CITIZEN SCIENCE:
THEORY AND PRACTICE

Data Management Documentation in Citizen Science Projects: Bringing Formalisation and Transparency Together

RESEARCH PAPER

GEFION THUERMER
ESTEBAN GONZÁLEZ GUARDIA
NEAL REEVES
OSCAR CORCHO
ELENA SIMPERL

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Citizen science (CS) is a way to open up the scientific process, to make it more accessible and inclusive, and to bring professional scientists and the public together in shared endeavours to advance knowledge. Many initiatives engage citizens in the collection or curation of data, but do not state what happens with such data. Making data open is increasingly common and compulsory in professional science. To conduct transparent, open science with citizens, citizens need to be able to understand what happens with the data they contribute. Data management documentation (DMD) can increase understanding of and trust in citizen science data, improve data quality and accessibility, and increase the reproducibility of experiments. However, such documentation is often designed for specialists rather than amateurs.

This paper analyses the use of DMD in CS projects. We present analysis of a qualitative survey and assessment of projects' DMD, and four vignettes of data management practices. Since most projects in our sample did not have DMD, we further analyse their reasons for not doing so. We discuss the benefits and challenges of different forms of DMD, and barriers to having it, which include a lack of resources, a lack of awareness of tools to support DMD development, and the inaccessibility of existing tools to citizen scientists without formal scientific education. We conclude that, to maximise the inclusivity of citizen science, tools and templates need to be made more accessible for non-experts in data management.

CORRESPONDING AUTHOR:
Gefion Thuermer
King's College London, GB
gefion.thuermer@kcl.ac.uk

KEYWORDS:
citizen science; data management; data management documentation; data management plans; data quality

TO CITE THIS ARTICLE:
Thuermer, G, Guardia, EG, Reeves, N, Corcho, O and Simperl, E. 2023. Data Management Documentation in Citizen Science Projects: Bringing Formalisation and Transparency Together. *Citizen Science: Theory and Practice*, 8(1): 25, pp. 1–13. DOI: https://doi.org/10.5334/cstp.538

Citizen science projects can help advance scientific knowledge, and educate participants about specific topics and the scientific process in general (Bonney et al. 2009). These projects occur at different scales, from local, such as the iSPEX project (http://ispex-eu.org), where citizen scientists use sensors to measure air quality (Volten et al. 2018), to international, such as eBird (https://ebird.org), an online platform used globally to record bird observations (Lagoze 2014). Citizens may create such projects from the bottom up, with or without the support of professional scientists; conduct data collection or analysis in scientist-led projects (Wiggins and Crowston 2011); or contribute to scientific publications (Tinati et al. 2015).

The implementation of data management policies can make data and projects more scientifically sound, improve data quality and accessibility, and increase reproducibility. In CS projects, data management is an essential activity that enables citizen scientists to produce data that can be relevant and useful for, and trusted by, researchers (Hunter, Alabri and Ingen 2013). However, in many projects, data management policies or documentation are not systematically applied, leading to the perception that the data they produce is of lower quality (Ponti and Craglia 2020), and limiting its impact (Fraisl et al. 2020). A lack of documentation also poses risks for data reuse, especially when missing contributor details make it impossible to apply best practice on data citations (Hunter and Hsu 2015). More sharing of guidance and best practice on open science and usage of existing infrastructure is one route to help alleviate this situation (Schade et al. 2017).

Professional researchers implement data management policies through data management documentation (DMD), most commonly in the form of data management plans (DMPs). These describe the projects' data lifecycle, from collection, through analysis, archiving, and publication, and comprise details on data processing, quality assurance, and privacy. DMPs are required by many research institutions as part of ethical and legal project approval processes, and by research funders for access to grants. As more CS projects apply for such grants, these funder expectations, addressed at professional researchers, collide with the limited practice of many CS projects, which have neither the expertise nor the resources to create or implement such plans. Other documentation formats, such as datasheets (Gebru et al. 2021), may be easier to use for CS projects, but are even less common practice.

In this paper, we examine the current use of DMD in CS projects. We reached out to 240 projects to enquire about their use of DMPs, or reasons for lack of one. Based on the 56 responses we received, we found that 62% of the projects did not use DMPs, mainly due to timing, and because they lack the resources to write and implement such a plan. This is problematic, as a lack of documentation for their data management implies that the data may not be as accessible and reusable as it otherwise could or should be. It may also point to insufficiencies in the data management itself, if the lack of a plan implies lack of structured data management practices. At the same time, many projects have established, informal data practices.

We conclude that the resources to address the barriers to good, proactive data management in CS projects are already available, but not framed in a way that CS projects find intuitive. Resources to support project leaders and participants in the development of DMD should consider the constraints and needs of CS actors, and educate not only about the best way to develop data management policies, but also why this is beneficial to them. Existing tools should be adapted to reflect this, and become more useful for CS projects.

BACKGROUND

In this section, we outline the benefits, structure, and common usage of data management documentation in citizen science projects.

DATA MANAGEMENT DOCUMENTATION

Data documentation has many benefits, from boosting productivity and preventing confusion and errors in data processing to reducing risks (Azhar 2021). Among the first decisions in the development of DMD is whether to apply a formal or informal approach (Atici et al. 2013). Informal documentation can consist of notes or emails, while formal documentation is based on standards or templates. The two approaches are not mutually exclusive and can be part of the same documentation process. Common formats to document data include README files or tabs, with information about the data; Data Dictionaries describing the variables used; or Codebooks, containing the layout, structure, and codes used during data analysis (University of Illinois Library n.d.). We will focus on DMPs as the most common documentation form.

DMPs are not new, although their creation has become more relevant in the public policy domain (Smale et al. 2020). They were first implemented as technical documents in complex projects, often with restricted access. This definition has evolved due to the promotion of Open Science in academia. Today, DMPs describe the origin of and quality assurance applied to data, whether and where it is published or shared, and under which licensing terms. However, a DMP does not imply that data will be open. Figure 1 depicts a typical data flow captured by a DMP.

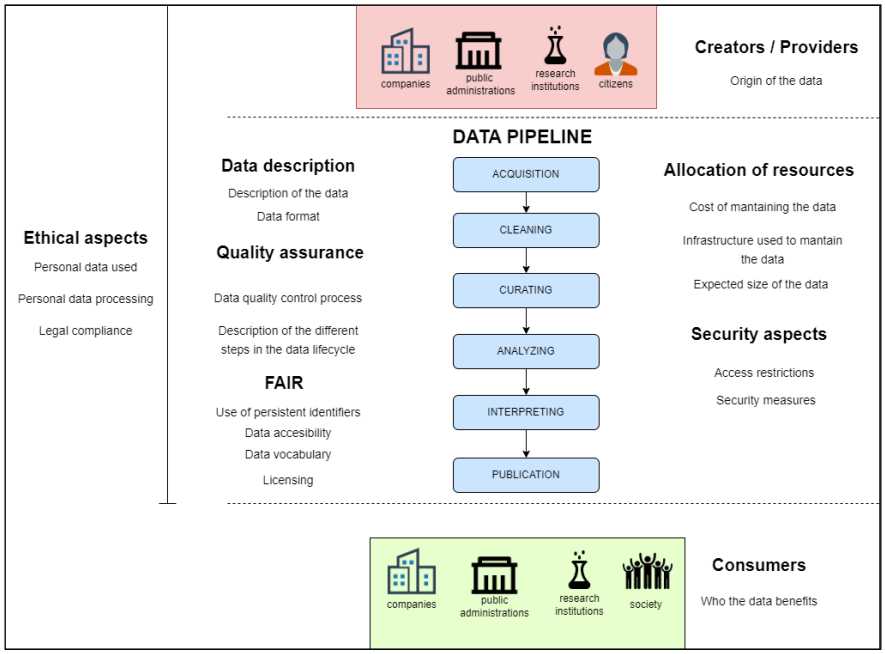

Figure 1 Description of a generic dataflow in a data management plan.

DMPs define the relation between the different stakeholders involved in data management and their respective responsibilities. This applies especially where personal data are concerned, as they define measures to ensure legal compliance. They outline how the research project host, staff, and citizen scientists work with the data, and how external researchers, policy-makers, or companies can interact with it. They are mandatory for research projects funded by most public funding agencies (e.g., the European Commission in the Horizon framework (EC n.d.), the Agency for Healthcare Research and Quality, the Department of Energy, or NASA; Adler 2015). Even where they are not required, they are recommended by funders like the Wellcome Trust (EAGDA 2017). Moreover, many research institutions require them as part of their data management strategy and risk assessments.

DMPs have three distinct benefits: They 1) increase efficiency for researchers by planning resources (Hudson-Vitale and Moulaison-Sandy 2019), 2) unlock economic value through sharing and reuse of data (Houghton and Gruen 2014), and 3) standardise institutional data practices and improve data quality (Burnette, Williams, and Imker 2016). They allow research funders to have a better overview and accountability of results beyond scientific publications and technical reports. They help researchers to actively consider how to make their data more reusable. Good data documentation can help both researchers and funders save resources by reproducing experiments (Koesten et al. 2020), a major problem in scientific research (Nature 2016).

The more data are made openly available, the more transparent and reproducible research becomes (Molloy 2011), whilst also enhancing benefits for wider society. This is particularly relevant in CS projects, which aim to make the scientific process more accessible to society at large (Geoghegan et al. 2016). Williams et al. (2018) identify the lack of reproducibility and reuse of data as one of the key factors limiting the impact of CS projects.

Making more thorough descriptions of data available, including a narrative of the data in context of the study in general, helps users make sense of data (Koesten et al. 2021). Providing context and metadata makes research more robust, ensures recognition (through citation), and increases the knowledge generated (Eynden and Bishop 2014). In disciplines where datasets are commonly shared, alternative documentation formats such as datasheets are used, which are also based on questionnaires about the data (Gebru et al. 2021).

Some publication venues have adopted models where researchers publish papers alongside data and/or code, for example in machine learning (https://paperswithcode.com), in Computer in Science & Engineering (CiSE) (https://www.computer.org/publications), or in IEEE (https://innovate.ieee.org/ieee-code-ocean/). Another key factor for data publication is interoperability, enabled through standards like metadata schemas. The Citizen Science Association in the United States has recently developed PPSR Core (https://core.citizenscience.org/), a metadata standard for public participation in scientific research projects. Consistent use of such schemas could alleviate insufficient documentation concerns.

DATA MANAGEMENT IN CITIZEN SCIENCE PROJECTS

CS projects can be led by professional researchers, organisations, or citizens themselves. Most citizens' contributions are related to data collection and/or processing, though they can occur at any stage of a CS project, including project development, analysis, or outreach (Thuermer et al. 2022).

Some CS projects use external platforms to manage their data, such as Epicollect (https://five.epicollect.net/), focused on data collection (Aanensen et al. 2009); or Zooniverse, focused on data processing (e.g., classification tasks; Simpson, Page, and De Roure. 2014); mailing list solutions like Google Groups; and ad-hoc solutions for various purposes. These projects may provide data for larger projects, such as bird observation projects that contribute data to the eBird platform, which manages data collected by users, publishes it in a repository, and exploits it with maps and visualisations. Researchers use eBird for its comprehensiveness and the high volume and quality of, as well as easy access to, the data (Sullivan et al. 2017). These options may entail costs for hosting and maintenance, which must be taken into account in projects' sustainability planning, especially in large CS projects performed over a long period of time (Locke et al. 2019).

Three of the Ten Principles of Citizen Science (ECSA 2015; emphasis ours) are related to data:

- **5**: Citizen scientists receive feedback from the project. For example, *how their data are being used* and what the research, policy or societal outcomes are.
- **7**: Citizen science *project data and meta-data are made publicly available* and where possible, results are published in an open access format.
- **10**: The leaders of citizen science projects take into consideration legal and ethical issues surrounding copyright, intellectual property, *data sharing agreements*, confidentiality, attribution, and the environmental impact of any activities.

Despite the large volume of data created by CS projects, data quality is a recurring issue, mainly due to lack of quality assurance mechanisms (Wiggins and He 2016). Ninety-four percent of CS projects implemented at least one data quality assurance mechanism, but these practices are poorly documented and publicised (de Sherbinin et al. 2021). The main barriers to publishing data in biodiversity projects are inconsistencies in the quality of data, and biases among scientists for citizens' profiles (age, education, etc.; Burgess et al. 2017). Discrepancies persist between the information reported by CS projects and their actual practices, with project leaders not always aware of how data are managed (Bowser et al. 2020). Stevenson et al. (2021) argue that today only specific types of citizen contributions—digital and physical object collection, digital classifications, and observations—are prone to data quality challenges, and that trust can be established through better data management.

DMD templates include questions that encourage project leaders to think about issues concerning their data, and implement data quality mechanisms. These could include additional steps to detect anomalies, validate observations by regional experts, apply analytical techniques (Hochachka et al. 2012), or specify requirements for participant training, which is one of the most effective strategies to improve data quality in CS projects (Freitag, Meyer, and Whiteman 2016).

Various tools already help scientists create DMPs (Gajbe et al. 2021). The Argos (https://argos.openaire.eu) and DMPOnline (https://dmponline.dcc.ac.uk) platforms especially can help project leaders and participants collaboratively manage their DMP over time, creating different versions as they develop. The Argos tool is integrated with others of the OpenAire ecosystem, especially the general-purpose repository Zenodo (https://zenodo.org). Its ease of use makes it ideal for CS project teams that lack skills or resources for data management.

PRIVACY POLICIES IN CITIZEN SCIENCE PROJECTS

A variety of policies and guidelines determine how a CS project interacts with its contributors (Bowser et al. 2014), including terms of use, legal and privacy policies. Terms of use establish the ownership of, access to, and use of data; legal policies outline the projects' adherence to relevant legislation in the projects' location; and privacy policies outline how they collect and process personal data (Costante et al. 2012).

On the one hand, privacy is an important part of DMD, and should be reflected in the design of applications that collect data (Sturm et al. 2017). Lack of consent and poorly documented policies, on the other hand, may lead to exploitation of participants, cause harm, or reduce the benefits for participating individuals or communities (Resnik, Elliot, and Miller 2015).

The introduction of the European Union's (EU's) General Data Protection Regulation (GDPR) has raised awareness for the necessity of both DMD and privacy policies (Kamocki, Mapelli, and Choukri, 2018). It defines personal data as "any information relating to an identified or identifiable natural person," and introduces the concept of the data controllers and processors. The data controller is responsible for compliance with the regulation, including overseeing activities of data processors who work with the

data. The regulation encourages ethical data processing and increased awareness among data subjects (those whose data are being processed).

CS projects are rarely designed with privacy in mind (Anhalt-Depies et al. 2019). Some potential threats for volunteers' privacy include their personal information (e.g., names listed as contributors; Bowser et al. 2017), triangulating of information (combined details from different sources revealing personal information), and geolocations, which are often submitted alongside observations, and can reveal home locations or activities (Tsai et al. 2010).

Guidance on how to manage personal data in CS is available (Bowser et al. 2014), and includes identifying which entries can be shared, anonymization techniques, informing contributors about risks and consequences, giving them the option to hide entries, and allowing them to modify or delete their data. Despite available support, data policies of many not-publicly-funded CS projects, which should inform users about threats and safeguards, are opaque or insufficiently documented (Bowser and Wiggins 2015).

In summary, there are many ways to implement and document data management policies, and if done well, they are very beneficial to make CS projects' data more accessible and reusable, and allow citizen scientists to better understand what data they contribute and what happens with it.

METHODOLOGY

The goal of the work described in this paper is to better understand the current use of data management policies and documentation in CS projects. We explore projects' practices with the aim to identify common issues, identify their causes, and provide recommendations to solve them.

We based our analysis on a database of CS projects developed within the ACTION project (https://actionproject.eu), created from two sources: SciStarter and the Wikipedia citizen science list. The combined list (Reeves 2021) was refined by removing duplicates, projects that generated no data, projects about which no further details could be found (e.g., no website), and projects that were not focused on pollution, since this was the focus of the initial analysis (see Roman et al. 2021). We supplemented the list with 15 CS projects from the H2020 SwafS group. The resulting dataset includes 330 projects. For each of these projects, we attempted to locate downloadable DMD on project websites, or else contact details or forms to reach out to them. Of the 330, 79 projects had neither DMD nor any contact details available on their website, and were therefore discarded. Five projects had a DMP available on their websites. To the remaining 235 projects, we reached out, via email or web forms, with these questions:

- Do you have a data management plan for your project?
- If so, would you be able to share a copy of it?
- If not, is there a specific reason for that?

We received responses from 52 projects. An overview of our outreach effort and results is provided in Table 1.

We collated all responses, conducted a qualitative content analysis, and inductively categorised all projects based on whether or not they had a DMP (21 did, 34 did not); why they did not have a DMP, if a reason was given; and where this was provided, alternative DMD used.

For the 17 DMPs we received, we conducted a thematic analysis (one document submitted as a DMP was excluded, as it was merely a documentation of data flows). All documents were coded both deductively, based on the criteria and common elements of DMPs listed in the background section, such as data acquisition, analysis and preservation; and inductively, to capture any elements that were not commonly expected. This led to a classification of 16 different topics, including data processing steps, licensing, participant engagement and consent, and (potential) impact of the data. An overview of the classification used and their frequency is provided in Supplemental File 1: Analysis Codebook.

To provide further context for the different practices, we prepared four vignettes or projects and their respective approaches to DMD. One project has a complete DMP, while the remaining three have varying perspectives on the need to formalise their data management, with one currently formulating a policy, one lacking the resources to do so, and one deeming it unnecessary. For each project, we removed identifying data, and summarised their responses, the DMD on their websites, and the prominence such materials were given. These are provided in Supplemental File 2: Vignettes.

Downloaded	5
Responded	52
Contacted	
Emailed	157
Web form	25
Bounced	11
Refused	1
No contact details	79

Table 1 Overview of outreach to projects.

RESULTS

In this section, we outline the results of the analysis of the responses to our queries of the projects, and the content analysis of the data management plans we received.

RESPONSES

Of the 21 projects that stated that they had a DMP, 13 were funded by the European Commission (EC), 5 by research institutions such as universities, 2 by the US government, and one by industry. Three of the projects, including the industry-funded one, have stated that their DMP is a confidential document and therefore cannot be shared. An overview of projects' DMP status is provided in Figure 2.

Only one of the responses—from a platform that hosts multiple projects—stated that they consider DMPs as good practice, and require all projects on the platform to have one that is understandable for citizen scientists.

The overwhelming majority of projects that have a DMP are institutionally obliged to it. The EC Horizon framework and US National Science Foundation require project DMPs at different stages: within the first six months of funded projects, with regular updates thereafter (EC n.d.), and as part of applications (NSF n.d.), respectively. The EC requires details about the data, its adherence to FAIR principles, and related resources, security, and ethical aspects; the US requires details on how data will be shared and made available to other researchers. While institutional pressure may not be the only driver for DMPs in projects, it is an efficient one: All H2020 projects that responded to our query had one.

An overview for why projects did not have a DMP is provided in Figure 3. The primary reason was that a DMP is still being developed. This was due to the newness of the projects, which were still developing in general, or were undergoing changes in data management that were yet to be documented. One of the projects explained that their

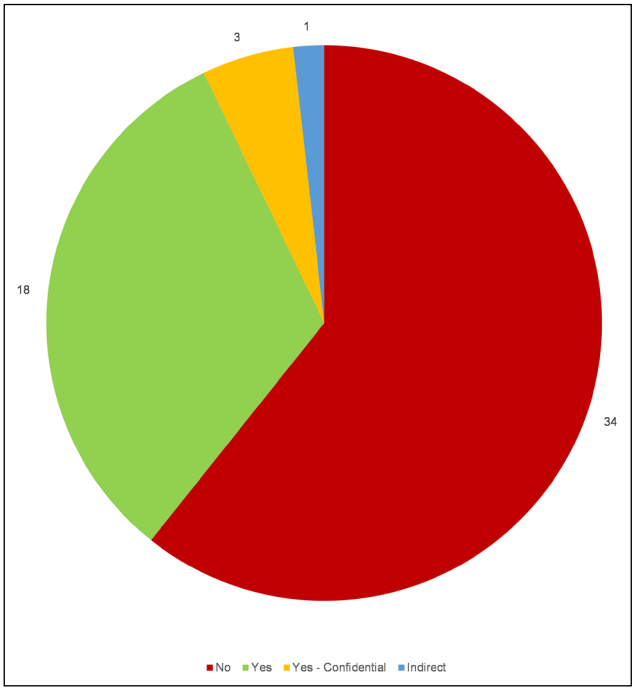

Figure 2 Responses to "Do you have a data management plan?" (N = 56).

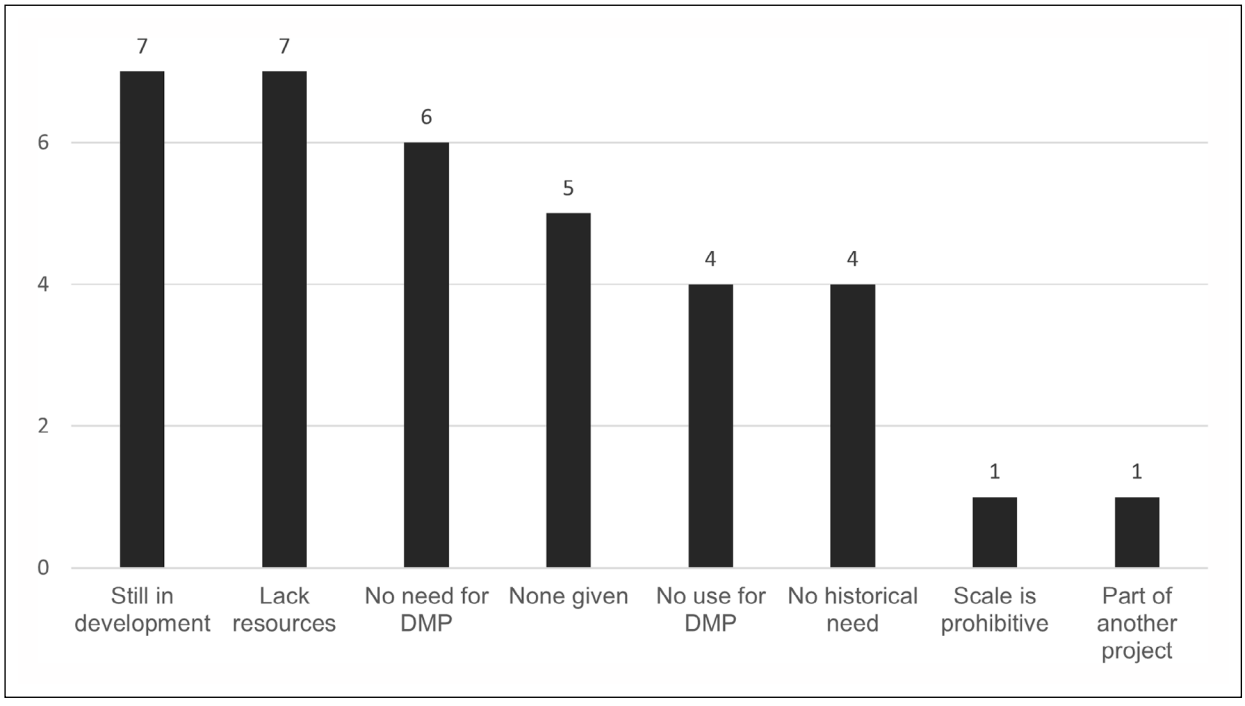

Figure 3 Categorised responses to "Why do you not have a data management plan?" (N = 35).

data management is undefined because the data itself is continuously changing, making it incompatible with their institutions' requirement to deposit a "final" dataset.

Another seven projects stated that they had no DMP because they lacked the resources to develop and/or implement one. While some of these projects had informal procedures, they were either small-scale with limited capacity among volunteers to take care of data management, or were hosted by companies who have not set aside resources for such tasks. Based on the anecdotal feedback we received, the primary issue here appeared to be that for small CS projects, which were run by volunteers, documenting data management is not a priority. While they had data management practice and a clear commitment to open data, formalising procedures was considered an unnecessary overhead. Projects may not only lack the resources and expertise to write DMD, but also have no clear understanding of why it would be relevant for them to have at all.

Six projects stated that they have no DMP because they have no need for one; four of these explained that their data were collected and stored on external platforms such as eBird, and thus they did not hold any of their own data that required management. These projects' primary focus was on data collection, to ensure the data were available to their members, to researchers, or to the general public. They did not conduct analysis or otherwise use the data.

The projects with no need for a DMP are distinct from another four who stated that they did not have a need for a DMP when they were created, on dates that spanned 2008 to 2018. Formal DMPs were not a requirement or common practice when the projects started, and they have either already concluded, or the data management is already established, with no perceived need to document it now. These projects did have informal procedures, or even public data, but no accessible documentation of its management.

Another four projects stated that they have no use for a DMP, as it was not necessary for them to document their data management processes. These projects were used for specific, closed purposes such as teaching, dealt only with their customers' data, or didn't perceive their activities as citizen science at all. Instead of a DMP, they offered privacy policies or user settings for their participants to be informed about and enact control over their data contributions. An overview of alternative DMD used by projects is provided in Figure 4.

DATA MANAGEMENT PLAN CONTENT

There are clear differences in the content of DMD over time: Plans developed at the beginning of projects were often preliminary, indicating what the projects intended to do, whereas updated and final plans from later stages include detailed overviews of collated datasets. Some plans included a complete overview of all datasets that were to be collected and how they would be used. However, not all projects updated their documents over time. We find this reflected in the vignette for Project B, which developed the DMD while already collecting data.

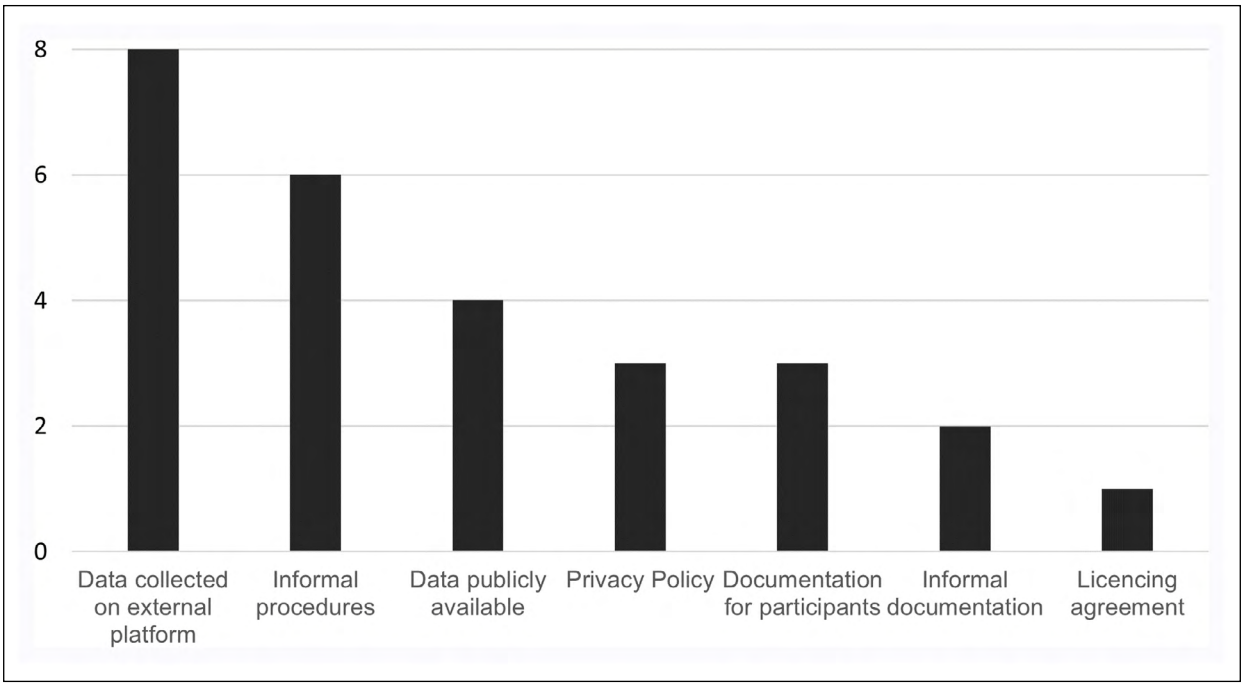

Figure 4 Alternative forms of data management documentation (N = 25, multiple mentions possible).

All DMPs included details on data acquisition, storage, and metadata. DMPs by projects funded through the EC focused on personal data, consent, costs, and licensing, likely due to the FAIR principles being part of the common template. However, the structure of the overall document differed a lot, showing how projects' thinking about these criteria differs, between being project-wholistic, and dataset-specific.

DMPs from projects funded through research institutions, while discussing personal data, included details on ethics and consent only occasionally. US government–funded projects (of which there were only two in the sample) did not include such details at all. They did, however, focus more on the outcomes of the research and the quality of the data they created. Less than half of the documents did not discuss data analysis at all, or only very vaguely. While it was often alluded to ("Results will be published in research papers"), the data analysis that was or would be applied to the data was barely discussed. An overview of the different content of DMPs by funders is provided in Figure 5.

The primary template used by projects funded by the EC is from DMPOnline, which the EC recommends. In comparison, plans developed by US institutions were much shorter and appeared less formalised, containing general summaries and responses to a small number of broad questions, as opposed to a long list of questions and subquestions, tables, and legal language in European plans. DMPs of US-based projects appeared to be written with a goal to be understandable by citizen scientists, while European plans tended to be written for institutions and experts. DMPs by Horizon projects were clearly designed for and used as institutional or bureaucratic tools, and were written for an audience of professional researchers. These templates and plans may therefore be much harder to grasp for CS projects that are not based in institutions, led by researchers, or supported by funders who prioritise data management, but still need to manage their data, and may want to make it available to others.

While many DMPs included details on privacy, pseudo- or anonymisation, data access etc., some did so at a level of detail that would be much more suitable to a privacy policy—especially when they discussed operational as opposed to research data. Where plans made a clear distinction between operational and research data, the applicability of FAIR principles and legal obligations was much more obvious: Operational data needs to be stored and processed but not analysed or published, hence GDPR applies but FAIR principles do not, while research data requires consideration of pseudo- or anonymisation and publication, including FAIR. These differences were mirrored in our vignettes for Projects A–C, where all projects' privacy policies were either not publicly accessible, or not suitable for the data collected.

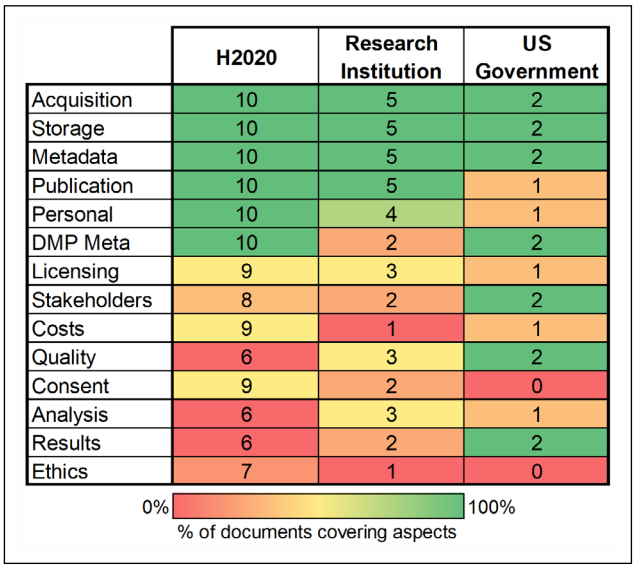

Figure 5 Percentage of codes applied in data management plan analysis, by funder (N = 17).

DISCUSSION

DMD can help CS project teams to implement professional data management practices, allow citizen scientists to understand, and the project host to monitor and track the research data lifecycle. In this section, we outline our insight from the analysis of projects' responses and DMPs.

SELECTING APPROPRIATE DATA MANAGEMENT DOCUMENTATION

Several project respondents stated they did not need a DMP because they used third-party platforms to collect and store data. While using existing platforms is a valid strategy to ensure data are accessible to those who are interested in it, especially for projects with limited resources, this practice may limit project teams' own agency in what happens with the data they contributed. This can obscure who is responsible for data management, because projects using these platforms are not the controllers of the data that is processed or collected through them. Documented roles and responsibilities can help to clarify boundaries between the project and the platforms that they use. This documentation should not only cover the management of research data, but also the management of any other data the project processes, such as potentially personal data of contributors. Even if these are not processed on platforms, or analysed, they still fall within the responsibility of project hosts to conduct ethical research and data management.

In terms of the DMD projects highlighted, some projects used privacy policies instead of DMPs. The two documentation

forms may indeed overlap where they touch on the processing and use of data, and ethical considerations of this processing. However, they serve different purposes: The privacy policy focuses on personal data, and explains the data processing to the people whose data are used. The purview of DMD is wider, no matter if the data are personal or not. Where DMD describes the lifecycle of data within a project, including its storage, access, and processing, a privacy policy addresses the participants, outlining the use of their data specifically. Where DMD is an internal management tool, privacy policies are outward facing agreements with contributors. Since DMD also covers personal data processing, including ethical aspects such as informed consent, DMD and privacy policies must be closely aligned. The synonymous use of both types of documentation among the projects indicates that they focus on the individual citizen scientist as a data contributor, instead of the aggregated data collected from all citizen scientists across the project. General guidance on the distinctions between DMD and privacy policies, and more adaptable templates that consider the differing requirements for personal and non-personal data, would go a long way to resolve such misconceptions.

CREATING DATA MANAGEMENT DOCUMENTATION

Many project respondents stated that they do not have a DMP because the teams were still in the process of developing it, or did not have the required infrastructure or resources to write and implement such a plan. Ideally, data management should be considered at the very beginning of a project, so that resources can be planned accordingly. While some of the projects that stated they were still developing their data management strategy were actually at the very beginning, others had already begun to collect data. We consider this a risk, as it implies that neither participants nor project coordinators have a full picture of the data flows at the time data collection starts, which means that participants cannot contribute data with fully informed consent.

In the initial development phase for DMD, there are resources available to support project teams, including tools to create DMPs. Data management documentation—DMP or otherwise—should not only be developed as a tick-box exercise, but treated as living documents, which cycle through different stages over the project lifetime. Tools like Argos can support CS project teams in the development of DMD. However, their use may still appear as an administrative burden, rather than a useful and necessary tool for good data management practice. One reason for this may be that these templates are very technically oriented (e.g., FAIR principles), and often include ethical and legal considerations that may not be at the forefront of citizen scientists' minds, or even be what CS project team or citizen scientists themselves are interested in. Other documentation formats, such as datasheets (Gebru et al. 2021), with a more specific focus on the data itself, and guidance that does not require expert language, may be easier to use and understand for citizen scientists. CS project teams will need accessible templates and guidance that allows them to use these formats.

Existing tools to support DMD development are mainly developed to support larger, funded projects, and structured in line with institutional templates. While the use of such tools will be quite obvious to large project teams who are required to develop DMD, and have a variety of skills and resources to do so, the tools may be more formal and extensive than small project teams find intuitively useful. However, they do make it easier for project teams to think about their data management in a structured way and to consider what data they have and what they do with it regardless of scale. Such DMD development tools should be adapted to make them more accessible for lay people with no formal training in scientific methods.

MATERIALISING BENEFITS OF DATA MANAGEMENT DOCUMENTATION

More accessible tools could also use the DMD development process to educate about the benefits of data policies. That way, smaller, un-funded project teams might use them to develop their planning despite limited resources. Developers of such tools should consider the benefits CS projects can gain from good data management, and frame the tools in a way that not only allows project teams to generate DMD, but simultaneously informs users about best practice in data management and how they benefit from it. This might be positioned together with mandatory data processing requirements, such as GDPR. While project teams will still be less familiar with these regulations, they will be aware of—and need guidance to comply with—their legal duties. One way to achieve this would be for CS platforms, such as Zooniverse or Epicollect, to support projects they host with data documentation templates. The Ecosystem Investigation Network (https://investigate.gmri.org/) have already done this, with a short and understandable documentation of datasets and processes available for each project, covering data quality, analysis, publication, and management.

We further found that many projects make data available online in order to make it reusable, following one of the ten CS principles (ECSA 2015). While this is encouraging, if data are meant to be useful to others, they should not only be made public, but be published following the FAIR principles, including a persistent digital object identifier (DOI), metadata based on common standards, a narrative of its generation (Koesten et al., 2021), and an appropriate licence (such as creative commons). General purpose repositories, such as Zenodo, freely create DOIs for files deposited on their platforms, and thus are good options for CS projects. Publishing CS project data in this way would

enable researchers not only to understand the data better and trust their quality, but also enable researchers to give appropriate credit to the CS project as the data source. Templates could help project teams to structure their own data management, and increase the reusability of the data.

We found from our analysis of the DMPs that the different plans serve different purposes: Some were written to meet institutional requirements, including assessments of legal risks inherent to the data; while others were meant to be read by lay people or citizen scientists themselves, laying out concisely and in non-expert language what data are collected and what happens with it. These two practices are not necessarily mutually exclusive: Projects can have extensive documentation of their data processes, adherence to FAIR principles, etc., and still provide understandable documentation for their participants. This also means that, when templates are developed, the different purposes and target groups need to be considered, both for the development and consumption of DMD.

CONCLUSION

In summary, we found that 62% of surveyed CS projects did not use a DMP, mainly due to lack of time or resources. However, many of these projects still use informal data practices. We conclude that

1. resources for data management are available, but not framed in a way that is useful for CS projects;
2. authors of such resources should consider the constraints and needs of CS projects and adapt them accordingly; and
3. such resources should educate not only about the best way to develop DMD, but also why they are beneficial.

In future research, we will investigate how DMD in participatory projects can be used to ensure that benefits of data are returned to contributors, and processes are sufficiently transparent for these contributors to make informed decisions and provide meaningful consent to the proposed uses of their data. We will also investigate who within CS projects is in charge of managing data, and to which degree citizen scientists are involved in these processes.

If one of the goals of citizen science is to democratise the scientific process and make it more accessible to citizens, then citizens need to be able to understand what happens with the data they contribute to these projects. Managing data is part of citizen science, because it is part of the scientific process, and citizen scientists need opportunities and motivations to learn about data management, if they are to contribute to the scientific process. DMD can help to make this data more accessible, understandable, and usable, and improve its quality. It is a way to systematise knowledge about project data, and allows CS to come even closer to professional research. DMD can mediate between scientists and civil society: Written by scientists, it can help citizens understand what data they contribute and what happens to it; written by citizen scientists, it would enable professional scientists to do the same.

DATA ACCESSIBILITY STATEMENT

The case selection in this paper is based on a published collection: https://doi.org/10.5281/zenodo.5101358.

The DMPs and responses cannot be published owing to lack of consent from respondents.

SUPPLEMENTARY FILES

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

- **Supplemental File 1.** Analysis Codebook. DOI: https://doi.org/10.5334/cstp.538.s1
- **Supplemental File 2.** Vignettes. DOI: https://doi.org/10.5334/cstp.538.s2

ETHICS AND CONSENT

The project is registered with the Ethics Committee at King's College London under Project ID 33468.

ACKNOWLEDGEMENTS

We are grateful to all citizen science projects that responded to our questions about their data management practices and documentation.

FUNDING INFORMATION

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 824603 and 101058677.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

GT and EGG conceptualised the paper. GT collected and analysed the data, with contributions from EGG and NR. GT and EGG wrote the paper, with contributions from NR. ES and OC made substantial comments to improve the paper before submission, and secured the funding for the study with contributions from GT and NR.

AUTHOR AFFILIATIONS

Gefion Thuermer orcid.org/0000-0001-7345-0000
King's College London, GB
Esteban González Guardia orcid.org/0000-0003-4112-6825
Universidad Politécnica de Madrid, ES
Neal Reeves orcid.org/0000-0002-1044-3943
King's College London, GB
Oscar Corcho orcid.org/0000-0002-9260-0753
Universidad Politécnica de Madrid, ES
Elena Simperl orcid.org/0000-0003-1722-947X
King's College London, GB

REFERENCES

Aanensen, DM, Huntley, DM, Feil, EJ, al-Own, F and **Spratt, BG.** 2009. EpiCollect: Linking smartphones to web applications for epidemiology, ecology and community data collection. *PLOS ONE*, 4(9): e6968. DOI: https://doi.org/10.1371/journal.pone.0006968

Adler, P. 2015. AHRQ, NASA, USDA release plans for public access to funded research. *Association of Research Libraries*. Available at https://www.arl.org/news/ahrq-nasa-usda-release-plans-for-public-access-to-funded-research/ (Last accessed 19 July 2021).

Anhalt-Depies, C, Stenglein, JL, Zuckerberg, B, Townsend, PA and **Rissman, AR.** 2019. Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. *Biological Conservation*, 238: 108195. DOI: https://doi.org/10.1016/j.biocon.2019.108195

Atici, L, Kansa, SW, Lev-Tov, J and **Kansa, EC.** 2013. Other people's data: a demonstration of the imperative of publishing primary data. *Journal of Archaeological Method and Theory*, 20(4): 663–681. DOI: https://doi.org/10.1007/s10816-012-9132-9

Azhar, A. 2021. 15 reasons why documentation is important? *Curious Desire*. Available at https://curiousdesire.com/why-documentation-is-important/ (Last accessed 23 May 2023).

Bonney, R, Cooper, CB, Dickinson, J, Kelling, S, Phillips, T, Rosenberg, KV and **Shirk, J.** 2009. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11): 977–984. DOI: https://doi.org/10.1525/bio.2009.59.11.9

Bowser, A, Cooper, C, Sherbinin, A. de, Wiggins, A, Brenton, P, Chuang, T-R, Faustman, E, Haklay, M and **Meloche, M.** 2020. Still in need of norms: the state of the data in citizen science. *Citizen Science: Theory and Practice*, 5(1): 18. DOI: https://doi.org/10.5334/cstp.303

Bowser, A, Shilton, K, Preece, J and **Warrick, E.** 2017. Accounting for privacy in citizen science: ethical research in a context of openness. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland, Oregon, USA on 25 February 2017, pp. 2124–2136. DOI: https://doi.org/10.1145/2998181.2998305

Bowser, A and **Wiggins, A.** 2015. Privacy in participatory research: advancing policy to support human computation. *Human Computation*, 2. DOI: https://doi.org/10.15346/hc.v2i1.3

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

Burgess, HK, DeBey, LB, Froehlich, HE, Schmidt, N, Theobald, EJ, Ettinger, AK, HilleRisLambers, J, Tewksbury, J and **Parrish, JK.** 2017. The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation*, 208: 113–120. DOI: https://doi.org/10.1016/j.biocon.2016.05.014

Burnette, M, Williams, S and **Imker, H.** 2016. From plan to action: successful data management plan implementation in a multidisciplinary project. *Journal of eScience Librarianship*, 5: e1101. DOI: https://doi.org/10.7191/jeslib.2016.1101

Costante, E, Sun, Y, Petković, M and **den Hartog, J.** 2012. A machine learning solution to assess privacy policy completeness. In: *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*. WPES '12. New York, NY, USA on 15 October 2012, pp. 91–96. DOI: https://doi.org/10.1145/2381966.2381979

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 and **Sivakumar, K.** 2021. The Critical Importance of Citizen Science Data, *Frontiers in Climate*, 3. DOI: https://doi.org/10.3389/fclim.2021.650760

ECSA. 2015. *Ten principles of citizen science*. Available at https://ecsa.citizen-science.net/wp-content/uploads/2020/02/ecsa_ten_principles_of_citizen_science.pdf (Last accessed 23 May 2023)

European Commission. n.d. *Data management*. H2020 Online Manual. Available at https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm (Last accessed 23 May 2023).

Expert Advisory Group on Data Access (EAGDA). 2017. *Data management plans: recommendations*. Available at https://cms.wellcome.org/sites/default/files/data-management-plans.pdf (Last accessed 23 May 2023).

Eynden, V and **Bishop, L.** 2014. *Sowing the Seed: Incentives and motivations for sharing research data, a researchers' perspective*. Available at: https://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf (Last accessed 23 May 2023).

Fraisl, D, Campbell, J, See, L, Wehn, U, Wardlaw, J, Gold, M, Moorthy, I, Arias, R, Piera, J, Oliver, JL, Masó, J, Penker, M and **Fritz, S.** 2020. Mapping citizen science contributions to the UN sustainable development goals. *Sustainability Science*. DOI: https://doi.org/10.1007/s11625-020-00833-7

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

Gajbe, SB, Tiwari, A, Gopalj and **Singh, RK.** 2021. Evaluation and analysis of data management plan tools: a parametric approach. *Information Processing & Management*, 58(3): 102480. DOI: https://doi.org/10.1016/j.ipm.2020.102480

Gebru, T, Morgenstern, J, Vecchione, B, Vaughan, JW, Wallach, H, Daumé III, H and **Crawford, K.** 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92. DOI: https://doi.org/10.1145/3458723

Geoghegan, H, Dyke, A, Pateman, RM and **West, SE.** 2016. *Understanding motivations for citizen science*. UKEOF. Available at https://www.ukeof.org.uk/resources/citizen-science-resources/MotivationsforCSREPORTFINALMay2016.pdf (Last accessed 23 May 2023).

Hochachka, WM, Fink, D, Hutchinson, RA, Sheldon, D, Wong, W-K and **Kelling, S.** 2012. Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, 27(2): 130–137. DOI: https://doi.org/10.1016/j.tree.2011.11.006

Houghton, J and **Gruen, N.** 2014. *Open research data. Report to the Australian National Data Service (ANDS)*. Available at https://apo.org.au/node/53613 (Last accessed 23 May 2023).

Hudson-Vitale, C and **Moulaison-Sandy, H.** 2019. Data management plans: a review. *DESIDOC Journal of Library and Information Technology*, 39(6): 322–328. DOI: https://doi.org/10.14429/djlit.39.06.15086

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

Hunter, J and **Hsu, C-H.** 2015. Formal acknowledgement of citizen scientists' contributions via dynamic data citations. In: Allen, RB, Hunter, J and Zeng, ML (eds.), *Digital Libraries: Providing Quality Information*. Lecture Notes in Computer Science. Cham: Springer International Publishing. pp. 64–75. DOI: https://doi.org/10.1007/978-3-319-27974-9_7

Kamocki, P, Mapelli, V and **Choukri, K.** 2018. Data Management Plan (DMP) for language data under the new General Data Protection Regulation (GDPR). In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan in May 2018.

Koesten, L, Gregory, K, Groth, P and **Simperl, E.** 2021. Talking datasets – Understanding data sensemaking behaviours. *International Journal of Human-Computer Studies*, 146: 102562. DOI: https://doi.org/10.1016/j.ijhcs.2020.102562

Koesten, L, Vougioukli, P, Simperl, E and **Groth, P.** 2020. Dataset Reuse: Toward Translating Principles to Practice. *Patterns*, 22. DOI: https://doi.org/10.1016/j.patter.2020.100136

Lagoze, C. 2014. eBird: curating citizen science data for use by diverse communities. *International Journal of Digital Curation*, 9(1): 71–82. DOI: https://doi.org/10.2218/ijdc.v9i1.302

Locke, CM, Anhalt-Depies, CM, Frett, S, Stenglein, JL, Cameron, S, Malleshappa, V, Peltier, T, Zuckerberg, B and **Townsend, PA.** 2019. Managing a large citizen science project to monitor wildlife. *Wildlife Society Bulletin*, 43(1): 4–10. DOI: https://doi.org/10.1002/wsb.943

Molloy, JC. 2011. The Open Knowledge Foundation: Open data means better science. *PLOS Biology*, 9(12): e1001195. DOI: https://doi.org/10.1371/journal.pbio.1001195

National Science Foundation (NSF). n.d. *Preparing Your Data Management Plan*. Available at https://new.nsf.gov/funding/data-management-plan (Last accessed 23 May 2023).

Nature. 2016. Reality check on reproducibility. *Nature*, 533(7604): 437–437. DOI: https://doi.org/10.1038/533437a

Ponti, M and **Craglia, M.** 2020. *Citizen-generated data for public policy. A brief review of European citizen-generated data projects*. Available at https://ec.europa.eu/jrc/communities/sites/jrccties/files/jrc120231_citizen-generated_data_for_public_policy.pdf (Last accessed 9 Feb 2021).

Reeves, N and **the ACTION Consortium.** 2021. *Database of Citizen Science Projects*. DOI: https://doi.org/10.5281/zenodo.5101358.

Resnik, DB, Elliott, KC and **Miller, AK.** 2015. A framework for addressing ethical issues in citizen science. *Environmental Science & Policy*, 54: 475–481. DOI: https://doi.org/10.1016/j.envsci.2015.05.008

Roman, D, Reeves, N, Gonzalez, E, Celino, I, Abd El Kader, S, Turk, P, Soylu, A, Corcho, O, Cedazo, R, Re Calegari, G, Scandolari, D and **Simperl, E.** 2021. An analysis of pollution Citizen Science projects from the perspective of Data Science and Open Science. *Data Technologies and Applications*, 55(5): 622–642. DOI: https://doi.org/10.1108/DTA-10-2020-0253

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

Simpson, R, Page, KR and **De Roure, D.** 2014. Zooniverse: observing the world's largest citizen science platform. In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. New York, NY, USA on 7 April 2014, pp. 1049–1054. DOI: https://doi.org/10.1145/2567948.2579215

Smale, NA, Unsworth, K, Denyer, G, Magatova, E and **Barr, D.** 2020. A review of the history, advocacy and efficacy of data management plans. *International Journal of Digital Curation*, 15(1): 30. DOI: https://doi.org/10.2218/ijdc.v15i1.525

Stevenson, RD, Suomela, T, Kim, H and **He, Y.** 2021. Seven primary data types in citizen science determine data quality requirements and methods. *Frontiers in Climate*, 3. DOI: https://doi.org/10.3389/fclim.2021.645120

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

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

Thuermer, G, Reeves, N, Baroni, I, Scandolari, D, Scrocca, M, van Grunsven, R, Maddalena, E, Simperl, E, Austen, K, Hoelker, F, Schroer, S, Grossberndt, S, Roman, D, Passani, A, Firus, K, Gonzalez Fuentetaja, R, González Guardia, E and **Corcho, O.** 2022. *Participatory Science Toolkit Against Pollution*. DOI: https://doi.org/10.5281/zenodo.6491235.

Tinati, R, Van Kleek, M, Simperl, E, Luczak-Rösch, M, Simpson, R and **Shadbolt, N.** 2015. Designing for citizen data analysis: a cross-sectional case study of a multi-domain citizen science platform. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Seoul, Republic of Korea on 18 April 2015. pp. 4069–4078. DOI: https://doi.org/10.1145/2702123.2702420

Tsai, J, Kelley, PG, Cranor, LF and **Sadeh, N.** 2010. Location-sharing technologies: privacy risks and controls. *ISJLP*, 6: 119.

University of Illinois Library. n.d. *Introduction to data management for undergraduate students: data documentation*. Available at https://guides.library.illinois.edu/introdata/documentation (Last accessed 23 May 2023).

Volten, H, Devilee, J, Apituley, A, Carton, L, Grothe, M, Keller, C, Kresin, F, Land-Zandstra, A, Noordijk, E, van Putten, E, Rietjens, J, Snik, F, Tielemans, E, Vonk, J, Voogt, M and **Wesseling, J.** 2018. Enhancing national environmental monitoring through local citizen science. In: Hecker, S, Haklay, M, Bowser, A, Makuch, Z, Vogel, J and Bonn, A (eds.). *Citizen science*. Innovation in open science, society and policy. UCL Press. pp. 337–352. DOI: https://doi.org/10.2307/j.ctv550cf2.30

Wiggins, A and **Crowston, K.** 2011. From conservation to crowdsourcing: a typology of citizen science. In: *2011 44th Hawaii International Conference on System Sciences*. January 2011 pp. 1–10. DOI: https://doi.org/10.1109/HICSS.2011.207

Wiggins, A and **He, Y.** 2016. Community-based data validation practices in citizen science. In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. New York, NY, USA on 27 February 2016, pp. 1548–1559. DOI: https://doi.org/10.1145/2818048.2820063

Williams, J, Chapman, C, Leibovici, DG, Loïs, G, Matheus, A, Oggioni, A, Schade, S, See, L and **van Genuchten, PPL.** 2018. Maximising the impact and reuse of citizen science data. In: Hecker, S, Haklay, M, Bowser, A, Makuch, Z, Vogel, J and Bonn, A (eds.). *Citizen science*. Innovation in open science, society and policy. UCL Press. pp. 321–336. DOI: https://doi.org/10.2307/j.ctv550cf2.29

TO CITE THIS ARTICLE:
Thuermer, G, Guardia, EG, Reeves, N, Corcho, O and Simperl, E. 2023. Data Management Documentation in Citizen Science Projects: Bringing Formalisation and Transparency Together. *Citizen Science: Theory and Practice*, 8(1): 25, pp. 1–13. DOI: https://doi.org/10.5334/cstp.538

Submitted: 01 July 2022 **Accepted:** 22 May 2023 **Published:** 05 June 2023