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*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/DTA-10-2020-0253

An analysis of pollution Citizen Science projects from the perspective of Data Science and Open Science

Dumitru Roman[1,*], Neal Reeves[2], Esteban Gonzalez[3], Irene Celino[4], Shady Abd El Kader[1], Philip Turk[1], Ahmet Soylu[5], Óscar Corcho[3], Raquel Cedazo[3], Gloria Re Calegari[4], Damiano Scandolari[4], Elena Simperl[2]

[1]SINTEF AS, Norway
{dumitru.roman, philip.turk, s.abdelkader}@sintef.no
[2]King's College London, United Kingdom
{neal.reeves, elena.simperl}@kcl.ac.uk
[3]Universidad Politécnica de Madrid, Spain
{egonzalez, ocorcho, raquel.cedazo}@fi.upm.es
[4]Cefriel, Italy
{irene.celino, gloria.re, damiano.scandolari}@cefriel.it
[5]OsloMet – Oslo Metropolitan University, Norway
ahmetsoy@oslomet.no

[*]Contact author, email: dumitru.roman@sintef.no

Abstract.

Purpose: Citizen Science – public participation in scientific projects – is becoming a global practice engaging volunteer participants, often non-scientists, with scientific research. Citizen Science is facing major challenges, such as quality and consistency, to reap open the full potential of its outputs and outcomes, including data, software, and results. In this context, the principles put forth by Data Science and Open Science domains are essential for alleviating these challenges, which have been addressed at length in these domains. The purpose of this study is to explore the extent to which Citizen Science initiatives capitalise on Data Science and Open Science principles.

Approach: We analysed 48 Citizen Science projects related to pollution and its effects. We compared each project against a set of Data Science and Open Science indicators, exploring how each project defines, collects, analyses and exploits data to present results and contribute to knowledge.

This is the author accepted version of an article published in
Data Technologies and Applications
http://dx.doi.org/10.1108/DTA-10-2020-0253

Findings: The results indicate several shortcomings with respect to commonly accepted Data Science principles, including lack of a clear definition of research problems and limited description of data management and analysis processes, and Open Science principles, including lack of the necessary contextual information for reusing project outcomes.

Originality: In the light of this analysis, we provide a set of guidelines and recommendations for better adoption of Data Science and Open Science principles in Citizen Science projects, and introduce a software tool to support this adoption, with a focus on preparation of data management plans in Citizen Science projects.

Keywords: Citizen Science, Data Science, Open Science, pollution projects, Data Management Plan, software

1. Introduction

Citizen Science (CS) describes the active engagement of volunteer participants within scientific research. Projects vary greatly in terms of the role that volunteers play and the degree of agency that they have, from more passive models where volunteers install software and sensors to more collaborative models where volunteers actively define problem spaces and research topics (Haklay, 2013). Nevertheless, CS commonly entails the gathering of data by volunteers for later dissemination and publication (Haklay, 2013; Pocock *et al.*, 2017).

CS data are of significant value not only in the projects in which they are gathered, but for subsequent analysis and re-use (Wang *et al.*, 2015). Volunteer-contributed data may complement and offer the opportunity to contextualise and expand upon existing monitoring efforts by professional organisations, without significant added cost (Hadj-Hammou *et al.*, 2017). In some contexts, the vast majority of available data are from CS sources (Groom *et al.*, 2017; Poisson *et al.*, 2020). This is equally true of software, with CS projects developing a wide variety of software programs which are of value not only to professional scientists, but also to lay people without the technical or scientific knowledge to develop tools of their own (Cooper *et al.*, 2018; Zaman *et al.*, 2020).

Despite the effectiveness of CS approaches, the impact of such projects remains limited. CS initiatives are often slow to publish their results – if indeed the results are published at all (Kullenberg and Kasperowski, 2016; Thoebald *et al.*, 2015). Data and conclusions from CS initiatives are under utilised in public policy spaces and small-scale projects aiming to have more immediate effects on local issues tend to be volunteer-led with little to no input from domain experts (Nascimento *et al.*, 2018; Wiggins and Crowston, 2011). Any use of Citizen Science data is also complicated by the ease with which any individual can set up such an initiative and the resulting lack of consistency and rigor that this entails and CS is no exception (Ponti and Craglia, 2020). Careful consideration is required throughout the research process, to design and consider how volunteer-generated data are to be gathered, evaluated and integrated within the scientific

workflow (McKinley *et al.*, 2017). Interoperability and integration concerns also heavily impact CS data and software, particularly for non-expert users (Simonis, 2018; Zaman *et al.*, 2020).

One possible solution to address issues of quality and consistency and to add greater value to CS and its outputs is adoption of principles from *Data Science (DS)* and *Open Science (OS)*. DS is an emergent interdisciplinary approach combining statistical analysis, computing and social sciences to develop methodologies and derive outputs, insights and decision making directly from data (Cao, 2017). Expertise and literacy surrounding DS has been identified as a crucial component of successful CS projects both within project management teams and participating volunteers, with the potential to improve the quality of data and outputs (Rambonnet *et al.*, 2019; Sagy *et al.*, 2019). Indeed, greater DS competency is essential if participants are to go beyond simple data collection or analysis exercises and conduct *authentic* scientific research (Wiggins and Wilbanks, 2019). Yet if such scientific and information literacy is to be fostered, it is essential that audiences are able to access and view scientific research and its outputs (Lana, 2019). These are the aims of OS, which cover not only the research itself, but also the tools, software and methods used to generate findings (Groom *et al.*, 2017; Wilkinson *et al.*, 2016).

In this paper, we present an analysis of CS projects with a focus on accepted principles within the domains of DS and OS. In sampling projects for analysis, we focus on projects related to pollution and its effects, drawn from two popular databases of CS projects – the *SciStarter* platform and Wikipedia's *List of Citizen Science Projects*. CS data are vital for pollution monitoring, representing not only a significant source of data, but also an opportunity for longitudinal monitoring (Poisson *et al.*, 2020). After devising a sample of 48 projects, we then compare each project against a set of DS and OS indicators, analysing how each project defines, collects, analyses and exploits data to present results and contribute to knowledge. Our findings explore the extent to which DS and OS principles are currently accounted for within CS, while identifying the steps and processes projects use to manage and ensure the quality of their data.

Finally, we present guidelines and recommendations for greater adoption of DS and OS principles within CS, before presenting a data management tool to support those with less experience of scientific research in conducting CS.

2. Related Work

2.1. Pollution Citizen Science

Pollution and environmental monitoring initiatives are a highly common focus of CS initiatives. An analysis of CS projects and outputs on the Web of Science platform as carried out by Kullenberg and Kasperowski (2016) identified geographic identified environmental monitoring as among the largest of CS topics, with environmental science being the second most commonly occurring discipline aligned with projects, second only to ecology. The adoption of CS within this area has been driven in part by opportunities associated with technological advances in recent years. The availability of low-cost and easily accessible sensors, as well as smartphones, has significantly

increased the quantity of data that can be gathered (Fishbain *et al.*, 2017). Secondly and as a result, CS can be employed to greatly increase the area over which data can be gathered, particularly in otherwise dangerous or inaccessible areas, at significantly lower cost (Paul and Buytaert, 2018; Rambonnet *et al.*, 2019).

Perhaps unsurprisingly, citizen-generated environmental monitoring data are increasingly being integrated with or considered alongside expert-generated data (Hadj-Hammou *et al.*, 2017). However, significant barriers remain to the adoption of such data. While integration of expert and non-expert gathered data sources has the potential to significantly increase data quality, the accuracy of data gathered from volunteers tends to be unknown and volunteers' understanding of data gathering methods conflict greatly with those of professional scientists (Jiang *et al.*, 2018). At the same time, as the availability of tools such as sensors has increased, so too has variety and researchers may be unwilling to employ data with such levels of uncertainty when used to highly calibrated instruments and research practices (Budde *et al.*, 2017).

Despite this, it is important to highlight that the adoption of CS in this area has not only been driven by research aims. Kullenberg and Kasperowski (2016) note that a large proportion of pollution monitoring projects have no scientific publications and instead were launched in response to citizen concerns around pollution effects on health. In some monitoring contexts–particularly air quality monitoring–projects have used CS not only as a source of data, but as a means to educate and raise awareness among citizen volunteers of the dangers posed by pollutants (Mahajan *et al.*, 2020). This capacity to create lasting changes to opinions and behaviours has been identified as one of the most powerful advantages of pollution-related CS (Forrest *et al.*, 2019). Similarly, citizen monitoring of fracking has been used to encourage policy change and to encourage civic participation in both science and local governance (Zilliox and Smith, 2018). These efforts are, however, strongly linked with and driven by data gathering efforts and are therefore subject to the same data concerns as more research-oriented efforts within pollution monitoring.

2.2. Data management in Citizen Science

Existing solutions to issues of data quality in Citizen Science are largely project- and domain-specific, but largely focus on gathering increased volumes of data, through additional evidence or approaches such as redundancy (Wiggins and He, 2016). It is important to note that data management and sharing is only one part of open science and open access – as Borgman (2015) notes, scholarly and scientific openness encompasses broader concepts such as technology, literature, communication and the dissemination of ideas, which in turn may feed into and become data. Nevertheless, there is a growing recognition of the importance of data and open science principles for the effective uptake of CS data. Bowser *et al.* (2020), conducted semi-structured interviews with representatives of 36 CS projects, to understand how the projects address key stages of the research lifecycle. The analysis identified a number of weaknesses within the sampled projects, including open access, documentation, interoperability and sustainability. Groom *et al.* (2017) analysed species monitoring datasets within the Global Biodiversity Information Facility, finding that while CS datasets made up the majority of the data sources, they

were significantly less openly accessible than commercial datasets and those provided by research institutions.

More broadly, an analysis of CS projects conducted by Schade and Tsinaraki (2016) found that over 50% of all projects had no data management plan in place, causing subsequent difficulties throughout the data analysis process. Similar issues were noted by Mckinney *et al.* (2017), who noted that data management plans are crucial to the effective and successful use of CS methods, particularly for subsequent acceptance and use of CS methodologies and data by professional scientists and other stakeholders.

Furthermore, issues stemming from improper planning and data management may arise throughout the lifespan of projects. Stakeholders storing and analysing data must be sure to record and track the provenance of – and transformations to – data if they are to be reusable by other stakeholders and this is a distinct challenge associated with CS (Tiufiakov *et al.*, 2018; Williams *et al.*, 2018). In a discussion of open data and data sharing practices among large-scale scientific projects, Nielsen (2011) noted the inherent difficulties associated with sharing data whilst ensuring it is sufficiently documented and contextualised to be useful to subsequent end users, whose needs may often be unknown. Nielsen further noted that while political pressures around funding often prompt larger projects to share their data, smaller projects are often not subject to the same expectations. Musto and Dahanayake (2020) analysed 38 CS platforms and identified a significant lack in the tracking and management of data provenance, including in platforms which supported the design and launch of subsequent projects. Anhalt-Depies *et al.* (2019) addressed similar concerns regarding data privacy and trust in the Snapshot Wisconsin project, developing recommendations and best practices for balancing data quality with issues of trust and the reusability of data resources. In contrast to these previous studies, we focus explicitly on the domain of pollution, looking at a larger selection of 48 projects and focusing on the FAIR principles to evaluate data management processes.

We additionally note a number of existing efforts that aim to support data management throughout the research process. Wang *et al.* (2015) developed CitSci.org, a web-based model that guides participants through the entire data analysis process aiming to enhance discoverability and reusability of research. Despite this, the platform assumes that some initial considerations have taken place to conceptualise and define the dynamics of the project and the platform mainly supports data collection and subsequent analysis (Gray *et al.*, 2017). Similarly, DataONE is a set of web-based resources to support data management and secure the provenance of scientific research data in the earth and environmental sciences (Cao *et al.* 2016; Mckinney *et al.*, 2017). While the service provides both tools and guidance, the underlying technology behind the platform is complex, technical and assumes datasets have already been gathered, making it less well suited for stakeholders who are not professional researchers with existing data collection and management knowledge (Cao *et al.*, 2016). In contrast to these platforms, we aim to develop a set of guidelines which support projects at the earliest, conceptual stages. These guidelines support CS researchers and citizen scientists to define and consider issues of data quality, as well as proposed data collection, analysis and dissemination efforts. In contrast to existing

methods, we do not assume any specialist knowledge of scientific research, data processing methods or software.

3. Methodology

3.1. Selection of pollution Citizen Science projects

The selection of CS projects for analysis was based on two primary sources - the most comprehensive catalogues of CS projects to date:

- SciStarter (https://scistarter.org) – a platform that connects volunteers with CS projects, listing more than 1200 projects. The organization's primary goal is to break down barriers preventing non-scientists from fully engaging in scientific research.
- Wikipedia's *List of Citizen Science Projects* (https://en.wikipedia.org/wiki/List_of_citizen_science_projects) lists approximately 300 projects. Some of these are completed or retired projects and there is also a degree of overlap with projects present in SciStarter.

It is worth noting that the two sources are not completely up-to-date, especially the Wikipedia list, so during the selection of relevant projects many of them have been discarded because their websites were not reachable or there was not enough information about them.

For the selection of the projects that primarily deal with pollution, we applied a filter on the pollution topic. The projects of interest were the projects that are built around the problem of "pollution" which could be, e.g., water, air or ground pollution. On the Wikipedia list we filtered the projects by the attribute "discipline", looking for those projects that have pollution or air quality or water quality in that field – in this way we reduced the projects from approximately 300 to 20 projects of interest. On SciStarter we applied a filter on the projects keeping only those which had pollution in the title, in the description or in one of the keywords – in this way starting from more than 1200 projects we reduced the projects of interest from SciStarter to 71. After removing the overlapping projects between Wikipedia and SciStarter we applied another filter, removing projects where pollution is not relevant in the analysis (e.g., https://scistarter.org/platypuswatch-gold-coast-2), projects expired with no data (e.g., https://scistarter.org/cyber-citizen) and duplicates. Also, for projects with the same domain but in different locations we kept only one of them (e.g., https://www.curio.xyz/explore/missions/67 and https://www.curio.xyz/explore/missions/66. At the end of this process and after merging the two lists (Wikipedia and SciStarter), the number of selected projects was 48 (the list of projects and the collected data for each of them is available via https://doi.org/10.5281/zenodo.3958853).

3.2. Approach

We take a Data- and Open-Science centric approach in analyzing CS projects, with the primary focus to identify to what degree pollution-related CS projects follow the DS principles/phases for the management and analysis of data, and to what degree they follow the OS principles/phases for publication of results.

3.2.1. Data Science perspective

From a DS perspective, we analyse the projects based on the main phases of a typical DS project:

1. *Problem definition:* In this phase, data scientists *identify a problem* or a subject of interest, and then by analysing the identified problem they *formulate a hypothesis* or a goal. Data scientists then plan the *experiment design* in order to solve the initial problem.
2. *Data management:* In this phase, data scientists focus on *data collection, preparation and storage* activities, with these activities influencing each other reciprocally. Choosing an unsuitable *data storage* tool based on the data available can lead to difficulties in further analysis (e.g., a good data management system for streaming data might not be good for static data). This must relate to the *data preparation* phase which is the process of cleaning and transforming raw data prior to analysis and often involves reformatting data, making corrections and combining data sets to enrich data. The data management phase is the most time consuming one.
3. *Hypothesis (HP) testing:* In this phase, data scientists apply *statistical evidence/data analysis/machine learning algorithms* in order to retrieve interesting information towards the initial hypothesis from the initial data. The possible techniques in this step are various and depend both on the data and on the hypothesis (e.g., if we want to understand what increases air pollution in a metropolitan area, a statistical model can be more useful than a predictive machine learning algorithm).
4. *Result evaluation:* In this phase, data scientists can extract *numerical values* that support or reject the initial hypothesis. Scientists usually write a report on the project with their *interpretation of results*, create *data visualizations* to communicate results to stakeholders and they *publish results* expanding the general level of knowledge on the matter.

3.2.2. Open Science perspective

The objective of this study from the OS perspective was to analyze if CS projects publish their outcomes as professional scientific research (using persistent identifiers to be cited, adding metadata to reference authors or dates, using domain specific/general repositories with the correspondent license) or more amateur. We analysed the projects following the Open Data FAIR principles (Wilkinson *et al.*, 2016):

1. *Findable.* Metadata is used to describe the data and facilitate its discovery. A unique persistent identifier has to be generated.

2. *Accessible.* Data has to use appropriate metadata to facilitate its use by humans and machines. Data should be deposited in trusted repositories.
3. *Interoperable.* To make data interoperable with other datasets, the data must use a vocabulary or an ontology.
4. *Reusable.* In order to increment the use of data, a license is needed.

Although FAIR principles are designed for data, we have extended them to the software domain similar to what was done in Lamprecht *et al.* (2019). Together with data and software, we have included more indicators in a third category related to other results such as number of scientific publications, media channels, etc. Also, we consider if all this information is open and transparent to the community for CS projects.

3.3. Analysis dimensions

3.3.1. Data Science dimensions

The analysis of the projects from the Data Science perspective follows the analysis of the main phases of a typical DS project, aiming to answer the following questions about the CS projects:

- *Problem identification:* Does the CS project deliver to citizen scientists a clear definition of the problem they should aim to work on?
- *HP formulation:* Does the project deliver a clear hypothesis on the problem in order to solve it through the help of citizen scientists?
- *Data collection:* How does the project perform the acquisition of initial/raw data?
- *Data preparation:* How does the project perform the process of cleaning and transforming the initial/raw data?
- *Data storage:* Where/how does the project store the data?
- *Data mining:* Which data mining activities are performed on the data?
- *Machine learning:* Which machine learning algorithms are performed on the data?
- *Statistical evidence:* Which activities are performed to extract statistical evidence on the data?
- *Publication of results:* How are the results published?

3.3.2. Open Science dimensions

We defined a set of indicators inspired in the FAIR Open Data principles and we grouped them into three categories based on the nature of the resource: *data*, *software* and *other results*.

The *data* indicators include:

- *Persistent identifier:* Is there a PID (Persistent Identifier) assigned to the dataset? This indicator shows if a persistent identifier or DOI was generated.
- *Metadata fields:* What metadata are used to describe the data? This indicator indicates the metadata fields filled in to describe the data.

- *Software related:* Is there any reference to the software used to produce these data?
- *Use of a public repository:* Is this data deposited on a public repository? This indicator indicates if the data generated in the project is published on a general public repository like Zenodo, in a domain-specific repository or is available on a website.
- *Openness of the data:* Is the data accessible without an authentication process?
- *Use of vocabulary:* Is a vocabulary/ontology used to describe the data or the metadata? For interoperability, it is important that data is defined by a specific vocabulary or ontology.
- *License used:* What type of license is used?

The *software* indicators include:

- *Persistent identifier:* Is there a PID assigned to the software?
- *Metadata fields:* What metadata are used to describe the software?
- *Data related:* Does this software generate/use data?
- *Use of a public repository:* Is this software registered on a public repository?
- *Openness of the software:* Is the software accessible without an authentication process?
- *Use of a vocabulary:* Is a vocabulary/ontology used to describe the software?
- *License used:* What type of license is used?
- *Software deployment:* Are there instructions to deploy the software? This indicator is related to the reproducibility of the project. It is important that the software is documented to deploy it.
- *Container technology:* Does the software use any specific container technology? Related with the previous question, if the creator has used the technology docker, it can facilitate the deployment of the system.
- *Reproducibility:* Is it present on a reproducibility platform such as CodeOcean (https://codeocean.com)?
- *Machine reading:* If the project provides an API, is it documented?

The *other results* indicators include:

- *Scientific publications:* If there is a scientific publication related with the project, what is its license?
- *Communication channels:* What communication channels does this project use?
- *Hardware:* If the project has developed a device (hardware), is its design openly published? What is the license used?

It should be noted that there is a partial overlap between the processes involved in DS and the principles that govern OS, particularly in terms of data storage and the dissemination of results. This is largely due to the complimentary recommendations that the two approaches provide: DS dictates that data should be stored and disseminated in a logical, systematic and scientific manner, while OS dictates that data must be stored in a transparent and openly accessible way and that such data and any resulting research outputs should be freely and openly disseminated. We consider both approaches and their implications for the surveyed projects separately, before synthesising our findings.

4. Results

4.1. Data Science results

We first sought to analyse the 48 sampled projects from a DS perspective. To achieve this, we divided the data collection and analysis process into distinct stages and analysed the extent to which each project provided evidence or details of how each stage was achieved.

4.1.1. Problem definition

We first analysed whether projects had a clearly defined problem which they sought to solve by gathering data or metadata from citizen volunteers. Within the 48 sampled projects, approximately half – 25 of 48 – included a statement or explanation clearly explaining the issue that the project aimed to resolve or improve, be that expanding research or combatting pollution in a given area. The clarity and specificity of these problems varied strongly, with the most clearly defined problems detailing very specific issues to be addressed – e.g., combating water pollution stemming from construction-induced erosion. Projects with more nebulous or abstract problems stemming from a pollution topic more broadly tended to have less specific goals and objectives.

However, almost half of all projects had no clearly defined problem as a starting point. Projects within this category included monitoring projects aiming to identify whether any pollution-related problems had arisen in a particular context, but also a group of very specific projects with highly specialised locations and data collection goals, seemingly designed in response to particular research issues. This suggests the possibility that while projects may stem from a particular problem, this may not be clearly communicated to potential volunteers and other stakeholders within project materials, if indeed these problems are communicated at all.

4.1.2. Hypothesis Formulation

To understand data collection and analysis processes, we examined whether each of the projects defined hypotheses for analysis, either through the presentation of the hypotheses themselves or through a description of the aims of the project. However, just three projects included some form of hypothesis formulation and none presented specific hypotheses that the project aimed to examine. Perhaps the clearest example of a hypothesis within the projects was the Loss of the Night project (data available through My Sky At Night: http://www.myskyatnight.com/#map) which noted an ongoing debate about whether LED streetlights would make the night sky brighter or darker, but did not explicitly note that the project aimed to explore these issues or identify specific associated hypotheses.

4.1.3. Data Collection

Data collection was the process common to the largest proportion of projects, with 45 of the 48 projects asking volunteers to gather data in some form (see Table 1). The specific data to be gathered and submitted by volunteers varied between projects, with many projects requiring

submission of multiple complementary sources of data. Photographs were the most common data type used within projects, with 14 of 48 (25%) of projects requiring photos, while 10 projects used sensors and physical samples and a further nine used simple, repeatable surveys to gather data.

Table 1 Aggregated count of data types collected in the sampled projects.

Collection	**Number of projects**
Pictures	14
Samples	10
Sensors	10
Survey	9
Record Data	4
Photo Description	4
Measurements	3
Task	3
Volunteers' Analysis	2
Communication	1
Hosting Sensors	1
Information Provision	1
Physical Collection	1

4.1.4. Data Preparation and Analysis

Data preparation is an essential process in any form of CS, where raw data gathered by citizens must be validated and any false data rectified or removed. If collected data are to be made publicly available in a usable format, then it is equally essential that this process is clearly documented, such that the providence of the data and validity of both the data and any conclusions drawn from them can be identified by distinct stakeholders. Despite this, the vast majority of the sampled projects offered little to no information regarding steps taken to prepare and validate any gathered data. 37 projects failed to mention data pre-processing, while 3 projects explicitly stated that no such processing occurred. Even within those projects where such processing took place, 4 projects failed to explain *how* such processing was achieved, leaving just 4 projects where validation was documented at all. In two of the projects validation was carried out by project scientists using lab-based chemical analysis, while a further project carried out expert validation

using images provided by volunteers. Finally, one project involved validation by citizens, using a majority voting mechanism by which aggregated results were validated through agreement.

We subsequently examined projects for the presence of provenance metadata documenting any preparation and processing. A total of 3 projects were found to store some form of versioning metadata – each of these projects used Zenodo to store their datasets and metadata were generated automatically. In all other cases, projects did not provide any such metadata or any human- or machine-readable descriptions of the relationships between datasets.

Furthermore, project documentation offered no information about the data analysis process used to produce and validate conclusions. This kind of analysis could only be found (in rare cases) in external documents such as scientific papers, theses, or annual reports. Initially, this may not appear overly surprising, given that analyses are inherently linked with specific research questions. However, only 19 of the 48 projects shared on their website at least one report or scientific publication to communicate the conclusions drawn. Such reports had a great variety in terms of quality, type, and conclusions, suggesting a lack of general guidelines in producing them, and the fact that they are often produced by external researchers rather than CS project owners.

4.1.5. Data storage

As a final step, we analysed whether projects stored their data in a publicly accessible manner, such that diverse stakeholders would be able to find, access and make use of the data without the need to request assistance from project scientists. We noted that twenty of the projects made all – or at least some – of their data available in an open and public manner, while a further 2 described the opportunity to access data upon request, although there was little indication of who could or should request the data and whether such requests would be granted.

4.2. Open Science results

Subsequently, we analysed whether projects adhered to commonly accepted OS principles for project outputs. Based on our analysis of the sampled projects, we identified three key categories for project outputs: *data* gathered either from volunteers or produced through pre-processing and analysis; *software* as well as associated documentation such as code and usage instructions and finally *other results*, including any scientific publications, social media dissemination and hardware produced by the projects.

4.2.1. Data

Despite gathering and analysing data being a central aim of all of the sampled projects (as discussed above in the case of the DS analysis), there was significant variation in the availability and format of the associated resources that each project made available. Generally, the data were made available through project websites, with three projects using the public repository Zenodo and one using the Anecdata repository.

This lack of use of public repositories was associated with a number of significant outcomes for the usability of the data. Only three projects had a persistent identifier to point to the underlying datasets, which in all three cases were automatically generated through Zenodo. In all other cases, the only material to point to the dataset was a URL and there was nothing to suggest that these URLs would remain static or even available should the underlying data change, be updated or removed. Similarly, only these three projects had metadata fields to indicate key details such as versions, version history and the date that the dataset was uploaded and edited, which were missing from the remaining datasets. Arguably most significantly was the lack of licenses specified along with the datasets, with only two projects with machine readable licenses. Again, both of these licenses were specified within Zenodo. In all other cases, insufficient details were offered to adequately identify which – if any – conditions were to be imposed on those seeking to reuse the data and for which purposes such reuse would be permitted.

More broadly, we note two additional details which harmed the value of the data for subsequent use. Only one project offered explicit detail of the software and hardware used to generate the data and none of the sampled datasets included an ontology or specified vocabulary to describe the data or how fields related to one another.

Table 2 shows the results for the *data* category.

Table 2 FAIR indicators for *data* with project counts and descriptions.

Indicator	**Number of projects**
Persistent identifier	3
Metadata fields	3 (Publication date, license, communities, version)
Software related	1
Use of a public repository	1 (in Anecdata) 3 (in Zenodo)
Openness of the data	4
Use of vocabulary	0
License used	2 (CC-BY-4.0)

4.2.2. Software

Similarly, the accessibility of software materials was highly variable and generally hampered reuse. Although a total of 17 projects had some form of software available – each with instructions to deploy the software for subsequent uses – there were significant details missing from each of the projects. None of these 17 projects used container technologies to package or deliver the software, increasing the effort required to initially set up the software. Similarly, none of the

projects allowed the software to be accessed without authentication and there were no options to authenticate new users without input and action from the project teams. Moreover, we found no persistent identifiers assigned to software and no metadata fields or pre-defined vocabularies to allow for easy identification of datasets. Finally, none of the software programs were open source and perhaps because of this, none of the available software included any details or instructions to facilitate reproducibility.

Table 3 shows the results for the *software* category.

Table 3 FAIR indicators for *software* with project counts and descriptions.

Indicator	**Number of projects**
Persistent identifier	0
Metadata fields	0
Data related	2
Use of a public repository	2 (in Anecdata) 3 (in SciStarter) 1 (in GitHub)
Openness of the software	0
Use of a vocabulary	0
License used	1 (open source)
Software deployment.	17
Container technology	0
Reproducibility	0
Machine reading	5

4.2.3. Other results

We further explored each project's web presence to identify other forms of outputs and noted three forms: i) scientific publications, ii) articles and iii) communication through social media and other channels. Only four of the projects mentioned any form of scientific publication, although we note that the lack of licensing and attribution details within datasets means that we cannot reliably state whether data had been published outside of projects. Additionally, given our choice of methodology we cannot rule out that additional publications exist, either under peer review or in a published form that is not mentioned in the public facing elements of projects.

Taking a broader view of publications, 40% of the projects had some form of published report describing either results of the project or the data gathered by volunteers. The number of reports published by projects ranged from just one report to over one hundred individual and distinct publications in the case of seagrass watch (https://www.seagrasswatch.org), which has published a significant range of scientific and miscellaneous publications relating to the seagrass using findings from the project. Nevertheless, most projects had published significantly fewer reports, with a mean average of 8.67 and a median average of just 3 reports per project. Moreover, in the majority of projects, no report documentation was available, whether scientific publications or otherwise.

Beyond this, we note three key forms of output and communication for projects. Websites and an associated web presence were important for many of the sampled projects, with 36 projects possessing a public facing website, five using a SciStarter page in lieu of a website and two using both types. 19 of the projects used social networks such as Facebook and Twitter. Links between different channels were often inconsistently presented and, in some cases, lacking entirely. While we also considered the possibility that projects may have hardware designs for dissemination, such as pollution sensors, we found no evidence for this in the surveyed projects.

Table 4 shows the results for the *other results* category.

Table 4 Miscellaneous FAIR indicators with project counts and descriptions.

Indicator	**Number of projects**
Scientific publications.	4 projects
Communication channels	19 projects use social networks 7 are indexed in SciStarter 36 projects have a website

5. Discussions and Recommendations

5.1. Implications of results

Our findings noted a number of significant weaknesses throughout the DS process, without clear definition and formulation of research problems and questions and with limited description of data management and analysis processes. On the one hand, this is likely a result of our chosen methodology, as given the availability of reports and preliminary visualisations across projects, it can be assumed that projects have at least some data analysis processes and aim to resolve specific research problem(s). Conversely, in all but four projects no scientific research publications were found, suggesting that these processes are not publicly documented and therefore that these issues pose significant difficulties for the reuse of data and project outputs.

As a research methodology reliant on participation of non-expert volunteers, CS research is particularly vulnerable to concerns of rigor and reliability. There is growing recognition of issues arising from a lack of reproducibility in academic research and with potentially lax data validation processes, there is little to prevent malicious or careless volunteers from submitting large numbers of invalid results (Rasmussen, 2019; Elliott and Rosenberg, 2019). Yet within the sampled projects, we observed the open availability of data without coherent and structured reporting of the processing carried out to gather, prepare and validate this data prior to release. Such a lack of clearly defined processes and documentation greatly exacerbates these concerns. Furthermore, the lack of clear metadata such as version histories and upload dates, along with the lack of persistent indicators such as DOIs gives way to queries of how results using such potential low quality data could be identified and modified should such issues arise.

Moreover, while we identified a significant number of potential data and software resources within the sampled projects, our analysis suggests that projects lack the necessary contextual information to allow reuse of such data for research purposes. If researchers are to be able to utilise CS data, it is essential that they are able to understand the context in which any datasets were gathered, produced and made available and how they may be used in future (Williams *et al.*, 2018). Within the surveyed projects, we found this contextual information to be significantly lacking, likely severely limiting the impact of the provided datasets – all but two of the projects did not include the necessary licensing details to specify whether reuse of such data would be permitted. This is notably a common issue with Citizen Science datasets as noted by Groom *et al.* (2017). Nevertheless, it should be noted that this poses particularly significant issues in areas such as environmental monitoring, where significant proportions of data are sourced from volunteer-generated datasets and where volunteer contributions may be crucial to fill gaps in existing monitoring programmes led by government organisations (Hadj-Hammou *et al.*, 2017; Poisson *et al.*, 2020).

Additionally, such barriers to reuse significantly harm the impact of CS projects. There is often a significant delay between the gathering of CS data and the formal publication of these data in scientific articles and papers and indeed many CS projects do not aim to achieve formal scientific publications of their work (Kullenberg and Kasperowski, 2016; Theobald *et al.*, 2015). Many CS projects achieve impact predominantly through the use and specifically re-use of gathered data, but such usage is hampered by questions of data quality and relies heavily on other researchers being able to find and evaluate such data (Burgess *et al.*, 2017). The failure to upload datasets to commonly available and accepted repositories and to follow best practices regarding data management are therefore detrimental not only to scientific research, but to the very initiatives which aim to facilitate the dissemination of research data.

Further to these concerns about the impact and reusability of data, we also note ethical concerns stemming from the inconsistent documentation and contextualisation of data. Prior research has demonstrated that citizen volunteers generally support openness and demonstrate a willingness to share potentially sensitive data as long as it is for scientific research purposes (Bowser *et al.*, 2017). Even so, analyses of biodiversity datasets suggest that the availability of volunteer-derived datasets does not adequately match the desires and motivations of volunteers and indeed is often

not recognised by potential end-users of the data (Groom *et al.*, 2017). Our findings support this concern – while we cannot speculate on whether the availability of the datasets matches the expectations of volunteers, it is nonetheless clear that the restrictions associated with the datasets are not clearly communicated to potential users or, in many cases, communicated at all.

Finally, a common issue identified across the projects is one of findability – the capacity for researchers to discover the datasets and findings as a first step towards reuse. Studies of environmental monitoring CS projects have identified an overlap between projects and research issues (Geoghegan *et al.*, 2016). We suggest that there is significant risk of repetition within projects as well as that projects will aim to reinvent the proverbial wheel in spite of existing software and data solutions being available, in part because of difficulties in finding and exploiting these existing resources stemming from the failure to follow best open access practices.

5.2. Recommendations

Based on our findings and on the issues discussed above, we developed a set of recommendations to support project administrators at all stages of the research process, from defining problems, to gathering, managing and analysing data, as well as successfully disseminating findings and data in an openly accessible manner.

1. Projects should start with a clear problem definition. This should identify not only the research questions and hypotheses for analysis, but also any stakeholder groups who may benefit from the data or outputs from the project.
2. The openness and availability of any data should be considered throughout the project and should guide many of the data collection, analysis and dissemination decisions. Projects should complete a full data management plan – however provisional – as early as possible, which should be updated or replaced as necessary throughout the life of the project.
3. Any preprocessing of data should be clearly and succinctly accessible alongside datasets and other research outputs. This may include a version history, methodological description or pre-processed version of the dataset.
4. To avoid unnecessary repetition of existing analyses and to provide contextual details on the intended purpose(s) of gathered data, projects should document completed and/or intended analyses alongside datasets. Wherever possible, this should include numerical results, data visualisations and the interpretation and analyses of results. Most optimally, this would include a text-based report, which would be stored and disseminated alongside datasets.
5. If stakeholders are to make use of data, then they must first be able to find and access any datasets and outputs. Projects should disseminate their results through openly accessible repositories wherever possible. Any dataset should include a permanent identifier such as a DOI, as well as a human readable and ideally machine-readable license.

6. Perhaps most crucially, there should be greater integration between external repositories and projects web-presences. Project web-pages should have clear, visible links to external datasets and other resources.
7. Use public repositories for software, like GitHub, to maintain a control version of your code and to describe how to deploy it by other users. Plus, there is a module in Zenodo that allows users to automatically generate a DOI for their software in GitHub every time the user creates a release.

In light of these recommendations, especially those related to management of data, we present in the next section a tool that helps users to create Data Management Plans in CS projects applying FAIR principles, and complying with the above recommendations (and, in particular, recommendations 2, 5, 6 and 7).

6. A solution for Citizen Science Projects Data Management Plans

6.1. Motivation and related work

As per our recommendations in the previous section, one way to ensure a coherent and successful plan for CS projects is to define a Data Management Plan (DMP). A DMP document is often the standard way to explicitly define a strategy for data collection, processing and archiving, in line with the best practices of data science and open science. In several contexts, including the research and innovation projects supported by the NSF and by the Horizon 2020 Programme, the DMP is also a contractual obligation and a reference template is available to guide the compiler to cover all the necessary aspects, from a summary of data, to FAIR principle compliance to security and other aspects. The DMP compiled at the beginning of the project serves also as a guide throughout the project execution and should be kept up-to-date as soon as the activities progress.

However, creating a DMP requires a good understanding of the best practices behind data management and also a good set of skills related to data handling. As our research showed, most of the analysed CS projects are not aware of the importance of data management and our findings suggest that it is unlikely that most CS project coordinators possess the needed competences to compile a DMP. Our finding is in line with previous literature results: half of the CS projects analysed by Schade and Tsinaraki (2016) had no DMP in place, even if it is recognized that a DMP is crucial both from a methodological point of view and for a scientific acceptance of project results (Mckinney *et al.*, 2017). Therefore, to be able to implement our recommendations, CS projects should be supported with tools and guidance that ensure a successful data management but that do not require a deep understanding of DS and OS. Some tools supporting the creation of a DMP already exist, examples including the Data Stewardship Wizard (https://ds-wizard.org), DMPTool (https://dmptool.org), Argos (https://argos.openaire.eu/splash) or DMPOnline (https://dmponline.dcc.ac.uk). The last two projects are mainly focused on the lifecycle of the DMP

and use the same questions as those in the official template for DMPs (in addition they provide some indications to answer the questions). However, they do not fully hide the complexity or sufficiently simplify completion to guide inexpert users, especially in the context of CS projects. The approach of our tool is different in the sense that we use a small set of questions specifically designed for CS projects, and provide an interactive mechanisms to recommend users the text they should include based on their answers.

6.2. Overview of the ACTION DMP Tool for CS projects

The ACTION DMP Tool is a software tool that we developed to facilitate the creation of DMPs in CS projects. Since each dataset generated in CS projects must have a DMP, this tool turns out to be very useful for simplifying the management of generated data.

The ACTION DMP Tool is a web application that provides the user with a simple form to be completed in order to create the DMP. The user is guided in the DMP drafting by a set of questions that facilitate the understanding of the required information to be entered into this document (see Figure 1).

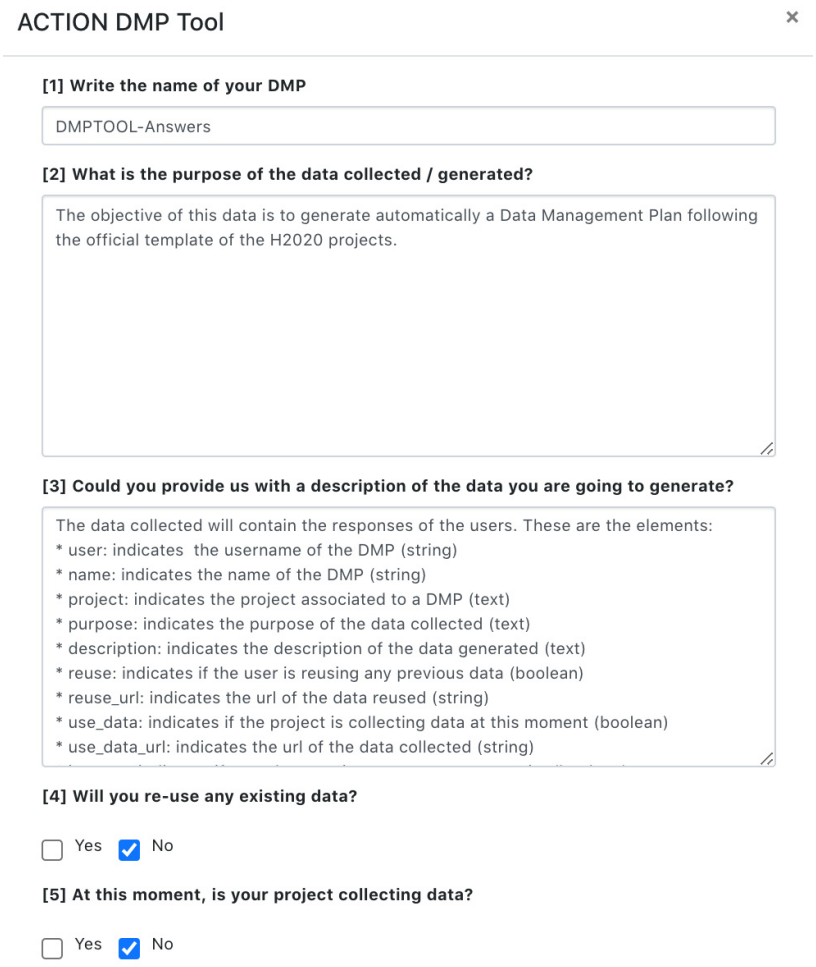

Figure 1 Example form in the ACTION DMP Tool.

Complex concepts are explained through an interactive interface that guides the user step-by-step in selecting the most appropriate answer. This step-by-step guide was developed using Coney (Celino and Re Calegari, 2020), a tool for creating chat-like conversations that can be customized according to users' answers. These "conversations" built with Coney were integrated as links in the form to help the users whenever they do not know how to fill in a specific section. It is indeed the case that people who are not data experts find it hard to understand what kind of content they should include in a specific section of the DMP. Users are supported to better understand their own data management practices, letting them choose between pre-defined options, to build the best possible answer to fill in the DMP tool fields.

An example of such a "tutorial" is the selection of the open data license for data release: we implemented the "decision tree" to choose a Creative Common license in a Coney conversation so that the user is guided through questions to understand which one best suites their case. Figure 2 shows two screenshots of this guiding "conversation" (the interested reader can also try this tutorial out at https://bit.ly/coney-license).

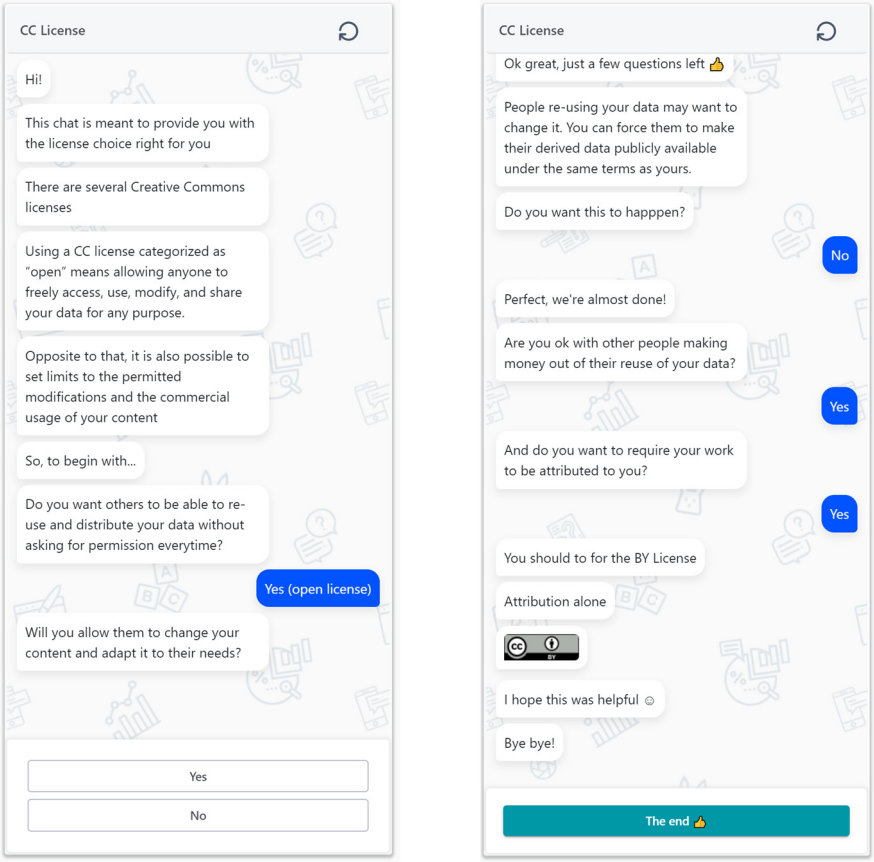

Figure 2 Coney tutorial about the selection of the Creative Common licenses.

Once the user answers the questions in the form, the user can have a DOCX file generated by the tool. This document, based on a template designed specifically for H2020 projects (https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf), is generated with the user-provided information. The document

shows the plan to publish data following the Open Data FAIR principles, incorporating the recommendations we outlined above. The user can further adapt the document by editing the generated version.

The first version of the ACTION DMP Tool can be found at https://dmptool.actionproject.eu. A technical description can be found at https://zenodo.org/record/3885566 and the software is available with an open source license at https://github.com/actionprojecteu/dmptool and https://github.com/actionprojecteu/dmptool-generator. The code of the Coney tool (integrated in the ACTION DMP Tool to create the tutorials) is also available with an open source license at https://github.com/cefriel/coney.

6.3. Evaluation and adoption of the ACTION DMP Tool

The ACTION project is a H2020 Research and Innovation Action aimed to set up an “acceleration” programme for CS initiatives related to pollution. ACTION supports CS projects with intensive training as well as co-design of tools and methodologies to successfully facilitate participatory data collection and analysis, according to the open science principles, including the provision of a digital infrastructure for data and result management. In other words, the target participants of the accelerator are representatives of the very type of project analysed in this paper.

As part of the acceleration programme, we had the chance to gather 14 representatives from 8 countries representing 11 CS projects (https://actionproject.eu/citizen-science-projects); their expertise range from air, water, chemical and light pollution to activism and inclusion. None of them was an expert in data management, although a minority of the participants had some limited previous experience with data collection and processing; all of them declared a lack of familiarity with DMP and an inability to compile it without some help. Therefore, we solicited them to act as early adopters of the ACTION DMP Tool and, following a co-design approach, we collected their feedback and suggestions. This happened in February 2020, during the kick-off meeting of the acceleration programme (https://actionproject.eu/action-short-workshop-report).

During this evaluation and co-design session, we went through various sections of the DMP, investigating which parts were more difficult to grasp for them and supported them in understanding the requirements as well as best practices in data management. For example, we discovered that citizen scientists have a hard time in understanding expert data topics, like open data licenses or quality assurance of data. The session proved to be very helpful because, together with the involved CS projects, we were able to improve our digital tool to create a DMP document to make it even more effective: the DMP sections are explained and rephrased in a language suitable for a non-expert audience and in which the compiling user is guided step-by-step in the understanding of more technical concepts.

The final version of the ACTION DMP tool is the result of that session. Thanks to this tool, the CS projects participating in the ACTION acceleration programme were able to create their DMP documents. As an example we provide reference to the DMP related to the CitiComPlastic project (https://actionproject.eu/citicomplastic): the raw version, generated by the tool is available at

https://doi.org/10.5281/zenodo.3958868, while the final version of the DMP (edited by the project coordinators starting from the raw version) is at https://doi.org/10.5281/zenodo.3885248.

7. Conclusions

In the context of CS projects moving towards a co-creation approach, understanding how various assets (including data, software, and publication of results) are handled and shared within and outside of the projects they co-create will become a relevant aspect of managing CS projects in the future. In this paper we took a Data- and Open-Science centric approach in analysing 48 pollution-related CS projects to better understand how they manage their assets, and based on the analysis we outlined a number of recommendations and proposed tools to improve the management of the assets.

As part of future work, we envision two possible directions. The first direction is related to further analysis of CS projects, including applying our analysis approach to analysing other types of CS projects (non-pollution), analysing projects led by scientists vs. non-scientists, and how to deal with complementary aspects such as managing data privacy in CS projects. The second direction is related to further validating our analysis approach for 16 CS project pilots covering agricultural, air, light, noise and water pollution. This validation is planned to take place in the context of the H2020 ACTION, including development of a data workflow tool (to implement small scientific workflows for CS projects to manage and publish their data following the DMP defined with the DMP Plan Tool presented in this paper), and raising more awareness on data management aspects through workshops, seminars, and interviews with citizen scientists.

8. Data Availability

- Raw data used for the analysis (containing data collected for the considered projects):
  - https://doi.org/10.5281/zenodo.3958853

- Software repositories:
  - https://github.com/actionprojecteu/dmptool
  - https://github.com/actionprojecteu/dmptool-generator
  - https://github.com/cefriel/coney

9. Acknowledgements

The work in this paper is partly funded by the H2020 project ACTION (grant number 824603). We thank the ACTION consortium partners for fruitful discussions related to analyzing Citizen Science projects and particularly to the pilots' providers in the project.

10. References

Anhalt-Depies, C., Stenglein, J.L., Zuckerberg, B., Townsend, P.A. and Rissman, A.R., 2019. Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. *Biological Conservation*, *238*, p.108195.

Borgman, C.L., 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, Cambridge, MA.

Bowser, A., Cooper, C., de Sherbinin, A., Wiggins, A., Brenton, P., Chuang, T.R., Faustman, E., Haklay, M.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), Vol. 5 No. 1, 18, pp. 1-16.

Budde, M., Schankin, A., Hoffmann, J., Danz, M., Riedel, T., & Beigl, M., 2017. Participatory sensing or participatory nonsense? Mitigating the effect of human error on data quality in citizen science. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1-23. DOI: http://doi.org/10.1145/3131900

Bowser, A., Shilton, K., Preece, J., & 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 on 25th February, pp. 2124-2136. DOI: http://dx.doi.org/10.1145/2998181.2998305

Burgess, H.K., DeBey, L.B., Froehlich, H.E., Schmidt, N., Theobald, E.J., Ettinger, A.K., HilleRisLambers, J., Tewksbury, J. and Parrish, J.K., 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

Cao, L., 2017. Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1-42. DOI: http://dx.doi.org/10.1145/3076253

Cao, Y., Jones, C., Cuevas-Vicenttín, V., Jones, M.B., Ludäscher, B., McPhillips, T., Missier, P., Schwalm, C., Slaughter, P., Vieglais, D. and Walker, L., 2016, DataONE: a data federation with provenance support. In *International Provenance and Annotation Workshop*, McClean on 7th June, pp. 230-234. DOI: https://doi.org/10.1007/978-3-319-40593-3_28

Celino, I. and Re Calegari, G., 2020. Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies*, *139*, p.102410.DOI: https://doi.org/10.1016/j.ijhcs.2020.102410

Cooper, S., Sterling, A.L., Kleffner, R., Silversmith, W.M. and Siegel, J.B., 2018. Repurposing citizen science games as software tools for professional scientists. In *Proceedings of the 13th International Conference on the Foundations of Digital Games*, Malmö in August, pp. 1-6. DOI: https://doi.org/10.1145/3235765.3235770

Elliott, K.C. and Rosenberg, J., 2019. Philosophical foundations for citizen science. *Citizen Science: Theory and Practice*, *4*(1). DOI: http://doi.org/10.5334/cstp.155

Fishbain, B., Lerner, U., Castell, N., Cole-Hunter, T., Popoola, O., Broday, D.M., Iñiguez, T.M., Nieuwenhuijsen, M., Jovasevic-Stojanovic, M., Topalovic, D. and Jones, R.L., 2017. An evaluation tool kit of air quality micro-sensing units. *Science of the total environment*, *575*, pp.639-648. DOI: https://doi.org/10.1016/j.scitotenv.2016.09.061

Forrest, S.A., Holman, L., Murphy, M. and Vermaire, J.C., 2019. Citizen science sampling programs as a technique for monitoring microplastic pollution: results, lessons learned and recommendations for working with volunteers for monitoring plastic pollution in freshwater ecosystems. *Environmental monitoring and assessment*, *191*(3), p.172. DOI: https://doi.org/10.1007/s10661-019-7297-3

Geoghegan, H., Dyke, A., Pateman, R., West, S. and Everett, G., 2016. Understanding motivations for citizen science. *Final report on behalf of UKEOF, University of Reading, Stockholm Environment Institute (University of York) and University of the West of England*.

Gray, S., Jordan, R., Crall, A., Newman, G., Hmelo-Silver, C., Huang, J., Novak, W., Mellor, D., Frensley, T., Prysby, M. and Singer, A., 2017. Combining participatory modelling and citizen science to support volunteer conservation action. *Biological conservation*, *208*, pp.76-86. DOI: https://doi.org/10.1016/j.biocon.2016.07.037

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

Hadj-Hammou, J., Loiselle, S., Ophof, D. and Thornhill, I., 2017. Getting the full picture: Assessing the complementarity of citizen science and agency monitoring data. *PLoS One*, *12*(12), p.e0188507. DOI: https://doi.org/10.1371/journal.pone.0188507

Haklay, M., 2013. Citizen science and volunteered geographic information: Overview and typology of participation. In Crowdsourcing geographic knowledge pp. 105-122. DOI: https://doi.org/10.1007/978-94-007-4587-2_7

Jiang, Q., Bregt, A.K. and Kooistra, L., 2018. Formal and informal environmental sensing data and integration potential: Perceptions of citizens and experts. *Science of The Total Environment*, *619*, pp.1133-1142. DOI: https://doi.org/10.1016/j.scitotenv.2017.10.329

Kullenberg, C. and Kasperowski, D., 2016. What is citizen science?–A scientometric meta-analysis. *PloS one*, *11*(1), p.e0147152. DOI: https://doi.org/10.1371/journal.pone.0147152

Lamprecht, A.L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., van de Sandt, S., Ison, J., Martinez, P.A. and McQuilton, P., 2019. Towards FAIR principles for research software. *Data Science*, Vol. 3 No. 1, pp. 37-59. DOI: https://doi.org/10.3233/DS-190026

Lana, M., 2019, January. Information literacy needs open access or: Open access is not only for researchers. In Italian Research Conference on Digital Libraries, pp. 236-247. DOI: https://doi.org/10.1007/978-3-030-11226-4_19

Mahajan, S., Kumar, P., Pinto, J.A., Riccetti, A., Schaaf, K., Camprodon, G., Smári, V., Passani, A. and Forino, G., 2020. A citizen science approach for enhancing public understanding of air pollution. *Sustainable Cities and Society*, *52*, p.101800. DOI: https://doi.org/10.1016/j.scs.2019.101800

McKinley, D.C., Miller-Rushing, A.J., Ballard, H.L., Bonney, R., Brown, H., Cook-Patton, S.C., Evans, D.M., French, R.A., Parrish, J.K., Phillips, T.B. and Ryan, S.F., 2017. Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, *208*, pp.15-28. DOI: https://doi.org/10.1016/j.biocon.2016.05.015

Musto, J. and Dahanayake, A., 2020. Improving data quality, privacy and provenance in citizen science applications. *Information Modelling and Knowledge Bases XXXI*, *321*, p.141.

Nascimento, S., Rubio Iglesias, J.M., Owen, R., Schade, S. and Shanley, L., 2018. Citizen science for policy formulation and implementation. London: UCL Press.

Nielsen, M., 2020. 6. All the World's Knowledge. In *Reinventing Discovery* (pp. 91-128). Princeton University Press.

Paul, J.D. and Buytaert, W., 2018. Citizen science and low-cost sensors for integrated water resources management. In *Advances in chemical pollution, environmental management and protection,* 3, pp. 1-33. DOI: https://doi.org/10.1016/bs.apmp.2018.07.001

Pocock, M.J., Tweddle, J.C., Savage, J., Robinson, L.D. and Roy, H.E., 2017. The diversity and evolution of ecological and environmental citizen science. *PLoS One*, *12*(4), p.e0172579. DOI: https://doi.org/10.1371/journal.pone.0172579

Poisson, A.C., McCullough, I.M., Cheruvelil, K.S., Elliott, K.C., Latimore, J.A. and Soranno, P.A., 2020. Quantifying the contribution of citizen science to broad-scale ecological databases. Frontiers in Ecology and the Environment, 18(1), pp.19-26. DOI: https://doi.org/10.1002/fee.2128

Ponti M.and Craglia M. 2020. *Citizen-generated data for public policy*. European Commission, Ispra 2020JRC120231, Brussels.

Rambonnet, L., Vink, S.C., Land-Zandstra, A.M. and Bosker, T., 2019. Making citizen science count: Best practices and challenges of citizen science projects on plastics in aquatic environments. *Marine pollution bulletin*, *145*, pp.271-277. DOI: https://doi.org/10.1016/j.marpolbul.2019.05.056

Rasmussen, L.M., 2019. Confronting research misconduct in citizen science. *Citizen Science: Theory and Practice*, *4*(1). DOI: http://doi.org/10.5334/cstp.207

Sagy, O., Golumbic, Y.N., Abramsky, H.B.H., Benichou, M., Atias, O., Braham, H.M., Baram-Tsabari, A., Kali, Y., Ben-Zvi, D., Hod, Y. and Angel, D., 2019. Citizen science: An opportunity for learning in the networked society. In *Learning In a Networked Society* pp. 97-115.

Schade, S., & Tsinaraki, C. 2016. Survey report: data management in Citizen Science projects. European Union Technical Report EUR 27920. European Union: Luxembourg. DOI: https://doi.org/10.2788/539115

Simonis, I., 2018. Standardized Information Models to Optimize Exchange, Reusability and Comparability of Citizen Science Data. A Specialization Approach. *International Journal of Spatial Data Infrastructures Research*, *13*, pp.38-47.

Theobald, E.J., Ettinger, A.K., Burgess, H.K., DeBey, L.B., Schmidt, N.R., Froehlich, H.E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M.A. and Parrish, J.K., 2015. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, *181*, pp.236-244.

Tiufiakov, N., Dahanayake, A. and Zudilova, T., 2018, September. Data Provenance in Citizen Science Databases. In *European Conference on Advances in Databases and Information Systems* pp. 242-253.

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. DOI: https://doi.org/10.1371/journal.pbio.1002280

Wiggins, A. and Crowston, K., 2011. From conservation to crowdsourcing: A typology of citizen science. In *2011 44th Hawaii international conference on system sciences*, Kauai on 4th January, pp. 1-10. In San Francisco on 27th February. DOI: https://doi.org/10.1109/HICSS.2011.2017

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*, pp. 1548-1559. DOI: https://doi.org/10.1145/2818048.2820063

Wiggins, A. and Wilbanks, J., 2019. The rise of citizen science in health and biomedical research. *The American Journal of Bioethics*, *19*(8), pp.3-14. DOI: https://doi.org/10.1080/15265161.2019.1619859

Williams, J., Chapman, C., Leibovici, D., Loïs, G., Matheus, A., Oggioni, A., Schade, S., See, L. and van Genuchten, P., 2018. Maximising the impact and reuse of citizen science data. London: UCL Press.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), pp.1-9. DOI: https://doi.org/10.1038/sdata.2016.18

Zaman, J., Kambona, K. and De Meuter, W., 2020. A reusable & reconfigurable citizen observatory platform. *Future Generation Computer Systems*, Vol. 114, pp. 195-208, DOI: https://doi.org/10.1016/j.future.2020.07.028

Zilliox, S. and Smith, J.M., 2018. Colorado's fracking debates: citizen science, conflict and collaboration. Science as Culture, 27(2), pp.221-241. DOI: https://doi.org/10.1080/09505431.2018.1425384