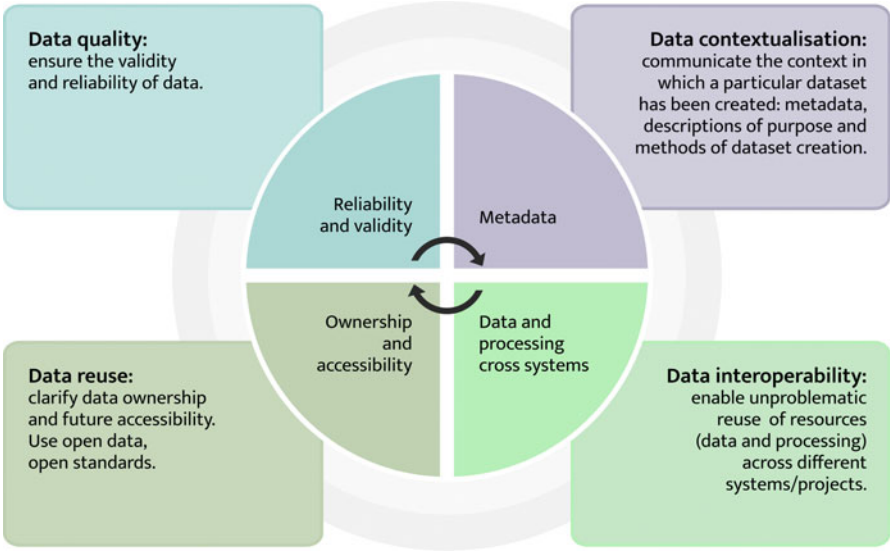

Fig. 8.1 Four aspects of data accuracy in citizen science

objectives of a project. A further aspect is *data interoperability* that enables consistent and straightforward handling of resources (data and processing) across different data sets, systems, and projects.

Citizen science often faces scepticism and distrust from professional scientists and significant resistance from policymakers (Kosmala et al. 2016; Bonney et al. 2014; Nascimento et al. 2018). The main prejudice against citizen science is that it is backward, marginal, and unprofessional; primarily this boils down to weakness in methodology, which can often be the case in professional science as well. On the positive side, citizen science has provided insights into fields such as biology and biodiversity and flora and fauna species and is complementary to traditional data collection methods. Therefore, citizen science as a proper research method should not be neglected by the professional scientific community. Instead, our classical scientific methods need to expand to allow citizen science data to be incorporated and used. This calls for holistic methodological approaches to accommodate citizen science approaches and data practices in the traditional way of studying scientific problems (see more in Pelacho et al., this volume, Chap. 4). In fact, citizen science, alongside technological advancement and increased availability and civic communities invested in solving real-life challenges, has revolutionised our access to more dimensional data. The transformative role of citizen science as an engine for addressing and monitoring Sustainable Development Goals (SDGs) should also be emphasised (Fritz et al. 2019).

For every stakeholder in citizen science, there appears to be a different definition of what constitutes data quality. Numerous terms are used in definitions of data quality, including completeness, availability, standards-based, validity, consistency,

timeliness, accuracy, and bias. This is an illustration of a socio-technical artefact with (hard) physical and (soft) social properties that gains acceptance from humans (volunteers) and machines (artificial intelligence). Several examples will be presented showing machine and human failure as well as soft and hard validation tools.

While it may be hard to agree on an acceptable level of data quality in any given citizen science project, in practical-methodological terms, we can start with known quality, fitness for purpose, and intended use (e.g. in operations, decision-making, or planning). However, from an epistemological point of view, the question is how accurately does the data represent the real-world constructs to which they refer. Real-world constructs are often not clearly defined at the project design stage, so *toolkits* that compare off-the-shelf protocols are helpful.

Data quality is valued from various perspectives, and its levels vary (Lewandowski and Specht 2015; Williams et al. 2018). In terms of data collection, precision and accuracy are the most important aspects. In data processing, it is vital to have consistency in data sets over time. For data analysis, data sets must have adequate representation and distribution of the target population or area. From a more general research design perspective, the validity and the reliability of data are most important (e.g. Lewandowski and Specht 2015).

Reliability implies long-term stability and consistency of data. Data results should be able to be replicated repeatedly; this is necessary in most citizen science projects operating large data sets. Reliability of data ensures citizen science is trusted and aligns with policy requirements and stakeholders' interests. However, citizen science data is valid only if it signifies what it is supposed to. Data validity in science has many aspects including accuracy, confidence, completeness, and error-freeness. There are an increasing number of articles on citizen science data quality in academic literature (Purdam 2014; Riesch and Potter 2014). Suggested data quality definitions converge around sets of characteristics; this leads to heuristic approaches that illustrate the need for a *data quality review toolkit* – a harmonised approach to data quality assurance across different citizen science projects.

Data Quality Issues in Citizen Science Projects

In this section, to illustrate the characteristics of data quality in citizen science, we present some examples of how and where data quality problems can arise in citizen science projects. In order to structure these examples in a meaningful way, we illustrate these data quality problems using the following categories:

1. Data collection protocols are not followed by participants.
2. Data collection protocols do not match the goals of the project or the probable participants.
3. Data collection protocols are incorrectly implemented.

4. Data collection protocols are not comprehensive and are used by stakeholders with different data quality expectation levels.
5. Data used are not fit for purpose.

While these five categories are by no means exhaustive, we believe that they represent a good cross-section of the most commonly encountered issues around data quality in citizen science (Lukyanenko et al. 2016).

Data Collection Protocols Are Not Followed by Participants

Citizen science projects must follow complex data collection protocols. In many cases, volunteers stop participating in projects as they do not know how to collect data using these protocols. Other authors have reported that participants often indicate that they are less concerned about the aims of the project or are unaware of the potential end uses of project data and are only interested in participation. This is obviously a training and communication issue. It is important to explain why a specific protocol has been chosen; what the project data can be used for; and what impact quality has on these end uses. In many cases, the best available strategy is to simplify *user interface design* in data collection tools and make these tools engaging and compatible with the variety of skills and motives of potential citizen scientists (Danielsen et al. 2014). Citizen science toolkits have been developed in many different contexts to facilitate better user engagement as well as the design and delivery of citizen science projects (Kelly et al. 2019). Finally, citizen science projects should incorporate more intuitive data practice considerations to allow users to directly or indirectly follow protocols.

Data Collection Protocols Do Not Match the Goals of the Project or the Probable Participants

Often, protocols for data collection are either too complicated or too simple. In the case of Galaxy Zoo, originally only three categories were listed, but later an additional two categories were added. The protocol did not allow for adding new values, such as discovering new shapes of galaxies; this oversight could have significantly diminished data quality (Lukyanenko et al. 2016). Citizen scientists can miss important data which should be recorded or observed if the protocols are inflexible. Overcomplicated protocols can result in reducing the sense of fun and participation for many citizen scientists by introducing seemingly onerous and systematic rules and tasks. A possible solution is to introduce a permanent channel or forum that participants can use to contact creators and provide input. Finally, making data collectors' tasks more straightforward by pre-filling files with

often-used values or providing examples for observations is an effective way to create better engagement and fulfilment for citizen scientists.

Data Collection Protocols Are Incorrectly Implemented

In citizen science, as in any research context, data quality can quickly deteriorate when the protocols are inaccurate and poorly implemented or do not reflect the relevant context. Often, the lack of 'do not know' or 'unsure' reporting options or fields can lead to false precision levels or recording of invalid values, for example, a value of 0 mm for a rainfall recording gauge which is broken and has not recorded any rainfall. This is a typical example when uncertainty is created without visibility. When devices or sensors are not well calibrated and present inaccurate observations, then data can be misplaced or misreported (Bell et al. 2013). This has severe downstream effects for the analysis of these data sets.

Many citizen science projects use *smart devices* for the collection of data. These devices can introduce technological problems such as the lack of a GPS signal or Internet connection and poor device quality (Bell et al. 2013) which can subsequently result in missing data. Different instruments and collection systems also often apply contrasting transformations to data before submission (e.g. automated altitude correction in some weather stations) which can hinder the accuracy of data (Bell et al. 2013). There are various solutions to these false protocol deployments, for example, by the thorough profiling of data scope, experimental pilots, and iterative development (see examples later in the chapter). Overall, it is essential to apply a common-sense approach to citizen science communities facilitating the reuse of successful data quality protocols. There is little value in constantly reinventing protocols for similar problems being tackled by other citizen science groups or projects.

Data Collection Protocols Are Not Comprehensive and Are Used by Stakeholders with Different Data Quality Expectation Levels

It is natural that authoritative bodies and other stakeholders seek the highest level of data quality for their applications and purposes. Different levels of data quality expectations can lead to tensions between the producers and consumers of citizen science data. Managing expectations of quality is a difficult proposition. Some authoritative bodies dealing with citizen science may only require a simple data protocol be used by citizen scientists. The reasoning for this is to maximise the data quality citizen scientists are capable of collecting. On the other hand, other authoritative bodies may implement complex scientific data collection protocols as they

require citizen scientists to collect detailed data. This has the effect of causing data quality to become a contested matter. Different stakeholders can claim that protocols are obsolete or irrelevant or that the data collected does not match the high expectations of more complex protocols.

The design of data collection protocols can also lead to spatial inequality where different geographical areas or regions receive proportionally more or less attention from citizen scientists, for example, urban areas being favoured over rural areas. Poorly designed or overly complex protocols can also create skill inequality if some protocols assume a specific level of scientific training before they can be used. This carries the risk of overly complex protocols excluding whole (social) groups and, in the case of international citizen science research, excluding countries or even continents.

Data Used Are Not Fit for Purpose

One of the most common and easily understood data quality issues is when data are used for purposes they are not suitable or fit for. This often happens with quantitative data. A phenomenon which is easy to measure may be inappropriately used as a proxy for the phenomenon that needs to be monitored (e.g. wetland acreage vs. wetland quality, Dale and Gerlak 2007). This misuse of data is not confined to the citizen science context, but it is more likely to occur where data documentation is imperfect or incomplete. Negative outcomes (Hunter et al. 2013) from citizen science projects can lead to overcorrection, which can in turn lead to errors and suspicion of all citizen science data. Misuse of citizen science data has caused many in the scientific community to perceive citizen science data as not worthy of being considered serious scientific research (Delaney et al. 2008). Appropriate documentation and metadata are the most effective and appropriate deterrents against using data for unsuitable purposes.

Validation and Verification of Data in Citizen Science Projects

Many citizen science projects collect valuable, high-quality scientific data. The data is subject to *validation* and *verification* before being used. Multiple socio-technical mechanisms can be deployed in citizen science projects to ensure the collection of high-quality data (Freitag et al. 2016). Validating the data in citizen science projects happens both during and after the project has generated data. Freitag and Pfeffer (2013) observe that often the process of a citizen science project is more successful than the product (data) – 'some citizen scientists point out that the data is "good enough" or "were not the main focus of the program"'. They further remark that this

is in stark contrast with many published studies, many of which discuss citizen science as a method, evaluated against traditional methods by the same metric of success – data quality (Riesch and Potter 2014). Therefore, validation or verification methods are required for the data generated, collected, and managed by citizen science projects. As for validation and verification methodologies, several prominent approaches have emerged. These approaches do not belong exclusively in citizen science projects but apply to a range of other application domains such as crowdsourcing, citizen sensing, et cetera. Consequently, we consider four approaches: peer verification, expert verification, automatic quality assessment, and model-based quality assessment.

Peer verification involves experienced project participants (*peers*) helping to identify and validate observations and data provided by new or inexperienced participants. Ideally, quality standards are maintained by the peers to improve performance and provide credibility. This approach is dependent on the community within the citizen science project. It can also have the effect of slowing down the process of data collection as extra time is required for peer verification. Similar to the process of peer review on Wikipedia, the main goal is self-regulation by qualified members within the relevant domain and a convergence towards shared narratives on data quality. For more examples see Liu and Ram (2018), Johnson et al. (2016), and Segal et al. (2015).

Expert verification differs from peer verification. Here, specific contributors or stakeholders are identified as experts within a citizen science project. These experts then verify the data which is generated or collected by other participants. This approach is frequently used by biological surveys. Once the needs of data usability are defined, solutions for data quality can be formulated for expert verification. Continuous expert assistance is required. Examples include iNaturalist, Young et al. (2019), Falk et al. (2019), and Bayraktarov et al. (2019).

Automatic quality assessment involves the use of software-based systems to automatically carry out a quality assessment of the data generated or collected by a citizen science project. There is a wide range of approaches, such as data mining algorithms, which filter and search for problematic data, statistical analysis (plausibility of data), and qualifying systems. As *artificial intelligence* (AI) approaches become more sophisticated and are more readily available in software, these can be used to carry out more resource-intensive automated quality assessments. Examples include Njue et al. (2019), Wiggins et al. (2011), and Wessels et al. (2019).

Model-based quality assessment goes beyond automatic filtering techniques which can address random variation (e.g. unsupervised data mining or naive outlier detection) and tackles residual errors using an explicit model of how the phenomenon of interest is expected to vary in space or time. This requires a concrete understanding of how the relevant phenomena behave and appropriate experts are required. This approach can be more effective in establishing the statistical relevance of false positives and false negatives and extreme or unexpected values in a data set. Examples include de-biasing procedures and generation of contributor ratings, based on identified sources of systematic errors in the archive of observations. Examples include Bamford et al. (2009) and Kelling et al. (2015).

When Does Validation Occur?

The methodologies described above must be applied at specific stages in the data collection or generation process within citizen science projects. There are a number of key stages where validation can occur. We summarise these below and indicate the type of validation methodology which can be used at each stage.

At the Project Planning and Design Stage At this stage, there is an opportunity to reduce the number of erroneous contributions. For example, is the accuracy of the location of an object to within 100 m acceptable, is a plant identification to genus level useful, etc. Approach used: expert verification.

During the Project While citizens are actively collecting and generating data, it can be difficult to validate data. However, a number of tactics can be used. These include flagging outliers or potentially erroneous contributions; providing useful and understandable help sections and guides within software apps and websites used by the contributors; access to online suggestion systems which can automatically suggest a class or label and provide automated feedback on submissions (van der Wal et al. 2016); and correcting or updating of contributions by peer contributors, for example, by requesting additional content (photos, free text, etc.) which might help with ambiguous contributions. Approaches used: peer validation, automatic quality assessment, and model-based quality assessment.

After the Project (Before Data Publication) At this stage, there are still opportunities and resolve to identify data quality issues. Remaining outliers can be automatically detected and flagged (e.g. by GeoWIKI, GBIF, eBird); experts can respond to requests for checking (iNaturalist, eBird); and estimates of observer skill or reliability can be calculated (this can be updated based on their history of contribution and used to weigh the value of their submitted data; see Kelling et al. 2015). Approaches used: expert verification, peer validation, automatic quality assessment, and model-based quality assessment.

After the Project (After Data Publication) While end users and stakeholders may already be using available versions of the data generated or collected by a project, post-activity quality assessment is still possible. Experts and peers can change or correct contributions on an ongoing basis (e.g. OpenStreetMap). Iterative corrections or changes can be applied to project design, for example, if data mining identifies a systematic bias in contributions. Indeed, iterative corrections can be also applied in the earlier project stages (via training materials, adapted keys, and applying improvement suggestions in real time). Approaches used: expert verification and peer validation.

Data Quality Assurance and Control in Citizen Science

Data are considered reliable if the methods by which they are collected and analysed remain stable over time. Data quality assurance plans and control are strategies implemented to reduce estimation error and bias; measurement error and bias; and data processing errors. In a survey of 30 citizen science project leaders, conducted by Freitag et al. (2016), 12 strategies for credibility building in citizen science were identified. Three of these are applied during the training and planning phase, four are applied during the data collection phase, and five are applied during the data analysis and project evaluation phase. The variation in the application of these strategies is due to factors including the number of participants in the project, the focus on group versus individual work, and the time commitment of participants. In this sense, data quality assurance and control must be adapted to the specific citizen science project under assessment. The literature indicates a number of different approaches to data quality assurance and control.

Meek et al. (2014) identify three types of quality assurance models: the *producer model*, the *consumer model*, and the *stakeholder model*. Their data quality assessment is based on seven steps in a workflow:

1. *Location-based services positioning* redirects users towards areas that are of interest to project organisers.
2. *Data cleaning* removes erroneous entries.
3. *Automatic validation* carries out preliminary credibility checks on the data collected.
4. *Comparison with authoritative data* improves the confidence and validity of collected data.
5. *Model-based validation* compares crowd data with data from models or previously validated crowdsourced data.
6. *Linked data analysis* combines the wealth of freely available data (big data) and associated data mining techniques to establish data confidence and quality.
7. *Semantic harmonisation* transforms input data to ensure conformance to or enrichment of an ontology.

All these steps produce inputs for each of the three (producer, consumer, stakeholder) models of quality assurance.

Clare et al. (2019) defined an iterative and adaptive data evaluation process in a six-step sequential framework (see Fig. 8.2). Three steps are about data quality assurance:

1. Define desired data quality explicitly in terms of study objectives grounded in specific analyses or estimates.
2. Estimate existing levels of accuracy or error within the data set.
3. Estimate a requisite level of accuracy or error within the raw data that allows study objectives to be achieved.

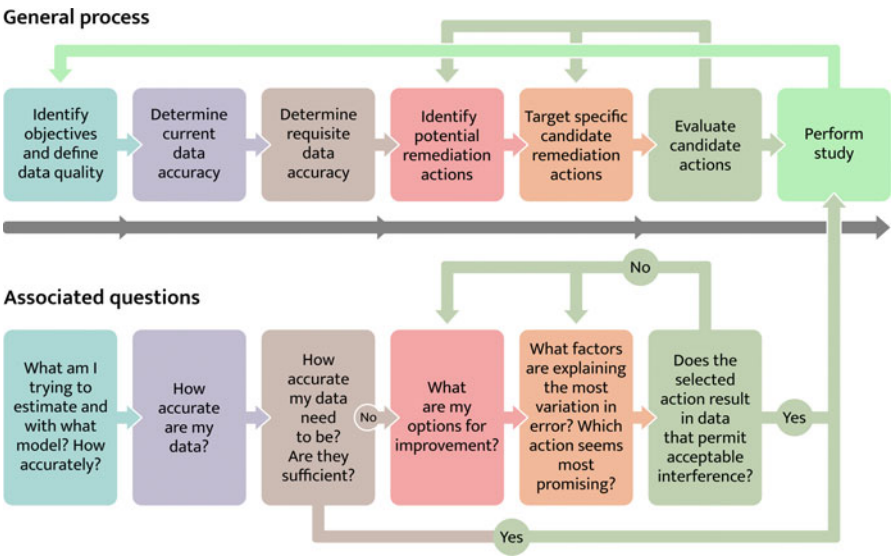

Fig. 8.2 Six steps of data evaluation from Clare et al. 2019

The remaining three steps are about data quality control:

4. Identify possible remedial actions.
5. Explore sources of variation in errors within a data set to target a specific action or set of actions to evaluate.
6. Implement and evaluate candidate actions to determine whether any meets the defined data quality objective.

Data quality assurance and control in citizen science can be conducted using two main strategies: (1) the *upstream* (assuring) strategy, which includes a set of actions that assure the quality of citizen science data to a certain level, or (2) the *downstream* (controlling) strategy, which includes a set of actions that controls the quality of citizen science and learns from earlier failures. Let us now consider some examples of both data quality assurance and control in order to illustrate these concepts more clearly.

Assuring data quality requires a set of criteria that pre-emptively restrict data inputs, such as:

- *Profiling* which assesses the data collectors to understand the quality challenges, including the impact of uncertainty in contributions and how it can be captured or traced.
- *Pre-testing* includes gathering sample data before a citizen science project begins using both expert and beginner contributors. This can help identify unforeseen sources of errors or other problems that can be fixed before the project starts.
- *Standardisation* ensures that expected data conform to quality rules and domain-relevant schemas.

- *On-the-fly data correction or cleansing tools* allow for auto-correction of some errors prior to reporting, for example, autocorrecting geocoding of address data, topology checks, and enforcing selection of an attribute value from a dictionary list.
- *Matching or linking* facilitates aligning or merging similar data records which can help avoid data redundancy.

Controlling data quality includes a set of actions that allow for controlling data quality after the project has started, such as:

- *Triangulation* which combines multiple criteria and methods to ensure data quality (Wiggins et al. 2011).
- *Recursive monitoring* keeps track of data quality over time and generates reports on uncertainties and variations. These reports can be used to maintain or improve data quality as well as provide feedback for project design.
- *Training participants* results in participants understanding data quality and appreciating the minimum data quality requirements for every citizen science project.
- *Protocols and standards for consistency* are followed to make the collected data consistent and homogeneous. Usage of protocols and standards should not adversely affect engagement levels of citizen scientists.
- *Compatible information systems* allow for long-term storage, curation, and archiving of data from citizen science projects.
- *Usage of international standards* such as ISO19115, ISO19157, and ISO8000 is recommended as a point of reference for quality control of citizen science projects.
- *Collect and release data under open science principles and open-access licences* which follow FAIR (findable, accessible, interoperable, reusable) principles. This allows for unrestricted data access and allows the data to be reused. Using FAIR principles maximizes the value of the data.
- *Record and communicate quality assurance practices*, as narrative descriptions of citizen science quality practices are often missing. This information should be provided in the description or metadata of a project or data set so that similar failures can be avoided in the future.

Conclusions and Recommendations

This chapter has discussed data quality in citizen science and approaches to ensure the validity and reliability of data generated by citizen scientists and citizen science projects. Data quality in citizen science has become a crowded and contested landscape in recent years, as various citizen science projects and their stakeholders often claim and seek different levels of data quality. Therefore, the meaning of data quality differs according to the type of project and its stakeholders. We certainly make no claims as to the exhaustive nature of the discussions in this chapter. Our

focus has been to consider what data quality is in citizen science and how data quality problems occur and to present some of the most popular and well-accepted mechanisms for assessing and verifying data quality. Most citizen science projects employ multiple mechanisms to ensure data quality. The selected mechanisms are driven in no small part by the resources available, the project type and structure, and the needs of stakeholders. Every project can seek to improve data quality. There are always places where one can improve the process to have better data quality (if it is needed).

Success criteria in citizen science are defined by mission statements that guide projects, which are more likely to emphasise the scientific process than the results (Freitag and Pfeffer 2013). Different disciplines will have different conventions around defining data quality and acceptable measures or levels of data quality. However, many scientific disciplines collect similar types of data but do so in varied ways. Consequently, there is no one-size-fits-all approach. It is this diversity and breadth of application which makes data quality in citizen science such a tantalising subject to tackle. Improving data quality always involves trade-offs. Given that there are many moving parts to any citizen science project, it can require additional resources (time, skills, technology, participants, etc.) to deal with the data quality issues identified. Overall, we find that most studies agree that to improve data quality, several approaches are necessary: adaptable project aims and survey protocols; volunteer training; the use of experts; automated and statistical analyses; and finding an appropriate project structure (e.g. volunteer recruitment and retention, overall management) (Lewandowski and Specht 2015).

With abundant literature and examples of data quality approaches in citizen science projects, how do we proceed in order to meaningfully contribute to the data quality discussion? We believe that problems about data quality are rarely shared between citizen science projects. There is often little scope for new projects to learn from existing projects in terms of best practice approaches. Avoiding the same pitfalls as previous or existing projects can go a long way towards ensuring the data quality goals of a project are achieved and maintained. There are many useful lessons relevant to data quality, for example, unforeseen problems with devices, suppliers, and volunteers or unintended consequences of training methods and use of advanced technologies such as AI. However, not only are these stories unlikely to be published in an environment where future funding depends on demonstrating success, but they are subjective narratives which do not clearly fit into the available structured options for data quality reporting. Unfortunately, this means that the same problems related to data quality continue to be repeated. As well as sharing insights on data quality pitfalls in citizen science projects, there is also a need to convey successful data quality approaches. Ensuring data quality in a citizen science project should not be regarded as a burden; it can enhance the reputation of the project, make the outputs (re)usable for a broad range of end users and applications, and contribute to higher levels of citizen engagement and long-term project sustainability. In addition to establishing credibility and trust, communicating data quality practices can help citizen science collaboration by identifying shared issues and concerns.

References

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.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*, 452–454. https://doi.org/10.1038/533452a.

Bamford, S. P., Nichol, R. C., Baldry, I. K., Land, K., Lintott, C. J., Schawinski, K., et al. (2009). Galaxy Zoo: The dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society, 393*(4), 1324–1352. https://doi.org/10.1111/j.1365-2966.2008.14252.x.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E. L., Nguyen, H. A., McRae, L., et al. (2019). Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution, 6*, 239. https://doi.org/10.3389/fevo.2018.00239.

Bell, S., Cornford, D., & Bastin, L. (2013). The state of automated amateur weather observations. *Weather, 68*(2), 36–41. https://doi.org/10.1002/wea.1980.

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Citizen science. Next steps for citizen science. *Science (New York, N.Y.), 343*(6178), 1436–1437. https://doi.org/10.1126/science.1251554.

Bowser, A., Shilton, K., & Preece, J. (2015). *Privacy in citizen science: An emerging concern for research & practice*. Paper presented at the Citizen Science 2015 conference, San Jose, CA.

Clare, J. D. J., Townsend, P. A., Anhalt-Depies, C., Locke, C., Stenglein, J. L., Frett, S., et al. (2019). Making inference with messy (citizen science) data: When are data accurate enough and how can they be improved? *Ecological Applications, 29*(2), e01849. https://doi.org/10.1002/eap.1849.

Dale, L., & Gerlak, A. K. (2007). It's all in the numbers: Acreage tallies and environmental program evaluation. *Environmental Management, 39*, 246–260. https://doi.org/10.1007/s00267-005-0332-x.

Danielsen, F., Topp-Jørgensen, E., Levermann, N., Løvstrøm, P., Schiøtz, M., Enghoff, M., & Jakobsen, P. (2014). Counting what counts: Using local knowledge to improve Arctic resource management. *Polar Geography, 37*(1), 69–91. https://doi.org/10.1080/1088937X.2014.890960.

Delaney, D. G., Sperling, C. D., & Adams, C. S. (2008). Marine invasive species: Validation of citizen science and implications for national monitoring networks. *Biological Invasions, 10*, 117–128. https://doi.org/10.1007/s10530-007-9114-0.

Falk, S., Foster, G., Comont, R., Conroy, J., Bostock, H., Salisbury, A., et al. (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.

Freitag, A., & Pfeffer, M. J. (2013). Process, not product: Investigating recommendations for improving citizen science 'success'. *PLoS One, 8*(5), e64079. https://doi.org/10.1371/journal.pone.0064079.

Freitag, A., Meyer, R., & Whiteman, L. (2016). Strategies employed by citizen science programs to increase the credibility of their data. *Citizen Science: Theory and Practice, 1*(1), 2. https://doi.org/10.5334/cstp.6.

Fritz, S., See, L., Carlson, T., Haklay, M., Oliver, J. L., Fraisl, D., et al. (2019). Citizen science and the United Nations Sustainable Development Goals. *Nature Sustainability, 2*, 922–930. https://doi.org/10.1038/s41893-019-0390-3.

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing geographic knowledge* (pp. 105–122). Dordrecht: Springer.

Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (Eds.). (2018). *Citizen science: Innovation in open science, society and policy*. London: UCL.

Hunter, J., Alabri, A., & van Ingen, C. (2013). Assessing the quality and trustworthiness of citizen science data. *Concurrency and Computation: Practice and Experience, 25*(4), 454–466. https://doi.org/10.1002/cpe.2923.

Johnson, I. L., Lin, Y., Li, T. J. -J., Hall, A., Halfaker, A., Schöning, J., & Hecht, B. (2016). *Not at home on the range*. In Proceedings of the 2016 CHI conference on Human Factors in Computing Systems, July 2016. https://doi.org/10.1145/2858036.2858123.

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

Kelly, R., Fleming, A., & Pecl, G. T. (2019). Citizen science and social licence: Improving perceptions and connecting marine user groups. *Ocean & Coastal Management, 178*, 104855. https://doi.org/10.1016/j.ocecoaman.2019.104855.

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.

Larson, L. R., Cooper, C. B., Futch, S., Singh, D., Shipley, N. J., Dale, K., et al. (2020). The diverse motivations of citizen scientists: Does conservation emphasis grow as volunteer participation progresses? *Biological Conservation, 242*. https://doi.org/10.1016/j.biocon.2020.108428.

Lewandowski, E., & Specht, H. (2015). Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology, 29*(3), 713–723. https://doi.org/10.1111/cobi.12481.

Liu, J., & Ram, S. (2018). Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering, 115*, 80–93.

Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2016). Emerging problems of data quality in citizen science. *Conservation Biology, 30*(3), 447–449.

Meek, S., Jackson, M. J., & Leibovici, D. G. (2014). A flexible framework for assessing the quality of crowdsourced data. In J. Huerta, S. Schade, & C. Granell (Eds.), *Connecting a digital Europe through location and place. Proceedings of the AGILE'2014 international conference on Geographic Information Science*. Castellón: AGILE Digital Editions.

Nascimento, S., Rubio Iglesias, J. M., Owen, R., Schade, S., & Shanley, L. (2018). Citizen science for policy formulation and implementation. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, & A. Bonn (Eds.), *Citizen science – Innovation in open science, society and policy* (pp. 219–240). London: UCL Press.

Njue, N., Stenfert Kroese, J., Gräf, J., Jacobs, S. R., Weeser, B., Breuer, L., & Rufino, M. C. (2019). Citizen science in hydrological monitoring and ecosystem services management: State of the art and future prospects. *Science of the Total Environment, 693*, 133531. https://doi.org/10.1016/j.scitotenv.2019.07.337.

Purdam, K. (2014). Citizen social science and citizen data? Methodological and ethical challenges for social research. *Current Sociology, 62*(3), 374–392. https://doi.org/10.1177/0011392114527997.

Riesch, H., & Potter, C. (2014). Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. *Public Understanding of Science, 23*(1), 107–120. https://doi.org/10.1177/0963662513497324.

Saltelli, A. (2018). Why science's crisis should not become a political battling ground. *Futures, 104*, 85–90. https://doi.org/10.1016/j.futures.2018.07.006.

Segal, A., Gal, Y., Simpson, R. J., Homsy, V., Hartswood, M., Page, K. R., & Jirotka, M. (2015). *Improving productivity in citizen science through controlled intervention*. In Proceedings of the 24th international conference on World Wide Web – WWW 15 Companion. https://doi.org/10.1145/2740908.2743051.

Turrini, T., Dörler, D., Richter, A., Heigl, F., & Bonn, A. (2018). The threefold potential of environmental citizen science – Generating knowledge, creating learning opportunities and enabling civic participation. *Biological Conservation, 225*, 176–186. https://doi.org/10.1016/j.biocon.2018.03.024.

van der Wal, R., Sharma, N., Mellish, C., Robinson, A., & Siddharthan, A. (2016). The role of automated feedback in training and retaining biological recorders for citizen science. *Conservation Biology, 30*(3), 550–561. https://doi.org/10.1111/cobi.12705.

Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., & Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLoS One, 12*(6). https://doi.org/10.1371/journal.pone.0178731.

Wessels, P., Moran, N., Johnston, A., & Wang, W. (2019). Hybrid expert ensembles for identifying unreliable data in citizen science. *Engineering Applications of Artificial Intelligence, 81*, 200–212. https://doi.org/10.1016/j.engappai.2019.01.004.

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011). *Mechanisms for data quality and validation in citizen science*. In 2011 IEEE seventh international conference on e-Science Workshops (pp. 14–19). https://doi.org/10.1109/eScienceW.2011.2.

Williams, J., Chapman, C., Leibovici, D., Loïs, G., Matheus, A., Oggioni, A., Schade, S., See, L., & van Genuchten, P. (2018). Maximising the impact and reuse of citizen science data. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, & A. Bonn (Eds.), *Citizen science – Innovation in open science, society and policy* (pp. 321–336). London: UCL Press.

Young, B. E., Dodge, N., Hunt, P. D., Ormes, M., Schlesinger, M. D., & Shaw, H. Y. (2019). Using citizen science data to support conservation in environmental regulatory contexts. *Biological Conservation, 237*, 57–62. https://doi.org/10.1016/j.biocon.2019.06.016.

Bálint Balázs is a senior research fellow and managing director of the Environmental Social Science Research Group (ESSRG). With a background in sociology and environmental sciences, he has research experience in sustainability transitions, policy analysis on sustainable food, cooperative research, public engagement, science-policy dialogues, and participatory action research.

Peter Mooney is an assistant professor in computer science at Maynooth University, Ireland. He has international research experience in a number of areas related to geocomputation and data science, including quality of geographical data, data mining techniques for pattern extraction, and engaged research between the general public, government, and academia.

Eva Nováková is a scientific worker at the Czech Academy of Sciences. Her research interest focuses on geography of energies, mostly renewable energy development and land use conflicts. Further she focuses on data visualisation, processing of qualitative data, and citizen involvement.

Lucy Bastin has a background in spatial ecology and remote sensing, software development for environmental applications, and interoperable standards to communicate uncertainty in scientific models and workflows. She leads development on DOPA (the Digital Observatory for Protected Areas) at the EC Directorate for Knowledge for Sustainable Development and Food Security.

Jamal Jokar Arsanjani is a professor of geographical information science at the Department of Planning, Aalborg University. His research interest is focused on using earth observations and citizen observations for land use/cover change modelling as well as exploring its driving forces. He serves as the editor-in-chief of *Data* journal and is on several journal editorial boards.