197 *Etica & Politica / Ethics & Politics*, XXV, 2023, 2, pp. 197-225
ISSN: 1825-5167

TALKING METADATA. UNDERSTANDING PRIVACY IMPLICATIONS OF VOLUNTEER CONTRIBUTIONS IN CITIZEN SCIENCE PROJECTS

GEFION THUERMER
King's College London
gefion.thuermer@kcl.ac.uk

LAURA KOESTEN
University of Vienna
laura.koesten@univie.ac.at

ELENA SIMPERL
King's College London
elena.simperl@kcl.ac.uk

ABSTRACT
Citizen science (CS) projects typically have citizen scientists with different levels of expertise and agency contributing data or knowledge. Every contribution leaves traces of their involvement, including metadata such as locations or emails. Through four case studies this paper explores the generation, use, and publication practices of CS projects' metadata. We use a mixed-method approach combining document reviews, interviews, and an online survey, to generate insights into current metadata practices and perceptions of project contributors and organisers. We identify several weaknesses in CS projects' data collection practices: Participants have only limited awareness of the metadata they contribute, and the privacy implications it can have. Matching expectations between project contributors and organisers regarding acknowledgement is crucial - and metadata play a key role. Projects need data processes and documentation aligned with open science principles, and clear communication to contributors about the data they collect and use. Finally, projects need to consider the mental models of contributors in relation to personal data and associated risks. We derive key considerations that data-intensive CS projects should make in their initial design phase, to generate consistent metadata in line with their participants' expectations, which in turn increases transparency and thus can increase data reuse.

KEYWORDS
Metadata, citizen science, mixed methods, licensing, privacy

INTRODUCTION - THE ROLE OF CITIZEN SCIENCE IN SCIENCE AND SOCIETY

In citizen science (CS) projects, volunteers collect and share data with the project staff, other volunteers, and the public. The European Commission defined citizen science as the "general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources" (2014, p. 6). CS encompasses diverse types of projects and data, and is reliant on public participation. Citizens in CS projects can be engaged in many different ways, and ideally, should be considered throughout the entire research lifecycle (Thuermer et al., 2022). This not only provides for actual rather than simulated participation (Arnstein, 1969) in the project and with the data itself (White, 1996), and more inclusion and representation in projects (Cooper et al. 2021), but also increases justice in the management and use of citizen generated data (Christine & Thinane, 2021). If science becomes more inclusive and open by adopting citizen and open science principles, it will be better able to respond to the needs of the communities it aims to serve. CS can advance science, contribute to innovation processes, and people in the scientific discourse (Bonney et al., 2009), and contribute to the United Nations' Sustainable Development Goals (Fraisl et a., 2020).

CS projects commonly collect data for various purposes, and 'data' in this context should be understood as pieces of information, whether these are images, observations, descriptions, categorisations, physical samples, audio files, or other data types. Collections of data are defined as datasets and might be published as the result of a CS project. Such datasets need to be described, to contextualise them, both for human as well as machine consumption. Any such description (which can be more or less structured) is understood to be metadata in the context of this work, as it constitutes "data about data". Here we investigate how CS projects collect and process metadata in practice and how formalised these work practices are.

Data in CS projects has mostly been discussed from the perspective of data quality, and how to make the datasets resulting from CS projects more fit for scientific reuse (Riesch & Potter, 2014), and little focus has been given to metadata. Data quality has been identified as a consistent issue (Ponti & Craglia, 2020), especially where data is meant for the use of academic research. Understanding and improving the data practices of CS projects can help mitigate the known issues of distrust in data and metadata quality, and in transparency of CS projects (e.g. Hunter et al., 2013, Ottinger 2010). Burgess et al. (2017) recommend that metadata on the collection protocols of data should be included in CS datasets to this end.

If citizen scientists are to be attributed for contributions that are subsequently used in publications, metadata about the contributors is required to enable this; so metadata can include personal data. Others have investigated potential privacy implications in the context of CS contributions, e.g., through location data

embedded in submissions (e.g. Bowser et al., 2014; Xia et al., 2017). Here we expand this focus by looking at contributors' awareness and expectations, not only regarding the protection of their privacy and their intellectual property rights, but also public attribution for their contributions.

Using a mixed methods approach we studied the contributors' awareness of metadata, its implications and potential risks. We conducted an analysis of project documentation, interviews with project organisers and contributors, and a survey with contributors, to triangulate different perspectives on the topic. We asked coordinators and contributors about the metadata that they provide - which may implicitly or explicitly include personal data. Furthermore, we investigate expectations on attribution in the projects' contexts, both for direct participation as well as for eventual future outputs (e.g., scientific publications, success stories, etc.) and how this is communicated as part of the projects' documentation. We do this by investigating four CS projects, all funded through the EC ACTION (Participatory science toolkit against pollution) project.[1]

Our findings point to several weaknesses in data collection practices, due to limited considerations of metadata, privacy risks and contributor acknowledgements. For instance, they show that participants have only limited awareness of the metadata they contribute and the privacy implications this metadata has. We argue that a thorough documentation of metadata would be useful to help participants understand exactly what data and metadata they contribute and what implications these data have, to make their contributions both more valuable for data users, and more ethical for participants, who would be fully aware of what data they contribute and to which end. We further find that expectations with regards to acknowledgement differ both between and within projects, and that appropriate communication strategies can pre-empt many of those risk factors. Lastly, we find that awareness of privacy implications and risks among project organisers can successfully be conveyed to participants through appropriate communication strategies.

BACKGROUND - WHY CITIZEN SCIENCE AND ITS DATA MATTER

Citizen science projects

Citizen science projects actively involve lay people in one or more aspects of research. This may include research design, data collection, recruitment, data analysis as well as interpretation of results, or publications (Riesch & Potter, 2014). CS projects may occur at small, local scale, or as international collaborative ventures, collecting hundreds of thousands of data points. One example of a large-scale CS project is eBird, which boasts 150,000 participants and contributes data for scientific

[1] https://actionproject.eu/

research in ornithology. Making the data usable for scientific research requires a consistent level of data quality, which is supported through a combination of an intuitive user interface for data entry, with automated filters that support participants' categorisations, and expert reviews of the data entries (Lagoze, 2014). Feedback and rewards have been shown to be effective tools to motivate citizen scientists to engage, and to enhance the quality of the contributed data (Reeves et al., 2017). CS has huge potential to support policy development (Hecker et al., 2019) and the UN Sustainable Development Goals, but to realise this potential, their output data quality has to become more consistent and reliable (Fraisl et al., 2020).

Citizen science is often conflated with data collection by citizens; in an ideal scenario, citizens should not only be contributing or collecting data, but be involved in the project in a broader sense. The European Citizen Science Association (ECSA) has developed ten guiding principles for CS, to ensure it is conducted responsibly, and achieves impact. These principles include the active involvement of citizens, genuine science outcomes as a goal, collaboration between scientists and citizens across project stages, and data made publicly available (Robinson et al., 2018). However, the principles also assume the projects to be led by professional scientists, which is not always the case - there are numerous bottom-up CS projects that are driven and implemented primarily by citizens (Miyashita et al., 2021; Oudheusden & Abe, 2021). Despite the crucial role of data and its contributors in CS, there is no overarching best practice for data use and attribution.

Metadata in CS projects

In CS projects, data can be many things, and there is no one definition. While traditional definitions commonly include the word "fact" (e.g. numerical facts, collected together for reference or information (OED)), critical discussions convene on a more representational view of data, emphasising data context and focusing on the "making of data" opposed to a positivist notion of data (Bokulich & Parker 2021, Leonelli 2020). Following this viewpoint there are definitions that also include relational properties of data (Borgman 2012) which take interactions around data as a part of context into account (Neff et al., 2017). This highlights the importance of metadata as an instrument for capturing context.

In this work, we understand data from the viewpoint of the citizen scientists, namely the pieces of information collected by citizen scientists for the purpose of generating insight for the CS project. Depending on the project, data could consist of images, observations, descriptions, categorisations, physical samples, audio files, or a variety of other details. For example, in the eBird project, contributors record observations, images and sounds of birds, all of which are entered into an online platform; in ACTION's Water Sentinels projects, participants collected water samples along with metadata, such as location and date/time.

Metadata is data about a dataset (or about data in a dataset). It describes properties of a dataset, such as its title and description, contributors, etc. Different metadata

schemes are developed for research data within and across disciplines, to make data interoperable and discoverable by machines (DataCite Metadata Working Group, 2019). The adoption of unified metadata practices improves data exchange possibilities and scientific transparency. Examples for general purpose metadata vocabularies are the Data Catalogue Vocabulary (DCAT[2]) or Schema.org[3]. There are many other, domain specific approaches aimed at improving and standardising metadata entries. Metadata often uses specific vocabularies and technical formats, which means its understanding can be challenging (Mayernik, 2011; Edwards et al., 2011).

Contributors to CS datasets supply a certain amount of metadata about themselves, depending on the project setup and structure (e.g. whether data is collected manually or online). Hence, the nature of the data type and format contributed in CS projects can lead to specific metadata challenges regarding privacy, data quality, and ownership, for example if location data is shared by participants unaware of potential privacy exposures (e.g. Bowser et al., 2017), or contribute data without awareness of the plans for its ownership and publication (Resnik et al., 2015). Access to data can be allowed at different levels, as researchers weigh which data to make open, when, for whom (Levin & Leonelli, 2017), and how sensitive data can be made available without posing privacy risks to contributors. Wong et al. (2022) suggest that involving data subjects in the co-creation of data protection regimes can enhance their effectiveness and alleviate potential power imbalances between stakeholders.

In the context of reusability of CS data for scientific research it has been pointed out that metadata should include details about data collection and analysis, to ensure scientists have sufficient confidence in data to actually use it for their research (Burgess et al., 2017). Projects such as CitSci.org have developed metadata documentation features that support different standards and community-driven metadata fields, and developed award schemes to incentivise people to supply comprehensive metadata (Wang et al., 2015). The US-based Citizen Science Association has recently developed a metadata standard for Public Participation in Scientific Research projects: PPSR Core[4]. While using these schemas could address issues such as insufficient documentation of the research design, implementation, or quality control, the application of schemata still requires expertise. We argue that while the meaningful use of metadata standards can be challenging even for experts (Attig et al. 2004; Koesten et al. 2020), they can present particular barriers to involving citizens in knowledge generation as they require expertise not necessarily available to bottom-up CS project teams and their contributors. In CS projects, decisions of what to capture, publish and report are often made without concrete guidelines on the potential risks and implications (Thuermer et al., 2023), which

[2] https://www.w3.org/TR/vocab-dcat-2/

[3] https://schema.org/Dataset

[4] https://core.citizenscience.org/

means that citizen scientists contribute without full awareness of what will happen with their contribution - and unable to question or fully consent to this use. The nature of the contribution of a specific project can lead to the collection of metadata that the contributors might or might not be aware of. This includes for instance the submission of geolocation data as part of data collection efforts in the real world, which has been pointed out as a risk for privacy (Bowser et al., 2017). Some projects explicitly require the contribution of personal data, including the contributors' identity, which mirrors the role of a "data publisher" in traditional metadata schemata (e.g. DCAT).

Aside from data about people, questions of intellectual property rights, such as copyright on contributed images, have also been discussed in the context of CS projects. This points to the fact that while rights vary with the contributed data type, it is essential to consider data ownership in advance, to avoid later issues with dissemination and use of research datasets that contain copyright-protected contributions without authorization (Scassa & Chung 2015; Resnik et al., 2015). Riesch & Potter (2014) raise the question whether contributors should be authors on outputs, which would in turn have implications for their privacy: if the licensing of their contributions requires acknowledgement, then their names (or pseudonyms) need to be collected and potentially published as metadata. All these issues culminate in questions of how CS projects collect and process both data and metadata, which we will explore in more depth in the following sections.

Methods

This study was conducted in the context of the ACTION project, which supported and co-designed tools with 16 CS pilot projects. Case studies were selected from the nine pilots that were active at the time of data collection (October 2020). We excluded pilots who worked with minors, as this kind of data has a different set of implications, or those that only collected data anonymously, as it would not have yielded insight on the privacy or acknowledgement issues we were interested in. Four pilots were selected for inquiry, which collected six types of data in total: i) images and contextual information of streetlights, ii) pictures of the night sky, iii) neighbourhood sound samples, iv) counts of dragonflies and butterflies, v) water samples, and vi) images of water bodies. Three of the projects were led by public authorities or professional scientists and primarily engaged contributors in data gathering, while one project was conceived and its design heavily informed by citizens. The projects' data is used to support policy decisions (such as environmental protection or traffic management), as well as research into different forms of pollution.

To understand the projects' metadata practices, we conducted an analysis of project documentation, interviews with project organisers, and a survey with contributors, to triangulate different perspectives on the topic. We used a mixed methods approach to gain insights into how data and metadata was conceptualised.

The documentation analysis gives us a non-intrusive way to learn about the projects' data practices which we could then expand on during the interviews. We used the survey to gain quantitative insight into the perspectives of a larger sample of contributors, to add more breadth to our inquiry.

For all projects, we conducted a document analysis (Bowen, 2009) on all relevant documentation that the projects shared with their participants. A list of all documents is provided in the appendix. This analysis aimed to understand, in as much detail as possible, how the projects conduct their data collection, what and how data and metadata is collected, what role participants play, and how they are informed about their role, contributions and attribution.

Based on these insights, we conducted semi-structured interviews with the project organisers, which allowed us to go into more depth around data and metadata collection, and considerations they made in the planning of these processes. Specific foci were privacy implications of data collection for the participants and any other risks inherent to the data, and the ownership and use of the research data.

Building on findings from both document analysis and interviews, as well as literature on CS, we designed a survey for project participants to explore how they perceive their engagement and the data and metadata they contribute. The survey was administered via MS forms, and available in three languages: English, Spanish, and Dutch[5]. It was structured in two sections: Participants' engagement and motivations, the data and metadata contributions they make, what role they believe metadata plays, the risks they associate with their activities, and how they expect their contributions to be used and acknowledged; and socio-demographic information, including age, gender, education, and country of residence. An overview of the whole survey can be found in appendix 2. Questions were a combination of Likert scales (for awareness / relevance), scales for motivations and risk perception, and single and multiple choice, with the option to add additional categories. Participants were also given the opportunity to volunteer for a short follow-up interview, and to add supplementary comments.

The survey was sent to all 334 participants associated with the three organisations engaging in the citizen science projects: UCM (Azotea and Street Spectra), DBC (butterfly and dragonfly monitoring; BDM), and BitLab (Noise Maps) (all described in the 'Findings' section below). While DBC has significantly more participants engaged in butterfly and dragonfly monitoring, the survey was only sent to those who also engaged in the water sample collection.

18 survey respondents volunteered for an interview. All of them were contacted, and three interviews were conducted. They were analysed together with the interviews with project organisers, but not included in this paper, as they only confirmed the insights from the survey.

[5] Translations from English were completed by project organisers who are native speakers in those languages.

Table 1: Overview of survey responses

	Sent	Recipients	Responses	Response rate
Street Spectra[6]	18-Sep-20	54	8	15%
Azotea[7]	02-Jul-21	13	11	85%
Noise Maps	10-Sep-20	12	5	42%
BDM	18-Sep-20	255	84	32%
TOTAL		**334**	**108**	**32%**

The survey results were analysed using descriptive statistics to identify differences between the views of project organisers and contributors, and chi-square tests to identify relevant correlations between participants' projects, views, and characteristics.

The study was approved by the institutional Research Ethics Office at King's College London, under reference MRA-19/20-20327. Informed consent was given by the participants through the survey as well as prior to the interviews.

FINDINGS - FOUR CASE STUDIES OF CITIZEN SCIENCE PROJECTS AND THEIR PARTICIPANTS

We spoke to four projects, **Street Spectra**[8], **Azotea**[9], **Noise Maps**[10] **and BDM**[11], hosted by three organisations, which we describe here in conjunction with the results of our analysis. **Street Spectra**, hosted by *Universidad Complutense de Madrid (UCM)*, engages a wide group of citizen scientists, who take photos of light spectra with their smartphone camera, and upload them to an open online database. The goal is to collect data on light pollution through streetlights over time. Contributors require a smartphone and a low budget handheld device, which the project provides them with, to participate. Contributors' main point of interaction is an app, which runs on the epicollect platform[12]. This helps to ensure consistency of contributions. Contributors can add, change or remove data. They are authenticated through

[6] Since it is impossible to reach out to Street Spectra participants directly, UCM sent the survey to astronomy clubs they told about the project to recruit citizen scientists; we do not have exact numbers of their members, or the proportion of members who engage in Street Spectra.

[7] As we only received one response to the survey from Azotea participants during the initial data collection phase, we decided to redistribute the survey during revisions of the paper.

[8] https://streetspectra.actionproject.eu/

[9] https://guaix.ucm.es/azoteaproject

[10] http://www.bitlab.cat/projectes/noise-maps

[11] https://www.vlinderstichting.nl/english/

[12] https://five.epicollect.net/

Google[13], so the platform itself does not process their personal details. It also provides detailed guides for participants, explaining how to collect and submit contributions, navigate the app, create new projects etc., and hosts the submitted data on a publicly accessible database.

The project provides an in-depth tutorial on how to take pictures of light spectra, and how to categorise them and identify the type of lamp that creates them. The goals of the project, as well as the data it collects, were well explained in the documentation. However, guidance of the app was limited to documentation from the app developers (which is independent of the project), and there was no specification of what would happen with the data, aside from it being published in a publicly available database. The privacy policy of the app suggests that all data is owned by the project, while users grant the project a licence to their contributions - which is contradictory in itself, and could not be clarified in our investigation.

The use of a project-external app means limited control over what data is collected in practice. This became clear when we attempted to distribute our participant survey, and UCM was unable to reach out to their participants directly, because they had no structured data, such as names or email addresses, about them. In the long-term, the project aims to develop their own app, which will also allow participants to identify the relevant light spectra on the photo, and categorise the lamp based on it. While the platform and project documentation both discuss different aspects of data collection and submission processes, as well as some of the metadata, they do not specify the use or ownership of the data, or privacy implications for participants. Combined with the above-mentioned contradiction in the platform policy, this may lead to licensing and/or privacy issues in practice.

One reason for a lack of privacy considerations became apparent when interviewing the project organisers: The use of the phones' geolocation could be a privacy risk, but since participants are expected to submit photos of *street*lights, this location should not be their home, and thus should not make personal information public. However, the public data shows that participants have uploaded geotagged photos of indoor lighting. Moreover, the geolocation submitted via the app is the one where the phone is located at the point of submission, which is not necessarily the same where the photo was taken. Three of the seven respondents to our survey who participated in Street Spectra also stated that they do provide their home address as part of their metadata. As one project organiser illustrated:

"What I have been doing is, at the moment that you take the picture, I just upload it to the server. I have not tested a use case where you go home and then upload the image.[...] Our interest is for public lampposts, not for indoor illumination [...] I have not seen it [images of indoor lighting], maybe this is for cosmetic reasons or someone wants to upload something, but it is not really our target. [...] I'll have to check that" (Project organiser, UCM)

[13] https://developers.google.com/identity

Azotea, also hosted by *UCM*, engages with hobby astronomers who set up their personal cameras to take regular pictures of the night sky throughout and beyond the lockdown period in Madrid in spring 2020. Their goal is to measure the effect of the lockdown on light pollution. The documentation of the project was not well developed, not least owing to its newness at the time. Azotea provides a guide for contributors that explains how to set up their camera and collect the image data, which also outlines the goal of the project, but gives no indication of what metadata would be collected. Contributors in this project have a close personal relationship with the research team and are heavily involved in the entire project, including the academic publication process. Hence a lot of information, while not part of the documentation, is passed on via direct email or conversations. This makes documentation less necessary, but simultaneously more vague and out of sync with the actual processes in the project. For example, the documentation was clearly written before the project launched, suggesting the processes were still being developed, even though the project had been actively collecting data for months at the time of our analysis.

Contributors provide their personal details, such as name and email address, which is also used for publications. All contributors submit photos from their home (but no location data), and upload the photos to a central server. They all have full read and write access to the entire dataset. A first academic publication is in progress, and the citizen scientists will be named on it. The close involvement in the project lifecycle also means that organisers set different expectations for contributions and acknowledgement:

"It is also a matter of science for us in Azotea. In the beginning it was really targeted at dedicated amateur astronomers, we have had more or less 15 of them. Maybe now about 5 are active. So there is a scientific paper coming out of this, and we want to give credit to these people with real names. [...] We knew from the beginning that this would be targeted at very few." (Project organiser, UCM)

Noise Maps is hosted by the Spanish NGO *BitLab*, and engages with citizens in Barcelona to record and map the soundscape of two local neighbourhoods. Sound sensors are installed in participants' homes, collecting regular sound samples. Participants can also record audio samples during walk-workshops, passing by pre-selected points of interest throughout the city. The sound samples are then processed automatically to generate insight into the types of sounds that are audible at different times and days, such as traffic, construction, people, or wildlife. This data can help maintain local cultural heritage, but could also be used in policy discussions about noise pollution.

The project was initiated and conceived by a local community, who then approached BitLab for support. Participants discussed the project and their contribution at a workshop, at which the goals of the project, its data collection and processing were explained, and privacy and security considerations discussed. The entire data collection and analysis protocol was co-created with the participants.

Further guidance is provided to contributors in the projects' documentation, and when sound sensors are set up at their homes. Noise Maps has extensive documentation about the project, its data collection and usage. They also have detailed information and a protocol for contributor consent. The project is very clear that potentially sensitive personal data will be collected in sound recordings, from both contributors and unrelated bystanders, and that this data is subject to special protection in raw form, and therefore undergoes anonymisation before it is published: "There is no way around putting some signposts, letting bystanders know that sound is being recorded." (BitLab)

All data is made publicly available; however not in its raw form. As recordings may include conversations of potentially sensitive nature, all human voices are distorted. As a further safety measure, sound sensors are set up such that the participants who collect the data do not have access to the raw sound files. This was a conscious decision by project organisers and contributors, in order to protect the privacy of those who may be unwittingly recorded. Since contributors cannot access the files, they cannot retract them directly; however, the project organisers would delete any data if this was requested. All these measures bring home the point to contributors that the data is sensitive and a potential privacy risk; if not to them, then to the people they record. Unsurprisingly, awareness for this risk among Noise Maps participants - the project that co-created the entire data collection and analysis protocol and specifically discussed these risks - was highest among our survey participants.

The **butterfly and dragonfly monitoring project (BDM)** hosted by the *Dutch Butterfly Conservation (DBC)* keeps track of the butterfly and dragonfly populations in the Netherlands. Contributors walk specified sectioned routes and count species they encounter, which they then enter into a central database. They are set up with a login to this platform when they start collecting data, and can use this to submit or amend their observations. Submissions can be made through an app or a website. The collected data is used by Statistics Netherlands (the national statistics office) to measure and predict butterfly and dragonfly populations, and inform environmental policy-making. As part of ACTION, BDM investigates the relationship between the occurrence of dragonflies and pesticides in local waters on some of their existing routes. In addition to species counts, contributors in this part of the study also collect water samples and photographs, which are analysed to understand the effect of pesticides on dragonfly populations. The data is published only in aggregated form, with no recognisable links to individual observations, as outlined by one of the organisers:

"We don't publish the data as is. The data is included in the national database for flora and fauna. But you can't recognise it as being from monitoring transects. [...] If you have an account, so someone who does one of these transects, and in there is a code, and from that you can look up who that is, the address and all that stuff." (Project organiser, DBC)

The monitoring has a thirty-year-long history, and BDM makes a large amount of guidance documentation for participants available, explaining how they can get involved, how to count the insects, and how to submit their data. Guidance we reviewed included tutorials for the count of dragonflies, the collection of water samples, and the submission system for counts. All documentation was clear on which data would be collected and how it would be used, but less clear about metadata. Our interviews showed that metadata was mainly limited to user access, which was linked to a database of contributors - however, in practice this personal data was used for authentication, but not linked to data submissions by those contributors. Our interviews with both the project coordinator and several participants showed that participants who join the project are on-boarded with a visit from DBC staff to help them set a personalised route, and create a login to the online platform which they then use to submit their observations. Interviews with participants of BDM showed that they were very aware of the data they supplied, such as insect counts. However, they did not consider personal information they provided to enable their participation in the first place, and that the organisation held about them irrespective of their individual data contributions. A participants account on their personal data:

"It is a set place, you always walk the same route, because if it is in another place you cannot compare [....] but yeah they know where the route is, they know exactly within 50 metres where I saw the butterfly. [...] They know who I am, they know how to reach me, they know my email address, my telephone number, my address, I think they know my age ... But you only give that once." (Participant, DBC)

In summary, our document analysis showed that all projects provided documentation of data collection processes to contributors, though in varying levels of detail. Our interviews showed that while the documentation may be extensive, three of the projects involved additional, equally extensive personal interactions to help contributors set up their equipment or route, or communicate details about the project. Thus, these contributors had more information about their engagement than the documentation suggests. However, this may also mean that contributors had different levels of knowledge and awareness, depending on their individual interactions with the projects. Furthermore, a requirement for personal interactions does not allow projects to scale up easily. Table 2 below summarises the different types of metadata that each project collected, according to the document analysis and interviews.

Table 2: Metadata collected by project

	Street Spectra	Azotea	Noise Maps	BDM
Contributor name		x		x
Nickname / ID	x		x	x
Date / Time of contribution	x	x	x	x
(GPS) Location of contribution	x	x	x	x

Measurement specifications (e.g. type of sensor / camera)	x	x	x	x

Following our document analysis and interviews with project organisers, we sent a survey to the projects' citizen scientists. It received 112 responses, with 108 related to the target projects, and had a cumulative response rate of 32% (see Table 1). The majority (76%) of respondents are engaged in butterfly and dragonfly monitoring, which is expected, due to it being the oldest and most established of the projects, with the largest pool of contributors. While the other projects have significantly less contributors, we have received responses from a sufficient proportion of their contributors to make comparisons viable. The average age of participants is between 56 and 60. 72% of participants identified as men and 74% of participants hold a university degree. While the education distribution reflects general trends in citizen science (e.g. Domhnaill et al., 2020), the gender balance among participants in our sample was more male than is typical for such projects (Paleco et al., 2021). The gender distribution was stable across projects, with 60-71 % identifying as men; the education distribution differed for NoiseMaps, where 40% of respondents held vocational degrees.

The survey showed that the data participants state to contribute is well aligned with what the project expects (see Figure 1 below): Primarily observations and images for BDM, images for Azotea and Street Spectra, and sound files for Noise Maps - although only a small proportion of BDM participants noted the physical (water) samples.

The picture is not so clear-cut for the metadata. While Noise Maps participants agree on exactly what metadata they contribute (timestamps, locations and specifications), there is some variation in Azotea and Street Spectra. Part of this can be explained by participants entering different details into the Street Spectra app, with date/time and location being required - and automatically collected - from all of them. The amount of metadata BDM holds about participants however is not consistently recognised by its' contributors: they report to submit observations, often combined with a location and other factors such as weather conditions; but 15% also say they do not contribute their name - which is associated with the account they use to submit observations. Similarly, participants in Azotea may not re-share their location data for each submission, with this data being on file with the project organisers already. Interviews with participants of BDM indicate that while contributors are aware this data is held by the organisers, they do not consider it part of their observation recordings, even though in practice they are linked.

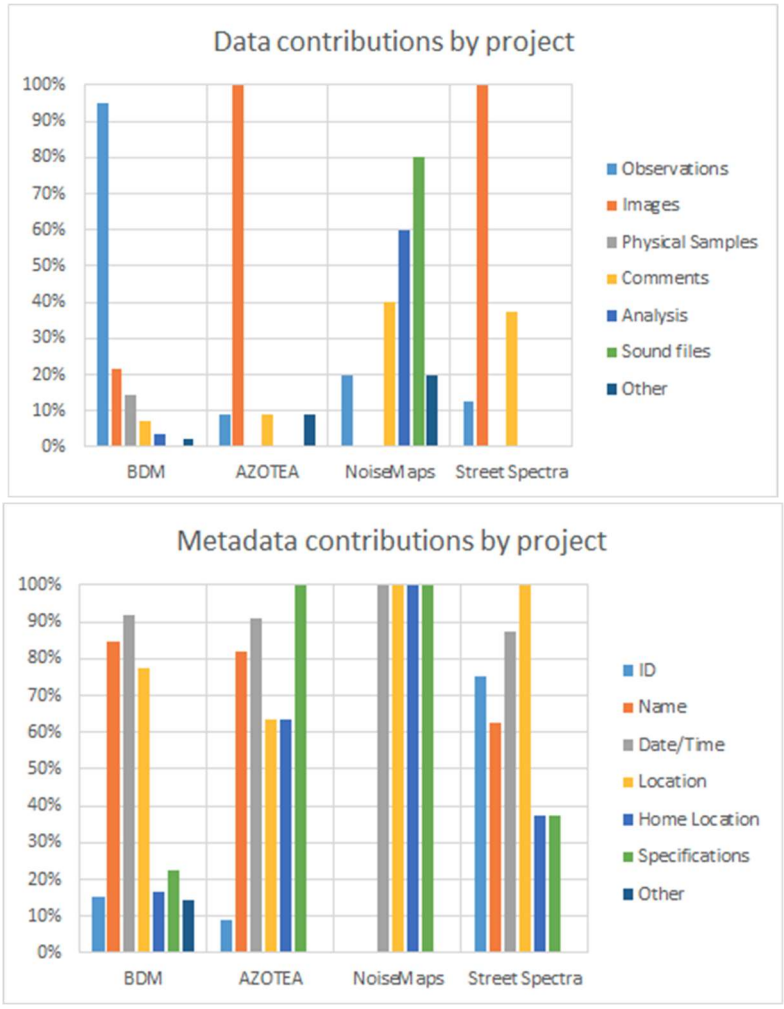

Figure 1: Survey result: What is your data contribution to the project? What metadata do you contribute to the project? (% of respondents by project, n=104)

Most participants (60%) do not expect to be acknowledged for their contribution, and acknowledgement is not important to them (54%). BDM participants show a surprising variety of acknowledgement expectations: 70% do not expect to be acknowledged, but only 58% say that acknowledgement is not important to them. None of the respondents contributing to NoiseMaps state that acknowledgement is important to them, and only one participant of Street Spectra has this expectation. However, some of these participants still expect to be acknowledged personally. Contributors can only be acknowledged personally for their contribution if the projects collect metadata on who made which contributions. We found a significant, though little surprising, correlation between participants' expectation of being acknowledged, and the relevance acknowledgement held for them (χ^2 (2) = 32.3, p < 0.000). Participants who identified as men had a significantly higher expectation of being acknowledged for their contribution than those who did not (χ^2 (1) = 8.9, p = 0.003).

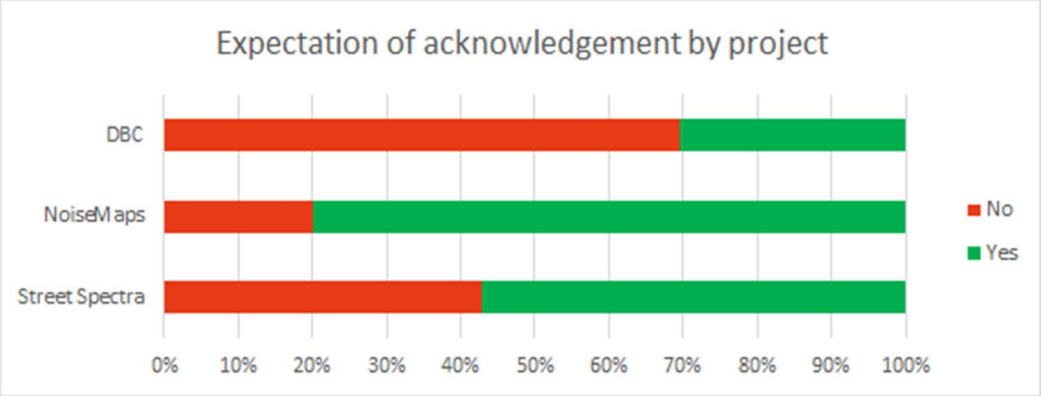

Figure 2: Do you expect your contribution to be acknowledged in project outputs? (Binary; n=94)

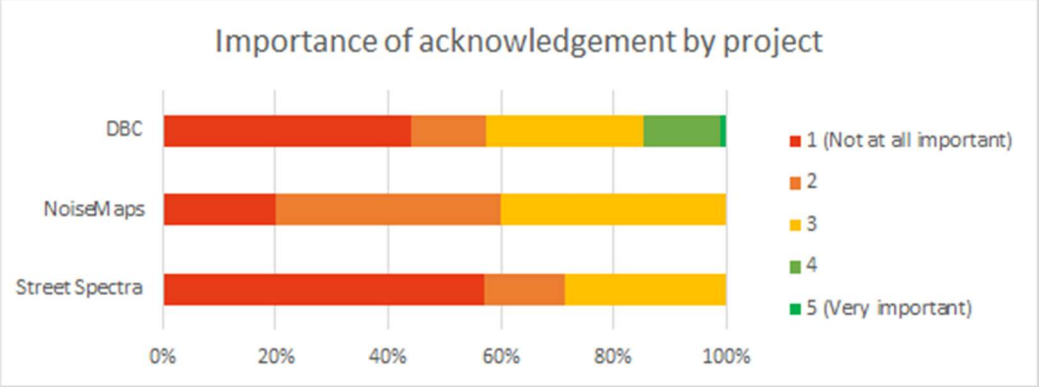

Figure 3: How important is it to you that your contribution is acknowledged? (Likert, n=94)

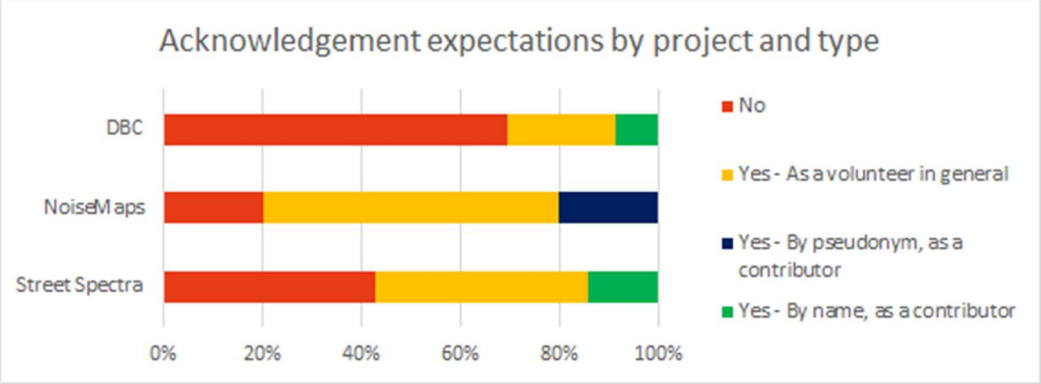

Figure 4: How do you expect to be acknowledged? (n=94)

We further found a significant correlation with participants' expectations and metadata: Those who expected their contribution to be acknowledged in publications cared more about what would happen with the metadata they contributed ($\chi 2$ (2) = 7.9, p = 0.019). This makes sense, given the above-mentioned necessity to record who contributed what in order to enable acknowledgement. We also found that participants who contributed images ($\chi 2$ (1) = 6.208, p = 0.013) and

observations ($\chi 2$ (1) = 11.708, p = 0.001) were significantly more likely to expect their contributions to be acknowledged in reports. Participants who provided measurement specifications as part of their metadata placed more value on being acknowledged for their contribution ($\chi 2$ (2) = 7.984, p = 0.018). There was no correlation between the motivations for participants to engage in CS projects and their expectations of acknowledgement.

Asked about perceived risks of their engagement, the primary risk participants acknowledged to even a small degree was with regards to privacy; both their own (16%) and others' (8%) (see Figure 5 below). BDM participants were least concerned about privacy, while participants in Noise Maps were mostly concerned about other people's privacy and reputation. This is likely due to the type of data collected in this project, as well as the intensive discussions within the project, and the coordinators' focus on this concern. Participants who recorded data at their home were significantly more concerned about their own privacy than those who did not ($\chi 2$ (1) = 4.654, p = 0.031).

Figure 5: Do you feel your contribution poses a potential risk to...? (n=108)

DISCUSSION - ETHICAL ISSUES WITH CITIZEN SCIENCE AND METADATA

Data- and risk-awareness

Our findings show that participants' awareness of the metadata they contribute varied widely. All the projects collect similar metadata for contributions, but the data that the projects objectively collect does not align well with participants' varied responses when being asked what metadata they contribute.

This difference was especially obvious in the case of the BDM project, where participants recognised the metadata they recorded, such as weather conditions or locations, very consistently. At the same time they did not consistently acknowledge that the login they use to submit observations is linked to their personal information. This personal metadata is not used or published, and our interviews showed that participants *are* aware that the organisation holds their personal details. While they did not acknowledge that the data was linked to contributions, they also did not mind this link when it was pointed out. Their awareness of risks on the other hand was a lot lower, although there are some risks related to the project's data collection practices. One such risk was reaching areas in which butterflies are common outside public pathways, which can require the use of a permit, lack of which might carry penalties.

The awareness of participants in BDM contrasts clearly with the contributions to Street Spectra, where participants and project organisers seemed to occasionally misunderstand one-another with regards to not only the metadata, but the contributed *data* itself. The project asks contributors to submit images of spectra of streetlights, but participants took this as an invitation to submit images of *any* light spectra, including living room lamps. This is problematic, as contributors submit photos from their home, which are then published in a public database without review or quality assessment, including their geolocation. This poses a major privacy concern: contributors may unwittingly broadcast their home location along with their nickname or even their full name. This is in line with Wang et al. (2015) who point out that many CS projects do not fully consider the procedures required to move from data collection to making data publicly available for reuse by others. Alongside the obvious privacy risk, these errant contributions also make the dataset less reliable: the locations of the source of light pollution may be incorrect; and images of light sources that do not meet the requirements become part of the dataset.

Yet another different perspective on risks is added by the Noise Maps project. Participants in this project contribute sound recordings, not of themselves but of their environment, which potentially includes bystanders' conversations. This is further exacerbated by the location of some of the sensors, in a neighbourhood associated with the local red-light district, which meant that the potential content of unwittingly recorded conversations could be highly sensitive. Specific risks the project had to consider included the privacy and reputation of both participants and

bystanders, and the security of their participants in case bystanders were unhappy about the sensors. The project organisers were very aware of these potential issues, and have taken steps from their set-up to mitigate any risks to contributors or bystanders. This included briefings on the risks and issues, signs informing about the nature of recursings, limiting access to the data, and developing mechanisms to anonymise the recordings, so they could be published openly without endangering sensitive personal information. This awareness of the project organisers has translated into processes for data collection, and recognition of the associated issues by contributors. Potential risks to both contributors and bystanders' privacy is summarised by one of the project organisers:

"This neighbourhood has had certain problems (...), for example drug sales or prostitution (...) Some of the volunteers that we have live in streets that are at the core of these problems (...) It is quite likely that we might inadvertently record conversations that people might not want to have recorded. (...) We have to discuss [this problem] with our volunteers: We have our own protocol that we developed [with an ethical review board] (...) with respect to the privacy of not only the active participants but also the passive volunteers." (Project organiser, BitLab)

Noise Maps show that awareness of privacy implications and risks among project organisers can successfully be conveyed to participants, if it is included in project documentation and processes, and communicated in a way that makes sense to the citizen scientists. CS projects rely on participation of non-expert volunteers, hence they are particularly vulnerable to concerns of rigour and reliability (Roman et al., 2020). In that context it is paramount to also consider implicit contributions that take place without contributors being aware, due to the project setup or the technology used, as well as contributions that happen by misunderstandings of the projects' goals, as was the case for Street Spectra. These unexpected and unwanted contributions and associated risks could be mitigated through clearer communication about what the project requires (spectra of outdoor lighting, with accurate locations) and what data is submitted (geolocation at upload time).

While the project organisers mostly stated that they would not use personal data about their participants they also did not have procedures in place to delete unnecessary metadata about their contributors. Clear processes documenting the expected activities to be carried out to gather, prepare and validate data before publishing it are needed to avoid unintended violations of ideals of ethical science. The projects' onboarding activities, some of which happen verbally through tutorials, conversations, and having contributors embedded in a personal network, served this purpose. However, these do not replace the need for formal processes to ensure consistency and transparency across the projects, and to other users of the data.

Overall, this means that more thorough documentation of both data and metadata collection, as well as associated risks, combined with participant introductions or training, serve not only to create better data and more transparency of the project to scientists, but also helps participants understand exactly what they

contribute, and conduct their own risk assessment and mitigation. This can be seen as the prerequisite of informed consent in such settings and is therefore a key element of a successful project to avoid exploitation, as mentioned by Resnik et al. (2015). Co-creation of data protection frameworks, such as suggested by Wong et al. (2022) could help address this issue.

While standards for CS metadata such as PPSR Core are being developed, and have been adopted by some of the large platforms, they may not be sufficiently accessible to smaller projects, especially those that are developed bottom-up and have limited or no dedicated technical expertise and resources. In order to be inclusive of *all* CS projects and citizen scientists, such solutions need to be made understandable *and usable* for lay people. Else such projects run a risk of conducting a form of data colonialism (Thatcher et al., 2016), where contributors contribute data that they are ultimately unable to understand or use themselves.

Moreover, CS projects, and those supporting or encouraging CS, should consider communication strategies that explain goals and requirements and why they matter to different stakeholders - especially citizen scientists - in clear and accessible language, to ensure that new projects, whether they be set up by experienced researchers or bottom-up by citizens, follow best practice and create datasets that are useful for the research they want to support. Confusion could be further mitigated through use of visuals elements (Erwig et al., 2017), indicating to contributors in an easily understandable format what data is required.

ACKNOWLEDGEMENT EXPECTATIONS

We find that expectations of acknowledgement differ both between and within projects. Acknowledgement is only possible in relation to the metadata that is captured: only where project organisers know who made individual contributions can they attribute individual participants. Only one project - Azotea - consistently captured participants' details with the intention to do so. It is clear from our survey results that participants heard this message loud and clear, and consequently submit the necessary metadata and expect to be acknowledged.

Street Spectra offers participants the opportunity to enter a name or nickname when they submit data, but it is not a requirement. In practice, that means that acknowledgement may be possible in some cases, however, not in a consistent manner. This becomes more complicated when considering the type of data: contributors submit not just observations, but also images. It is not clear exactly how submitting them affects the ownership of those images, though documentation suggests the images are owned by the individual users and licensed to the project. This information is only listed in the privacy policy of the app, which is independent of the project. Combined with the above mentioned contradiction in the platform policy, this may lead to licensing and/or privacy issues in practice: If researchers wanted to use images in publications (as opposed to analysis), they would *have* to

name the contributor, but their data would not allow for that: The Street Spectra team do not know the names, and if they did, they would have no way to get consent from contributors. Our survey results suggest that only half of participants submit these details. A look at the public database[14] suggests that 21% of contributions come with identifiable names, 72% are made under pseudonyms, and only 7% are made completely anonymous.

Two out of eight of the survey respondents of Street Spectra expected to be acknowledged by name - a sizable proportion for a project that does not advertise this opportunity. Being named as the licence holder of an image, even if it were the right approach for the use of the licence, might still make the contributors who neither expect nor want to be acknowledged uncomfortable. This limits potential reuse of the contributed data considerably, especially given that our survey results suggest that the majority of participants are motivated by their support of the goal and contribution to the research, and not as interested in acknowledgement as professional scientists might be.

While the BDM team does know which observations are contributed by which volunteer, the data only become valuable in aggregate and so individual contributions are rarely highlighted. Participants have individual routes for their observations, but they all make the same *type* of observations. And yet a surprising proportion of participants expect to be acknowledged personally.

What this means for projects overall is that they need to be clear about their intention to acknowledge volunteers, and how they enable this in practice through data collection and licensing. This should, as a minimum, include a detailed privacy policy, outlining to contributors what data is collected, why, when, and what for, who owns the contributed data, and on which level of detail attribution will be given. One aspect of this discussion concerns the dynamic nature of CS datasets, which might be used or cited while still evolving; a problem related to the duration of projects, some of which run over several years, which Hunter & Hsu (2015) have discussed.

LIMITATIONS

Although we report on several case studies, the projects we present are limited in diversity, hence generalisability may be limited. While the four projects we analysed covered a wide range of topics and data types, and are broadly representative of many of the common activities in citizen science, they cannot represent the whole breadth and variety of citizen science engagement. We also only focused on projects with non-anonymous data collection, and projects with anonymous data collection are likely to face a different set of complications and challenges. Similar studies with other project types and in other scientific domains would be an interesting avenue

[14] https://five.epicollect.net/project/action-street-spectra/data

for future work. We also plan future work on the perspective of CS data users, rather than contributors and organisers.

Regarding our methodology, as with every survey, there is a self-selection bias, which excludes information about non-respondents. However, we received a high response rate and additionally triangulated our findings by using a mixed-methods approach. Nevertheless, we found that the interpretation of project documentation can be challenging, which we mitigate by conducting interviews with project organisers to clarify open questions from the document analysis.

CONCLUSION & RECOMMENDATIONS

Through the context of the ACTION project and its CS pilots, we explored the generation, use, and publication practices of CS project's metadata. We used a mixed-method approach combining insights from structured reviews of documentation, online surveys with contributors, and interviews with organisers and participants of CS projects, to generate insights into their metadata practices and perceptions. Our findings point to several weaknesses because of limited considerations of metadata, privacy risks and contributor acknowledgements. Our findings further show the importance of matching expectations between project contributors and project organisers regarding acknowledgement. They emphasise the importance of clear data processes and documentation in line with open science principles, to enhance transparency and facilitate data reuse (e.g. Burgess et al., 2017). Beyond this the findings also highlight the need to consider the expectations and mental models of users for their contributions; their internal explanations of how the project works and what their contribution is used for. This is relevant in relation to personal data and associated risks, for explicit and implicit contributions submitted by citizen scientists. This has so far been often overlooked in CS projects focused on the final data, rather than the process of creating the dataset.

We infer the following key considerations from our findings as recommendations for CS projects:

(i) explicit data and metadata contribution and associated risks;

Only if CS projects make their data and metadata collection procedures explicit, and flag potential harm to participants or others, can their contributors make informed decisions about how they contribute. This includes careful considerations of the platforms they use to collect data, and the implications these may have for their data and contributors.

(ii) implicit contributions and associated risks;

As CS projects collect data in different forms, they must highlight implicit, collateral metadata that is collected potentially without contributors being fully

aware, and flag any potential harm. This will allow citizen scientists to adapt their behaviour.

(iii) data licensing and acknowledgement schemes.

Projects need to lay out to their contributors from the very beginning how contributions they make are licensed, and what implications this has, for example with regards to CC-BY licences, requiring researchers to name contributors, vs. CC-0 licences that can be used without acknowledgement. Ideally, projects should give contributors different options of attribution, depending on preference and project output types.

All these concerns need to be considered at the project design stage rather than retrospectively, as they influence choice of tools or task setup, as well as how citizen scientists engage, and the long-term (re-)usability of the data the projects collect. CS projects, and those wanting to create or support such projects, should especially consider different expertise among their target groups - from professional researchers, through administrative staff, to inexperienced enthusiasts - and identify suitable formats and ways to communicate these details to all of them. If we want to use citizen science as a way to make science more accessible to all of society - including marginalised groups - and establish it as a research methodology, we need strategies to enable all of society to implement such projects and still deliver ethical and good quality data.

ACKNOWLEDGEMENTS

We thank all the study participants for their time. We thank UCM, DBC and BitLab for distributing the survey and participating in the organiser interviews. This study was supported by the ACTION project, which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 824603, and the IMPETUS project, which received funding from the Horizon Europe programme under grant agreement number 101058677.

APPENDIX 1: LIST OF DOCUMENTS

All documents have been collected on 5th August 2020; we cannot account for later changes to these documents.

STREET SPECTRA

Manual https://guaix.ucm.es/wp-content/uploads/2020/01/StreetSpectra_manual.pdf

What is Epicollect5 - Epicollect5 Data Collection User Guide https://docs.epicollect.net/

Add an entry - Epicollect5 Data Collection User Guide https://docs.epicollect.net/web-application/adding-data

Upload entries - Epicollect5 Data Collection User Guide https://docs.epicollect.net/mobile-application/upload-entries

Privacy Policy - Epicollect5 Data Collection User Guide https://docs.epicollect.net/about/privacy-policy

AZOTEA

Project website https://guaix.ucm.es/azoteaproject

English Manual (v2) https://zenodo.org/record/4680191

NOISE MAPS

Project Website (Spanish; analysed with Google Translate) http://www.bitlab.cat/projectes/noise-maps

Protocol for citizen science experiment_v1 (not public)

Project guide for participants (Spanish; analysed with Google Translate; not public)

Workshop slides (Spanish; analysed with Google Translate; not public)

BDM

App guide 2020_03 (Dutch; analysed with Google Translate) https://assets.vlinderstichting.nl/docs/f59bf0e9-ba74-441a-b60b-4763da820aa8.pdf

Manual online import Guide (Dutch; analysed with Google Translate) http://www.vlindernet.nl/doc/Handleiding_meetnetten.pdf

ACTION D2.3 Making a tutorial for water sampling dragonflies https://zenodo.org/record/4980410

ACTION D2.3 Tutorial for Water Sampling and Transect Selection https://zenodo.org/record/3885721

APPENDIX 2: SURVEY

About your engagement

1. Which citizen science project are you engaged in?
   - AZOTEA
   - Dutch Butterfly Conservation (Vlinderstichting)
   - NoiseMaps
   - Street Spectra
   - Other
2. What motivates you to take part in the project?

I support the goals of the project
I am interested in the research
I want to contribute to the research
I am interested in what I contribute specifically
I have a personal relationship to the project / team
I enjoy the competition with other participants
I am rewarded for my contributions

Not at all Somewhat Very much

3. What is your data contribution to the project?

Images
Observations
Sound files
Physical samples (e.g. specimen)
Comments
Analysis or interpretation
Other

4. What metadata do you contribute to the project?

Your name
Nickname / ID
Date / Time of contribution
(GPS) Location of contribution
Measurement specifications (e.g. type of sensor / camera)
Other

5. Is the location your home for any of your contributions?

Yes No I do not provide location data

6. Are you aware of what is done with the metadata you contribute?

1 (Not at all aware) 2 3 4 5 (Very much aware)

7. How important is it to you to know what is done with the metadata?

1 (Not at all important) 2 3 4 5 (Very important)

8. Do you feel your contribution poses a potential risk to

Your privacy
Other people's privacy
Your personal safety
Other people's safety
Your reputation
Other people's reputation
Other

Not at all Somewhat Very much

9. What is your expectation about what will happen with your contribution / the data you contribute?

Will be used for analysis
Will be used for (academic) publications
Will be used to influence policy

Will be used for campaigns (e.g. on social media)
Data will be published
Metadata will be published
Other

10. Do you expect your contribution to be acknowledged in project outputs (e.g. reports)?
    Yes No
11. How do you expect to be acknowledged?
    As an author
    By name, as a contributor
    By pseudonym, as a contributor
    As a volunteer in general (without explicit mention of yourself)
    Other
12. How important is it to you that your contribution is acknowledged?
    1 (Not at all important) 2 3 4 5 (Very important)
13. Do you expect to be notified of project outputs or results?
    Yes No

About you

14. How old are you?
    <18 18-25 26-30 31-35 36-40 41-45 46-50 51-55 56-60 61-65 66-70 70+
15. What is your gender?
    Female Male Non-binary Prefer not to say Other
16. What is your country of residence?
    Germany Netherlands Spain Other
17. What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.
    Primary education (School)
    High school graduate, diploma or equivalent
    Trade/technical/vocational training
    Bachelor's degree
    Master's degree
    Doctorate
18. Is there anything you would like to let us know?
    * free text *

REFERENCES

Arnstein, S.R., 1969. A Ladder Of Citizen Participation, Journal of the American Institute of Planners, 35(4): 216–224. DOI: https://doi.org/10.1080/01944366908977225.

Attig, J., Copeland, A. and Pelikan, M., 2004. Context and meaning: The challenges of metadata for a digital image library within the university. College & Research Libraries, 65(3), pp.251-261.

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

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059-1078.

Bowser, A., and Shanley, L. (2013). *New Visions in Citizen Science*. Woodrow Wilson Center.

Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method, 9(2), 27–40. Retrieved from https://doi.org/10.3316/QRJ0902027

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. Retrieved from https://doi.org/10.1145/2540032

Bowser, A., Shilton, K., Preece, J. and Warrick, E. (2017). February. 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* (pp. 2124-2136).

Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. European Journal for Philosophy of Science, 11, 1-26.

Burgess, H. K., DeBey, L. B., Froehlich, H. E., Schmidt, N., Theobald, E. J., Ettinger, A. K., HilleRisLambers, J., Tewksbury, J., Parrish, J. K. (2017). The science of citizen science: Exploring barriers to use as a primary research tool. Biological Conservation, 208, 113–120. Retrieved from https://doi.org/10.1016/j.biocon.2016.05.014

Christine, D.I. and Thinyane, M., 2021. Citizen science as a data-based practice: A consideration of data justice, Patterns, 2(4): 100224. DOI: https://doi.org/10.1016/j.patter.2021.100224.

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

DataCite Metadata Working Group. (2019). DataCite Metadata Schema documentation for the publication and citation of research data v4.3 [Application/pdf]. 73 pages. https://doi.org/10.14454/7XQ3-ZF69

Domhnaill, C.M., Lyons, S. and Nolan, A., 2020. The Citizens in Citizen Science: Demographic, Socioeconomic, and Health Characteristics of Biodiversity Recorders in Ireland, Citizen Science: Theory and Practice, 5(1): 16. DOI: https://doi.org/10.5334/cstp.283.

Erwig, M., Smeltzer, K., and Wang, X. (2017). What is a visual language? *Journal of Visual Languages & Computing*, 38, 9–17. https://doi.org/10.1016/j.jvlc.2016.10.005

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., and Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. Social Studies of Science, 41(5):667–690.

European Commission. (2014). Green Paper on Citizen Science. Citizen Science for Europe [Green Paper]. https://ec.europa.eu/digital-single-market/en/news/green-paper-citizen-science-europe-towards-society-empowered-citizens-and-enhanced-research

Fraisl, D., Campbell, J., See, L., Wehn, U., Wardlaw, J., Gold, M., and Fritz, S. (2020). Mapping citizen science contributions to the UN sustainable development goals. Sustainability Science. Retrieved 9 September 2020 from https://doi.org/10.1007/s11625-020-00833-7

Hecker, S., Wicke, N., Haklay, M., and Bonn, A. (2019). How Does Policy Conceptualise Citizen Science? A Qualitative Content Analysis of International Policy Documents. Citizen Science: Theory and Practice, 4(1), 32. Retrieved from https://doi.org/10.5334/cstp.230

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), pp.454-466.

Hunter, J. and Hsu, C.H. (2015). December. Formal acknowledgement of citizen scientists' contributions via dynamic data citations. In International Conference on Asian Digital Libraries (pp. 64-75). Springer, Cham.

Koesten, L., Simperl, E., Blount, T., Kacprzak, E. and Tennison, J., 2020. Everything you always wanted to know about a dataset: Studies in data summarisation. International Journal of Human-Computer Studies, 135, p.102367.

Lagoze, C. (2014). eBird: Curating Citizen Science Data for Use by Diverse Communities. International Journal of Digital Curation, 9(1), 71–82. Retrieved from https://doi.org/10.2218/ijdc.v9i1.302

Levin, N., and Leonelli, S. (2017). How does one "open" science? Questions of value in biological research. Science, Technology, & Human Values, 42(2), 280-305. Retrieved from https://doi.org/10.1177/0162243916672071

Leonelli, S. (2020). Scientific research and big data. The Stanford Encyclopedia of Philosophy

Mayernik, M. (2011). Metadata realities for cyberinfrastructure: Data authors as metadata creators.

Miyashita, E.-F., Pernat, N., and König, H. J. (2021). Citizen science as a bottom-up approach to address human–wildlife conflicts: From theories and methods to practical implications. Conservation Science and Practice, 3(3), e385. https://doi.org/10.1111/csp2.385

Neff, G., Tanweer, A., Fiore-Gartland, B., & Osburn, L. (2017). Critique and contribute: A practice-based framework for improving critical data studies and data science. Big data, 5(2), 85-97.

OED Oxford English Dictionary, second edition (1989). Oxford University Press 1989. Retrieved June 2023 from https://www.oed.com/oed2/00057804;jsessionid=F1B19E9B19C4EA7FB6B0CE2A68B04E4B

Oudheusden, M. V., and Abe, Y. (2021). Beyond the Grassroots: Two Trajectories of "Citizen Sciencization" in Environmental Governance. Citizen Science: Theory and Practice, 6(1), 13. https://doi.org/10.5334/cstp.377

Ottinger, G. (2010). Buckets of resistance: Standards and the effectiveness of citizen science. Science, Technology, & Human Values, 35(2), pp.244-270.

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

Ponti, M., and Craglia, M. (2020). *Citizen-generated data for public policy. A brief review of European citizen-generated data projects.*

Reeves, N., Tinati, R., Zerr, S., Van Kleek, M. G., and Simperl, E. (2017). From Crowd to Community: A Survey of Online Community Features in Citizen Science Projects. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2137–2152. https://doi.org/10.1145/2998181.2998302

Resnik, D.B., Elliott, K.C. and Miller, A.K. (2015). A framework for addressing ethical issues in citizen science. Environmental Science & Policy, 54, pp.475-481.

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

Robinson, L. D., Cawthray, jade L., West, S. E., Bonn, A., and Ansine, J. (2018). Ten principles of citizen science. In Citizen Science: Innovation in Open Science, Society and Policy. UCL Press. Retrieved 3 March 2020 from https://doi.org/10.2307/j.ctv550cf2

Roman, D., Reeves, N., Gonzalez E., Celino I., Abd El Kader, S., Turk, P., Soylu, A., Corcho, O., Cedazo, R., Gloria Re Calegari, Damiano Scandolari, Elena Simperl (2020). Preprint: An analysis of pollution citizen science projects from the perspective of data science and open science. *Submitted to: Citizen Science: Theory and Practice*

Scassa, T., and Chung, H. (2015). Typology of citizen science projects from an intellectual property perspective: Invention and Authorship Between Researchers and Participants (Policy Memo Series).

Thatcher, J., O'Sullivan, D., and Mahmoudi, D. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data. Environment and Planning D: Society and Space, 34(6), 990–1006. https://doi.org/10.1177/0263775816633195

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.

Thuermer, G., Guardia, E.G., 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. DOI: https://doi.org/10.5334/cstp.538.

Wang, Y., Kaplan, N., Newman, G. and Scarpino, R. (2015). CitSci. org: A new model for managing, documenting, and sharing citizen science data. PLoS Biol, 13(10), p.e1002280.

White, S.C., 1996. Depoliticising development: The uses and abuses of participation, Development in Practice, 6(1): 6–15. DOI: https://doi.org/10.1080/0961452961000157564.

Wong, J., Henderson, T., & Ball, K. (2022). Data protection for the common good: Developing a framework for a data protection-focused data commons. *Data & Policy*, *4*. https://doi.org/10.1017/dap.2021.40

Xia, H., Wang, Y., Huang, Y., and Shah, A. (2017). 'Our Privacy Needs to be Protected at All Costs': Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–22. Retrieved from https://doi.org/10.1145/3134748