Perspective: The Power (Dynamics) of Open Data in Citizen Science

Caren B. Cooper [1]*, Lisa M. Rasmussen [2] and Elizabeth D. Jones [1]

[1] Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, United States,
[2] Department of Philosophy, University of North Carolina, Charlotte, NC, United States

In citizen science, data stewards and data producers are often not the same people. When those who have labored on data collection are not in control of the data, ethical problems could arise from this basic structural feature. In this Perspective, we advance the proposition that stewarding data sets generated by volunteers involves the typical technical decisions in conventional research plus a suite of ethical decisions stemming from the relationship between professionals and volunteers. Differences in power, priorities, values, and vulnerabilities are features of the relationship between professionals and volunteers. Thus, ethical decisions about open data practices in citizen science include, but are not limited to, questions grounded in respect for volunteers: who decides data governance structures, who receives attribution for a data set, which data are accessible and to whom, and whose interests are served by the data use/re-use. We highlight ethical issues that citizen science practitioners should consider when making data governance decisions, particularly with respect to open data.

Keywords: ethics, volunteer monitoring, data stewardship, data privacy, community science

INTRODUCTION

One aspect of open science involves sharing scientific data broadly, maximizing its power to benefit society through use and re-use in other research. In conventional environmental research, professional scientists generate data and make decisions about stewardship of resulting data sets. In contrast, in research through citizen science, those who generate data are not likely to be those making stewardship decisions about it. Consequently, the loss of volunteer control could lead to greater potential harms to data producers in citizen science from decisions about data use/re-use. Ethical conundrums arise when different parties (scientists and volunteers) have conflicting interests in relation to the data governance. Given the power differentials between scientists and volunteers, and irrespective of whether some parties have legal rights to control access to and use of the data, responsible research requires attention to the interests of all stakeholders (Ballantyne, 2018).

In this Perspective, we adopt the premise that professional scientists should steward data for its maximal use in advancing science via open data practices. We advance the proposition that stewarding data sets generated by volunteers involves the typical technical decisions in conventional research plus a suite of ethical decisions stemming from the relationship between professionals and volunteers. Differences in power, priorities, values, and vulnerabilities are features

of the relationship between professionals and volunteers. Thus, ethical decisions about open data practices in citizen science include, but are not limited to, questions grounded in respect for volunteers: who decides data governance structures, who receives attribution for the data set, which data are accessible and to whom, and whose interests are served by the data use/re-use.

In our recent work, supported by the National Science Foundation, we aim to provide practitioner-built tools to identify and facilitate ethical data practices in citizen science. In 2017, we held an interdisciplinary workshop about ethics in citizen science (Lisa M. Rasmussen: NSF SES-1656096, Filling the "Ethics Gap" in Citizen Science Research). The workshop brought together nearly three dozen attendees involved with citizen science, research ethics, and Science and Technology Studies to consider the novel ethical challenges posed by citizen science research. Workshop aims included identifying ethical issues in citizen science, articulating conceptual frameworks for them, and brainstorming possible solutions. The workshop yielded a list of over 40 ethical issues related to the practice of citizen science, many of which were explored in papers in two special collections: one in the journal of the Citizen Science Association, *Citizen Science: Theory and Practice* (Rasmussen and Cooper, 2019), and one in *Narrative Inquiry in Bioethics* (Rasmussen, 2019). Some of the topics related to different aspects of data acquisition and management.

The workshop findings informed a plan for research and facilitation to develop norms, and resources and tools to support those norms, around responsible, trustworthy data practices in citizen science (Caren B. Cooper: NSF CCE-STEM-1835352, *Cultivating Ethical Norms in Citizen Science*). Our aim with the grant is for the field of practitioners to expand their understanding of trustworthy data to include ethical practices related to data acquisition and management. In citizen science, there are unique ethical issues with open data practices. We begin from the assumption that data quality and data ethics are equally important, as both center on actions related to rigorous field methodology by volunteers and appropriate practices by data stewardship, such as attribution, accessibility, confidentiality, and transparency.

Citizen science produces scientific data. Practitioners of citizen science therefore have the same data stewardship obligations as conventional scientists. In addition, however, management decisions about citizen science data may include consideration of a unique set of risks and benefits for volunteers. For example, anonymity in projects, datasets, or contributions is not always possible, and can run counter to desired interests of attribution. Data stewardship in citizen science has a broader scope than in conventional science, including reporting back to volunteers so that they can make meaning of the data, respecting how volunteers want sensitive data to be handled, recognizing contributions in a manner preferred by volunteers, and communicating clearly and transparently with volunteers about the above. We expand on these issues below.

OPEN DATA DECISIONS

Data governance can be responsive to concerns about protecting sensitive and personally identifiable information, treatment of indigenous knowledge, and intellectual property. Making data open is the act of making data available for others to freely use and re-use. However, the appropriate form that "open data" takes varies with the context of a given citizen science project. The majority of projects identified as citizen science have goals of advancing scientific research, and as such, practitioners should make data open to maximize the scientific value of the data. At the same time, we recognize that some projects have specific, action-oriented goals other than the general advancement of science, such as directly informing policy or social action. Given varied uses of scientific data and interests served, making data open is not always or automatically the most appropriate choice. We emphasize that ethical practices for establishing open data involve decisions about what should, and what should not, be shared, and what restrictions are warranted.

A misperception of "open data" is that posting data to the Web implies making it available for free use. However, the concept of "open data" is much more complex than the seemingly binary decision to make data "open" or "closed." Complexity stems from the numerous motivations for, approaches to, and justifications for making data open in the first place. When making and sharing content, copyright is a traditional mechanism to clarify restrictions on data use/re-use. However, according to US law, copyright applies to "creative works" and thus does not often apply to databases unless there is some creativity in their compilation (Miller et al., 2008; Kristof, 2016). However, there are alternative approaches to data stewardship besides copyright.

In 2010, the Panton Principles launched a guide for open data practices (Molloy, 2011). The Panton Principles recommended public domain licenses via the Open Data Commons Public Domain Dedication and License (PDDL–http://opendatacommons.org/licenses/pddl/1-0/) or Creative Commons Zero (CC0–http://creativecommons.org/publicdomain/zero/1.0/) which waive copyright. Such public domain licenses allow free, unrestricted use of the data for any purpose. While these might be a viable guide for datasets produced by conventional science, licensing in this way does not necessarily provide an open data solution for citizen science if volunteers want attribution. For example, the ODC PDDL and CC0 licenses do not require any attribution; however, one can use CC0 "with attribution appreciated." CC-BY allows free use of the data for any purposes with the requirement of attribution and allows attribution to extend to groups such as members of a citizen science project. Open data practices are further complicated when citizen science databases include photographs and/or open text, each a creative product with potential claim to copyright. Such licenses may not be entirely sufficient for these datasets. Groups that have historically experienced data inequities, exploitation by scientists, and/or intimidation by powerful interests may have heightened concerns about data access, data re-use, and the distribution of benefits. Thus, there

can be varied circumstances where volunteers want to restrict data use, rather than adopt free, unrestricted licensing options.

Nevertheless, persistent interest in open data for citizen science has led to nuanced applications of licensing options and exploration of unique challenges that public data archiving presents to the sustainability of long-term citizen science projects (Pearce-Higgins et al., 2018). In light of this complexity, it is essential to recognize that regardless of what approach one takes to make data open, and the benefits and challenges associated with it, the process of making data accessible for third parties to use requires active steps by a data steward (Miller et al., 2008). Next in this Perspective, we highlight ethical issues that citizen science data stewards and practitioners should consider when making data governance decisions, particularly with respect to open data.

Decision-Makers and Data-Producers

In citizen science, data stewards and data producers are often not the same people. When those who have labored on data collection are not in control of the data, ethical problems could arise from this basic structural feature. Power differentials arise because practitioners may have more education and institutional resources than project volunteers, and when practitioners are the sole data stewards, the power differentials are amplified. Thus, in these cases, data stewards (practitioners) need to properly consider the interests of the data producers (volunteers). For example, a data steward may view sharing geo-located data produced by volunteers as a way to maximize scientific goals, but data producers may view sharing as increasing their risks of physical, economic, or emotional harm. Insofar as datasets are monetizable, some communities may want to retain control over them for the benefit of those who have compiled the data or may be directly affected by it. Alternatively, volunteers may want to ensure that a dataset cannot be used for any commercial purposes (e.g., CC-NC restricts uses to non-commercial purposes).

Few studies have examined volunteer perspectives on the handling of citizen science data. Fox et al. (2019) found that volunteers in a large-scale UK project supported open access in principle but for its practice supported cautionary actions to protect sensitive information and restrict commercial reuse of data. Groom et al. (2017) reviewed the open access nature of biodiversity observation data contributed to GBIF (one of largest biodiversity data repositories). Contrary to what many people assumed, the datasets generated by citizen scientists were actually among the most restrictive in how they could be used. A further study examined the challenges and opportunities presented digital platforms that host citizen science data. In this case, Lynn et al. (2019) described the technology of the CitSci.org platform that allows project managers to choose different data governance options, some of which allow volunteers to make data governance choices themselves. We found no work yet addressing the challenges presented by the involvement of other third-party organizations (e.g., schools, museums) that manage volunteers in citizen science without involvement in making decisions about data stewardship.

Attribution and Acknowledgment

Attribution is the act of recognizing an individual's or group's contribution and appropriately acknowledging it. There are different forms of attribution, including non-monetary recognition such as authorship, acknowledgment, and citation. Accountability may also be associated with some forms of attribution, and involves an individual or group taking responsibility for some or all of the work. For example, in authorship, one is taking credit for the work and also taking responsibility for its quality and integrity.

In conventional and citizen science, publishing datasets is an old practice modeled after systems for publication of research results. For research papers, there are generally accepted standards for authorship when someone has made a substantial intellectual contribution to a project, or acknowledgment for contributions that are significant but not rising to that level (Brand et al., 2015; International Committee of Medical Journal, 2015). For citizen science papers, mirroring conventional approaches to authorship of papers is probably not meaningful, appropriate, or always possible for volunteers (Ward-Fear et al., 2020). For datasets, we are not aware of widely accepted standards for levels of contribution that warrant authorship or licensing attribution. Given the absence of norms, we encourage the data stewardship practice of licensing a dataset to foster intentional deliberation and decisions related to attribution.

Data Accessibility

Considerations of data accessibility should address the question, "accessible by whom?" Open data practices generally involve datasets being documented, discoverable, and allowing use by other scientists. In citizen science, however, data accessibility extends beyond engagement by scientists to practices that ensure that the data producers (volunteers) can make meaning of the datasets and use them for their own goals. With origins in environmental health, a standard practice of citizen science practitioners is the provision of "report-backs" to volunteers (Brody et al., 2007). Report-backs typically include personalized summaries of data (e.g., placing the individual contributor's data in context within the project) and/or excellent visualization of the collective data. Although report-backs are an important component of citizen science projects, they can raise privacy concerns if they disclose sensitive or private data to project participants or the public.

An additional consideration of data accessibility is the question, "accessible for what purpose?" Open data practices involve making datasets useable by other scientists for purposes similar to the original collection effort as well as re-use by other scientists for other, perhaps unanticipated, current or future purposes. In a citizen science context, when data producers are not data stewards, they have limited control of data re-use (Ganzevoort et al., 2017). In this light, it is important to note that currently, there is no open data license that can restrict data use in cases where it might harm data producers. Instead, case-by-case assessment to determine the potential for harm would require a closed license. Alternatively, an approach could be built around a framework of ethical principles guiding data use. For example, in considering indigenous data sovereignty, Carroll et al.

(2020) presented a framework that combined FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management and stewardship with the CARE (Collective benefit, Authority to control, Responsibility, Ethics) principles for Indigenous Data Governance. This kind of framework could help meet challenges of operationalizing "Open by default" (Stone and Calderon 2019) and give clarity on sensitive data and mechanisms to minimize harms and maximize benefits to data producers.

Data Confidentiality

Decisions about what data to share rely on considerations about the project's context and the types of other publicly available data. There are numerous instances in citizen science in which confidentiality of volunteer data should have primacy over open data sharing. This might include the collection of sensitive data based on location (e.g., volunteer location or protected species location), the collection of other sensitive data based on the subject of research (e.g., health), the unintentional collection of data from other people (e.g., photographs), or the possibility of combining data sets which could identify volunteers. For example, data collected by volunteers about corporate polluters may, if publicly released, identify and endanger those who have collected it (e.g., Wing, 2002). Additionally, in conjunction with existing data sets such as tax and real estate data or voter lists, new volunteer-collected data sets may enable re-identification of individuals or their locations via data triangulation (Golle, 2006). Even when researchers anonymize environmental health data by removing overt identifiers such as names and addresses, risks to re-identification of participants remain (Boronow et al., 2020).

Nissenbaum's privacy framework (2004), called Privacy 3.0, is helpful for navigating the various contexts and potential concerns that may arise through the data collection and management process more generally. Privacy 3.0 emphasizes the importance of (1) data minimization, (2) user control of personal information disclosure, and (3) contextual integrity (Nissenbaum, 2004, 2010, 2019). The concept of contextual integrity is particularly important; it focuses on understanding the flow of data from the sender to the recipient with attention to the subject matter, information type, and transmission principle (Nissenbaum, 2019). In a citizen science context, this might involve (a) not collecting personal data that should be confidential or (b) ensuring that if personal data must be collected that it remains confidential throughout the data lifecycle (i.e., ensuring that those portions of the dataset never go into open license or public domain). Further, Bowser and Wiggins (2015) have suggested the importance of viewing data privacy as involving a volunteer's right to manage access to their own voluntarily contributed personal data, which includes identified or identifiable information.

In certain types of projects, however, volunteers have no choice in the handling of their data or the protection of their privacy (Cooper et al., 2019). For example, in a sample of projects in which volunteers contributed data that unwittingly contained personally identifiable information, none involved volunteers in data governance decisions, and only half of the projects informed volunteers about data stewardship decisions, mostly related to privacy, liability, and copyright, typically through Terms of Service agreements (Cooper et al., 2019). Furthermore, even the professional scientists do not always play an active role in stewardship decisions of citizen science data, instead leaving decisions to the hosting platforms or institutional IT support (Bowser et al., 2020). Digital platforms that host citizen science projects, however, can enable joint decision making. For example, the platform CitSci.org supports preferences of both project managers and volunteers for customized levels of access to data (Lynn et al., 2019).

Transparency

The success of science, as well as citizen science, rests on the transparency of its technical and ethical practices. Transparency can be understood as the act of "making implicit and explicit values known or potentially discoverable by providing accessible information about research methods and data" (Elliott, 2017). There are two types of transparency that are especially important for discussing ethical data practices in citizen science. In the first instance, scientifically relevant transparency "refer[s] to efforts designed to assist scientists in achieving their goals, such as promoting new scientific discoveries and maintaining the reliability of scientific research" (Elliott and Resnik, 2019). Meanwhile, socially and ethically relevant transparency is more "focused on providing information that enables decision makers and members of the public to make effective use of scientific research" (Elliott and Resnik, 2019). These two understandings of transparency are not mutually exclusive of one another; they are two sides of the same coin. Both are important to consider when making decisions about how to collect and steward citizen science data in the most effective and ethical manner. In other words, transparency is an overarching obligation that applies to data accessibility, data confidentiality, and volunteer attribution and acknowledgment.

CONCLUSION AND RECOMMENDATIONS

"Thinking like a scientist" refers to higher order reasoning that distinguishes evidence from opinion and uses formal tools like statistics to minimize biases in human judgements (Kahneman, 2011). Scientific methods often include hypothesis testing that will ideally produce replicable conclusions. A scientific question can result in an agreed upon scientific answer. In contrast, "thinking like an ethicist" often means identifying ethical issues and using ethical frameworks to weigh a variety of options for addressing the issues. An ethical question can result in many ethical answers, each with equal validity. When there are competing values among those with valid interests in a dataset, there can be multiple ethical (and unethical) decisions about data governance (Ballantyne, 2018). Because of the pluralism of moral values, it may be impossible to offer, in the abstract, a set of ethical prescriptions that will be true for all citizen science research. Context matters, and what arises as an ethical issue and appropriate solution in one project might not in another almost-identical project.

Thus, ethical issues in citizen science have many solutions, most often including open data practices. When practitioners

opt for open data, they can do so effectively and responsibly by communicating intentions clearly with volunteers. For example, in considering the content of consent statements, Meyer (2018) recommended avoiding promises not to destroy data (which runs counter to expectations of some IRBs), not to share data, to restrict data analysis to the focal topic, and to obtain re-consent for additional sharing. Although Resnik et al. (2015) suggested all data sharing requests should go through the lead investigator of citizen science projects, Meyer (2018) recommended that practitioners can provide maximal access by working with a data repository that provides the desired governance options. Similarly, selection of the appropriate IT platform for the administration of the project should consider whether there are the desired data governance options (e.g., Lynn et al., 2019).

Open data is not a "one-size-fits-all" answer to the challenges of every project. A key variable to consider when deciding on data restrictions is the interests of the volunteer data producers, especially if they are not also part of the data stewardship team, with regard to accessibility, confidentiality, and attribution. Data stewards should listen to data producers, which may dictate more openness, or less, depending on a variety of circumstances. With transparency, practitioners can help data producers make highly informed decisions. Our dual hopes for citizen science are first, that a better understanding of the issues, risks, and stakes in decision making about open data in citizen science may help project leaders navigate these ethical decisions; and second, that by incorporating ethical rigor into data science practices from the outset, work in citizen science will be deeply informed by ethical practice.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CC lead the compilation, organization, and writing of the paper. LR and EJ helped with the compilation, organization, and writing of the paper. All authors contributed to the article and approved the submitted version.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Ballantyne, A. (2018). Where is the human in the data? a guide to ethical data use. *GigaScience* 7:giy076. doi: 10.1093/gigascience/giy076

Boronow, K. E., Perovich, L. J., Sweeney, L., Yoo, J. S., Rudel, R. A., Brown, P., et al. (2020). Privacy risks of sharing data from environmental health studies. *Environ. Health Perspect.* 128:017008. doi: 10.1289/EHP4817

Bowser, A., Cooper, C., de Sherbinin, A., Wiggins, A., Brenton, P., Chuang, T. R., et al. (2020). Still in need of norms: the state of the data in citizen science. *Citiz. Sci. Theory Pract.* 5:18. doi: 10.5334/cstp.303

Bowser, A., and Wiggins, A. (2015). Privacy in participatory research: advancing policy to support human computation. *Hum. Comput.* 2, 19–44. doi: 10.15346/hc.v2i1.3

Brand, A., Allen, L., Altman, M., Hlava,. M., and Scott, J. (2015). Beyond authorship: attribution, contribution, collaboration, and credit. *Learn. Publ.* 28, 151–155. doi: 10.1087/20150211

Brody, J. G., Morello-Frosch, R., Brown, P., Rudel, R. A., Altman, R. G., Frye, M., et al. (2007). Improving disclosure and consent: "Is it safe?": new ethics for reporting personal exposures to environmental chemicals. *Am. J. Public Health* 97, 1547–1554. doi: 10.2105/AJPH.2006.094813

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., et al. (2020). The CARE principles for indigenous data governance. *Data Sci. J.* 19, 1–12. doi: 10.5334/dsj-2020-043

Cooper, C., Shanley, L., Scassa, T., and Vayena, E. (2019). Project categories to guide institutional oversight of responsible. *Citiz. Sci. Theory Pract.* 4:7. doi: 10.5334/cstp.202

Elliott, K. C. (2017). *A Tapestry of Values: An Introduction to Values in Science.* Oxford: Oxford University Press.

Elliott, K. C., and Resnik, D. B. (2019). Making open science work for science and society. *Environ. Health Perspect.* 127, 1–6. doi: 10.1289/EHP4808

Fox, R., Bourn, N. A., Dennis, E. B., Heafield, R. T., Maclean, I. M., and Wilson, R. J. (2019). Opinions of citizen scientists on open access to UK butterfly and moth occurrence data. *Biodivers. Conserv.* 28, 3321–3341. doi: 10.1007/s10531-019-01824-6

Ganzevoort, W., van den Born, R. J. G., Halffman, W., et al. (2017). Sharing biodiversity data: citizen scientists' concerns and motivations. *Biodivers. Conserv.* 26, 2821–2837. doi: 10.1007/s10531-017-1391-z

Golle, P. (2006). "Revisiting the uniqueness of simple demographics in the US population," in *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society* (Alexandria, VA), 77–80.

Groom, Q., Weatherdon, L., and Geijzendorffer, I. R. (2017). Is citizen science an open science in the case of biodiversity observations? *J. Appl. Ecol.* 54, 612–617. doi: 10.1111/1365-2664.12767

International Committee of Medical Journal (2015). *ICMJE Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals.* Available online at: http://www.icmje.org/recommendations/ (accessed May 26, 2021).

Kahneman, D. (2011). *Thinking, Fast, and Slow.* New York, NY: Macmillan.

Kristof, C. (2016). Data and copyright. *Bull. Assoc. Inf. Sci. Technol.* 42, 20–22. doi: 10.1002/bul2.2016.1720420607

Lynn, S. J., Kaplan, N., Newman, S., Scarpino, R., and Newman, G. (2019). Designing a platform for ethical citizen science: a case study of CitSci.org. *Citiz. Sci. Theory Pract.* 4, 1–15. doi: 10.5334/cstp.227

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Adv. Methods Pract. Psychol. Sci.* 1, 131–144. doi: 10.1177/2515245917747656

Miller, P., Styles, R., and Heath, T. (2008). Open data commons, a license for open data. *LDOW* 369–374.

Molloy, J. C. (2011). The open knowledge foundation: open data means better science. *PLoS Biol.* 9:e1001195. doi: 10.1371/journal.pbio.1001195

Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. Law Rev.* 79, 119–158.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Palo Alto, CA: Stanford University Press.

Nissenbaum, H. (2019). Contextual integrity up and down the data food chain. *Theor. Inq. Law* 20, 221–256. doi: 10.1515/til-2019-0008

Pearce-Higgins, J. W., Baillie, S. R., Boughey, K., Bourn, N. A. D., Foppen, R. P. B., Gillings, S., et al. (2018). Overcoming the challenges of public data archiving for citizen science biodiversity recording and monitoring schemes. *J. Appl. Ecol.* 55, 2544–2551. doi: 10.1111/1365-2664.13180

Rasmussen, L. M. (2019). When citizens do science: stories from labs, garages, and beyond. *Narrat. Inq. Bioeth.* 9, 1–4. doi: 10.1353/nib.2019.0001

Rasmussen, L. M., and Cooper, C. (2019). Citizen science ethics. *Citiz. Sci. Theory Pract.* 4:5. doi: 10.5334/cstp.235

Resnik, D. B., Elliott, K. C., and Miller, A. K. (2015). A framework for addressing ethical issues in citizen science. *Environ. Sci. Policy* 54, 475–481. doi: 10.1016/j.envsci.2015.05.008

Ward-Fear, G., Gregory, P. B., Vendetti, J. E., and Shine, R. (2020). Authorship protocols must change to credit citizen scientists. *Trends Ecol. Evol.* 35, 187–190. doi: 10.1016/j.tree.2019.10.007

Wing, S. (2002). Social responsibility and research ethics in community-driven studies of industrialized hog production. *Environ. Health Perspect.* 110, 437–444. doi: 10.1289/ehp.02110437

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

The handling editor declared a past collaboration with one of the authors CC.