European Commission

JRC TECHNICAL REPORTS

Survey report: data management in Citizen Science projects

Sven Schade, Chrysi Tsinaraki

2016

EUR 27920 EN

Joint Research Centre

Contact information
Name: Sven Schade
Address: Via Enrico Fermi, 2749, I-21027 Ispra (VA) Italy
E-mail: sven.schade@jrc.ec.europa.eu
Tel.: +39 0332 78 5723

JRC Science Hub
https://ec.europa.eu/jrc

JRC101077

EUR 27920 EN

ISBN 978-92-79-58387-2 (PDF)

ISSN 1831-9424 (online)

doi:10.2788/539115 (online)

How to cite: Schade S, Tsinaraki C.; Survey report: data management in Citizen Science projects; EUR 27920 EN; Luxembourg (Luxembourg): Publications Office of the European Union; 2016; doi:10.2788/539115

Table of Contents

Acknowledgements 3
Abstract 4
1 Introduction 5
2 Survey methodology 6
3 Survey launch and dissemination 7
4 Data pre-processing 9
5 Survey results – per question 9
5.1 General information about the project 9
5.2 Discoverability of Citizen Science data 13
5.3 Accessibility of Citizen Science data 14
5.4 Re-use conditions of Citizen Science data 16
5.5 Usability of Citizen Science data 19
5.6 Preservation and curation of Citizen Science data 21
5.7 Additional Information 24
6 Cross-question analysis 25
6.1 Investigations related to data granularity 25
6.2 Findings related to Open Access 28
6.3 Funding-related analysis 30
6.4 Relations to Data Management Plans 33
6.5 Characteristics of projects running for more than four years 37
6.6 Selected details about projects with long-term data curation 39
7 Discussion of survey results 40
8 Conclusions from the survey 42
List of figures 45
List of tables 48
ANNEX A: Survey introduction and questions 49

Acknowledgements

This work would not have been possible without the support of the participating Citizen Science projects, we are very satisfied and grateful about the many replies that we received. The high number of responses could only be achieved thanks to the many organizations and individuals that helped distributing our call for participation, here we particularly acknowledge the support of the INSPIRE, PEER and EUROGEO communities, the Software Sustainability Institute, ECSA, CSA, ACSA and the national Citizen Science networks of Germany, Switzerland and Austria, as well as the projects SOCIENTIZE, EUBON, Citizen Sense and the Citizens' Observatories. Special thanks also to the following power users on twitter: @EUexpo2015, @MozillaScience, @SciStarter, @ICTscienceEU and @john_magan. The preparation of the questionnaire was greatly supported by Jose Miguel Rubio, Anne Bowser, Muki Haklay, Claudia Göbel, Arne Berre, Jaume Piera, Max Craglia, Andrea Perego and Cristina Rosales Sanchez.

Abstract

The Citizen Science and Smart City Summit, organised by the European Commission's Joint Research Centre (JRC) in 2014, identified the management of citizen-collected data as a major barrier to the re-usability and integration of these contributions into other data-sharing infrastructures across borders. We followed-up on these findings with a survey with Citizen Science projects, experiments on a repository for EU-funded Citizen Science projects and discussions with European and international Citizen Science communities. This report summarises the outcomes of the survey. Amongst other findings, the 121 responses clearly underlined the diversity of projects in terms of topicality, funding mechanisms and geographic coverage, but responses also provided valuable insights related to the access and re-use conditions of project results. While, for example, 60% of the participating projects follow a dedicated data management plan and the majority of projects provide access to raw or aggregated data, the exact re-use conditions are not always present or miss well-defined licenses. Apart from replies from all across the globe, this activity also helped us to connect to the relevant players, helping to initiate discussions about data management for Citizen Science with representatives of the European, American and Australian Citizen Science associations.

The anonymised dataset of survey replies, together with a script (in R) that executes the main analysis steps are published as JRC Open Data, in accordance with the Commission decision on Reuse of Commission Documents (2011/833/EU).

1 Introduction

Voluntarism and participatory approaches have a long tradition in environmental and ecological sciences, especially in biodiversity[1]. With the advent of new technologies, public participation in research (Citizen Science) has just entered into a new era of possibilities, including its latest applications in areas such as astronomy[2] and earth observation[3]. As well as these developments, the European Commission (EC) white paper on Citizen Science already identifies general public engagement in scientific research activities as a mechanism to improve science-society-policy interactions, alongside democratic research based on evidence and informed decision-making[4]. Accordingly, citizen engagement in science and policy emerged as a prominent topic in areas such as Better Regulation[5], Responsible Research and Innovation[6], and environmental policy[7] (see Annex A for details). Until now, however, operational approaches of integrating citizens in the provision of scientific evidence for policy making are yet to be established at the European level.

In 2014, the EC's Joint Research Centre (JRC) organised the Citizen Science and Smart City Summit, which identified the management of citizen-collected data as a major barrier to the re-usability and integration of these contributions across borders[8]. In 2015, we followed up on these findings with a survey of Citizen Science projects, experiments on a repository for EU-funded Citizen Science projects and discussions with the European and international Citizen Science community. This deliverable summarises the outcomes of the survey.

Between 13 July and 4 September 2015, Citizen Science projects were asked to answer a questionnaire to provide the JRC and Citizen Science practitioners around the globe with insights into their data management approaches and best practices. With this activity, we wanted to better understand examples of current practice and use this as a basis for discussions with Citizen Science communities world-wide. Beyond the aims of pure stocktaking and awareness-raising, this survey aimed to also establish a baseline for prioritising subsequent actions and provide a means to measure progress.

The overall approach of the survey and the dissemination procedure are presented in the next sections (Section 2 and Section 3), followed by an explanation of the applied data cleaning and pre-processing in Section 4. A presentation of the results for each question in Section 5 is followed by some analysis on the interdependencies of answers between several questions (Section 6). We provide some discussion of the results (Section 7) before concluding in Section 8.

---

[1] See for example, the Citizen Science Best Practice Guide - https://www.ceh.ac.uk/citizen-science-best-practice-guide

[2] The most prominent example being Galaxy Zoo - http://www.galaxyzoo.org/

[3] For example, the use of citizen to identify land use from satellite imagery in GeoWiki - http://www.geo-wiki.org/

[4] Available from: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=6913

[5] Communication COM(2015) 215 final "Better regulation for better results - An EU agenda"

[6] Rome Declaration on Responsible Research and Innovation in Europe, Friday, 21 November 2014 -https://ec.europa.eu/research/swafs/pdf/rome_declaration_RRI_final_21_November.pdf

[7] Regulation (EC) No 1367/2006 of 6 September 2006 on the application of the provisions of the Aarhus Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters to Community institutions and bodies, and also the EC In-Depth Report on Environmental Citizen Science http://ec.europa.eu/environment/integration/research/newsalert/pdf/IR9_en.pdf

[8] Citizen Science and Smart Cities, Report of Summit, Ispra 5-7 February 2014 - http://publications.jrc.ec.europa.eu/repository/bitstream/JRC90374/lbna26652enn.pdf

2 Survey methodology

We decided to run the survey with an open call, in which anybody who considered his/her project to deliver Citizen Science data was invited to share their experiences. Although initially intended to have a European focus, discussions with members of the European Citizen Science Association (ECSA)[9] and the international Citizen Science Association (CSA)[10], led to the decision to open the survey to the international community, so that non-EU and globally acting organisations could also contribute and benefit from the results. As well as targeting Earth science and environmental communities, responses were sought from projects in other domains, including the social sciences.

The survey involved the workflow as depicted in Figure 1).

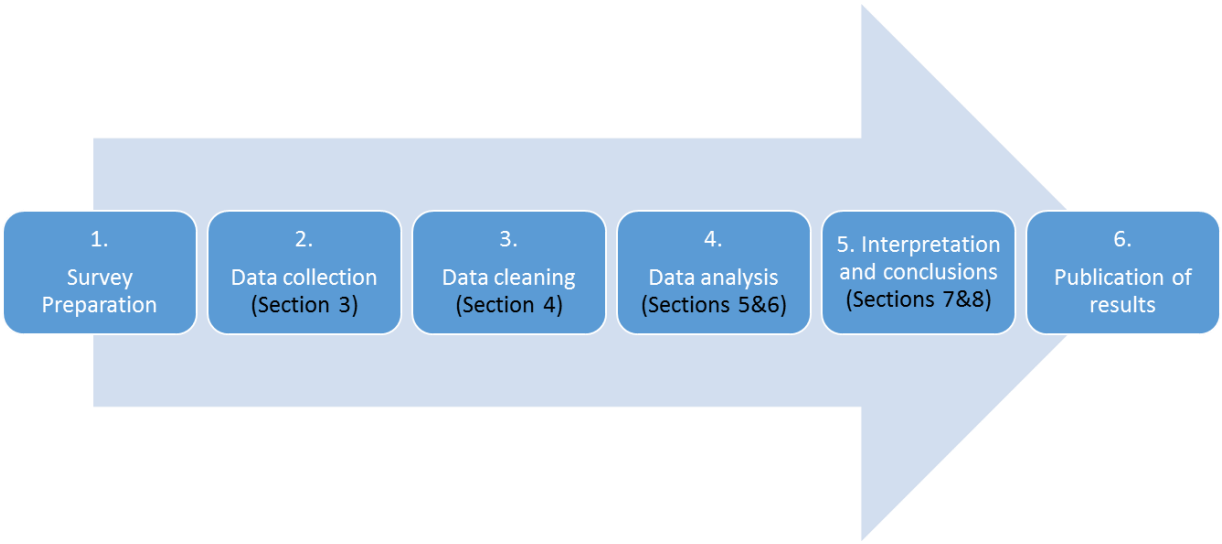

Figure 1: Overview of methodology.

1. **Preparation of the questionnaire**: The questions (see also Annex C) are inspired by data management principles being discussed in the Group on Earth Observation (GEO) and the Belmond Forum. Generic information about projects (e.g. about topical and geospatial coverage, duration and funding mechanisms) were complemented with specific questions about data set discoverability, accessibility, re-use conditions, usability, and data preservation. Finally, participants could identify themselves and their project(s) and provide final remarks. ECSA and CSA members, together with additional experts from the field and colleagues from DG Research and Innovation (RTD) and the European Environment Agency (EEA) provided valuable suggestions to shape the final questionnaire, which was then implemented as an EUSurvey[11].
2. **Data collection**: Interested parties were invited to provide their inputs over the summer of 2015. Our dissemination approach is outlined in Section 3.
3. **Data cleaning**: After closing the call, we cleaned-up the data set by harmonising the spelling of words, ensuring the correct use of separators to be able to import the data in analytical software, and examining the selection of pre-defined categories. Details about this pre-processing are further outlined in Section 4.
4. **Data analysis**: The clean results were then analysed in two phases. Firstly, we analysed the replies per individual question (see Section 5). Secondly, we investigated dependencies between selected questions (see Section 6).

[9] Official web page - https://www.naturkundemuseum.berlin/en/forschung/ecsa-european-citizen-science-association

[10] Official web page - http://citizenscienceassociation.org/

[11] Official web page: https://ec.europa.eu/eusurvey with our survey (last accessed on 29 March 2016) still available - https://ec.europa.eu/eusurvey/runner/CSDataManagement

5. **Data interpretation and conclusions**: On this basis, we interpreted the results and drew our main conclusions from the survey (see Sections 7 and 8, respectively).
6. **Publication of the results**: Given our interest in open research methodologies, we have prepared this report and also publish the clean and anonymised data, as well as the scripts that were used for our analysis. All are free for either re-use or additional analysis for both commercial and non-commercial purposes.

3 Survey launch and dissemination

The survey (see also Annex A) was officially launched on 13 July 2015, as a JRC news item[12]. It was then disseminated via the e-mail lists of the ECSA, CSA and the newly formed Australian Citizen Science Association (ACSA)[13], colleagues at DG JRC, DG RTD, and DG Communications Networks, Content and Technology (CONNECT), as well as a list of approximately 200 individuals including representatives of the EU-funded Citizens' Observatories[14], Collaboration and Support COST Actions ENERGIC[15] and "Mapping and the Citizen Sensor"[16], International Institute for Applied Systems Analysis (IIASA), European Space Agency (ESA), European Network of Living Labs (ENoLL)[17], Mozilla Foundation[18] and many more.

A rich set of projects and initiatives followed the request for dissemination and included items about the survey in their newsletters and web pages. In addition to the previously mentioned ECSA, ACSA and CSA, the survey was also disseminated to the following initiatives, projects and networks (see also Figure 2):

- Infrastructure for Spatial Information in the European Community (INSPIRE),
- Partnership for European Environmental Research (PEER),
- EU-funded project EUBON,
- past EU project SOCIENTIZE,
- European Association of Geographers (EUROGEO),
- European Research Council (ERC)-funded project Citizen Sense,
- Software Sustainability Institute,
- Austrian Citizen Science network (Österreich forscht)
- German Citizen Science network (Bürger schaffen Wissen), and
- Swiss Citizen Science network.

The initiative also triggered significant traffic on social media platforms. Aside some attention on Facebook and LinkedIn (some 350 views), most people were reached via Twitter, where the message was, for example, promoted by the Director General of the JRC (@VladimirSucha), the EU account for the work exhibition EXPO 2015 (@EUexpo2015), Mozilla Science Lab (@MozillaScience), SciStarter (@SciStarter) and Digital for Science (@ICTscienceEU) - see also Figure 3. In this way, the request reached more than 40.000 Twitter users.

[12] News item on JRC Science Hub - https://ec.europa.eu/jrc/en/news/citizen-science-survey

[13] Official web page: http://www.citizenscience.org.au/

[14] Overview web page: http://www.citizen-obs.eu/

[15] Official web page: http://vgibox.eu/

[16] Official web page: http://www.citizensensor-cost.eu/

[17] Official web page: http://www.openlivinglabs.eu/

[18] Official web page: https://www.mozilla.org/en-US/foundation/

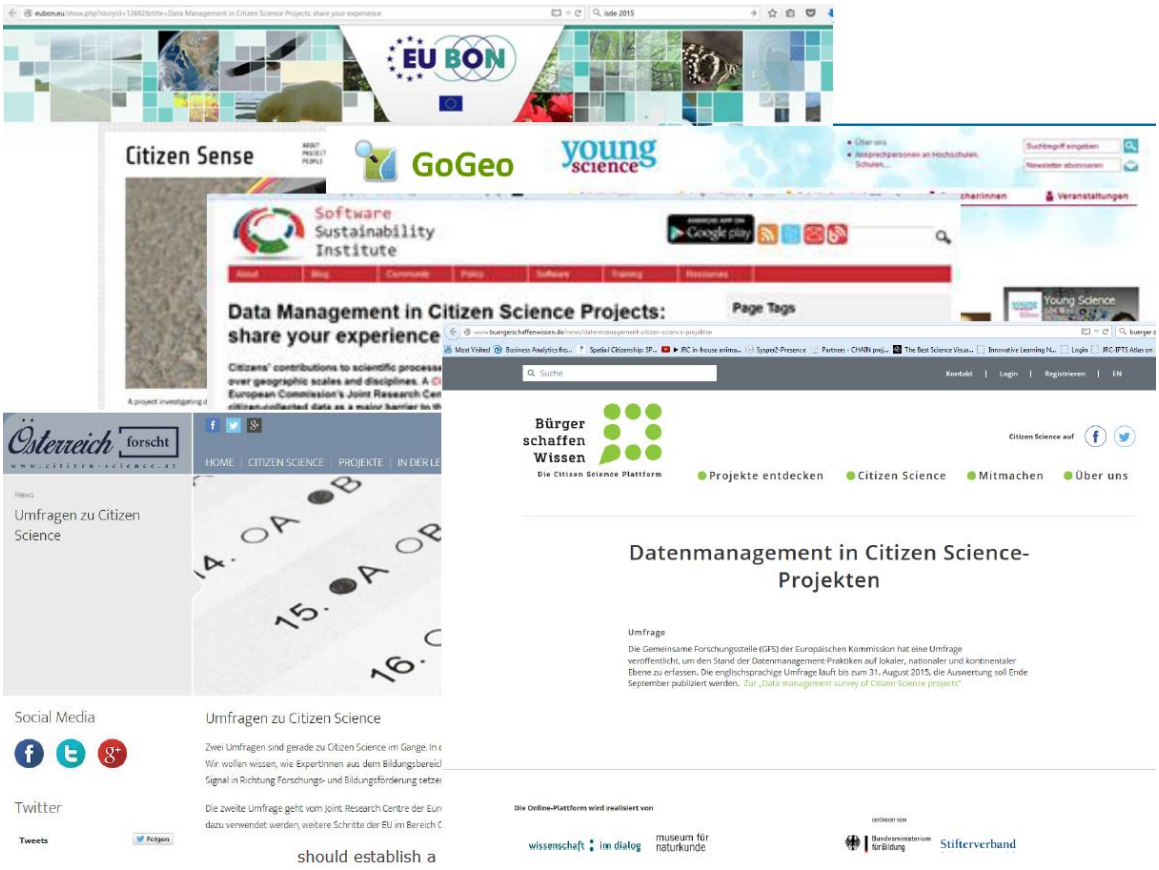

Figure 2: examples of the survey featuring on community web pages.

Figure 3 examples of the survey featuring on social media platforms.

4 Data pre-processing

As noted above, before initiating the analysis, we created a cleaned and pre-processed version of the received raw data. In particular, we:

- Harmonised the use of small and capital letters in the question responses that asked for free text in order to semi-automatically analyse some of this text.
- Harmonised the words that was used for describing countries. (In the responses to the open question 22. "In which country/countries do you store the data?")
- Ensured the correct use of separators, i.e. using ";" to be able to import the data in analytical software.
- Extended the assignment of "topics" (in the responses to question 1. "Which topic areas does your project cover?"), if the response to question 1.1 "Which 'other' topic(s) does the project cover?" suggested we do so. For example, where "Meteorology" was given as other topic, we also selected the topic "Earth science", applying common understanding of these fields[19].
- Extended the assignment of "themes" (in the responses to question 1.2 "Which environmental theme(s) does it cover?"), if the response to question 1.3 "Which 'other' environmental theme(s) does the project cover?" suggested to do so. For example, where "Endangered species" was given as other theme, we also selected the theme "Biodiversity".
- Harmonising the parts of free-text replies in order to cluster responses. For example, the harmonisation to abbreviations of licences in the responses to question 13.2 "Which license do you use?"

For the free-text analysis, we also removed indications that would have allowed to identify the respondent.

Before publishing the data set, we anonymised it completely including the removal of project names, web pages and e-mail addresses.

5 Survey results – per question

By 4 September 2015, a total of 121 projects completed the survey. In this section, we visit all questions in the survey. The sub-sections are organised according to clusters of questions as presented in the original survey. Interpretations and cross-comparisons are discussed in later sections[20].

5.1 General information about the project

The survey began with a block of questions to collect general information about the project.

5.1.1 Topic and Theme

As projects had the opportunity to select multiple topics, the majority of replies (84%) covered environmental topics, followed by Earth science (23%), social sciences (10%) and space science (6%), see also Figure 4. Other topics that have been mentioned are summarized in Table 1. It should be noted that a project could cover multiple topics. Here, we omit categorising the other topics any further due to arbitrariness of higher level categories. The table gives an overall impression of the diversity of the participating projects.

[19] As published on Wikipedia: https://en.wikipedia.org/wiki/Earth_science

[20] An anonymized and cleaned version of the original replies is publicly available, together with an R script that can be used on top of the survey data in order to re-produce these tables and graphics: http://data.europa.eu/89h/jrc-citsci-cs-survey-15.

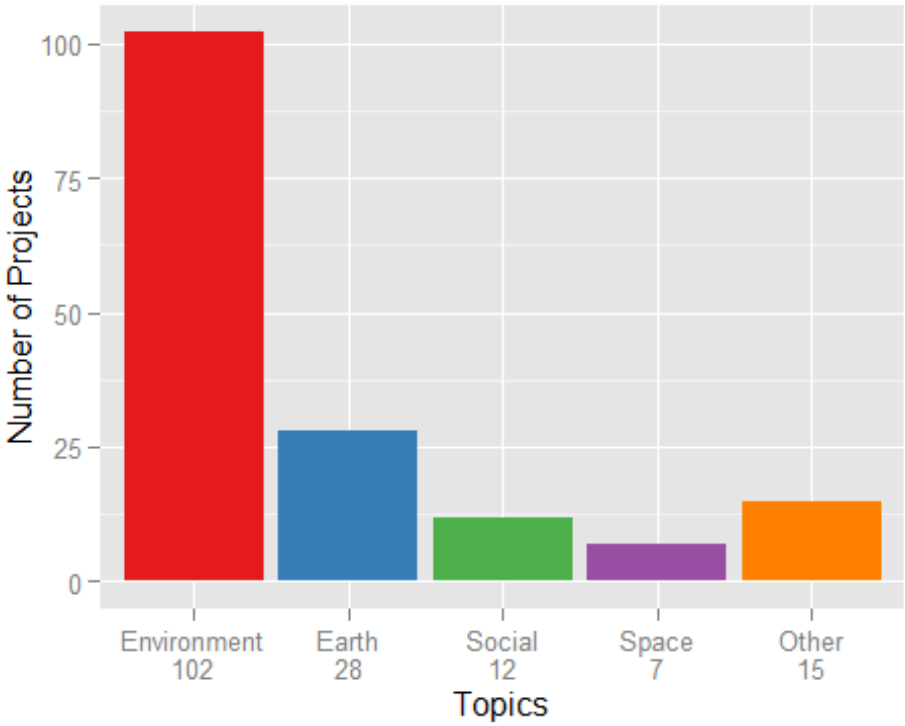

Figure 4: Answers to Question 1 - "Which topic areas does your project cover?"

Table 1: Answers to Question 1.1 – "Which 'other' topic(s) does the project cover?"

Other topic listed	Frequency
Humanities	**3**
Arts	2
Agriculture	1
Air pollution monitoring	1
Anthropocene	1
Biology	1
Biomarker studies	1
Biotechnology	1
Conservation	1
Cultural heritage	1
Cultural sciences	1
Ecology	1
Economics	1
e-Government	1
(Geographic) information science	1
Healthy Aging Science	1
Mapping	1
Materials	1
Meteorology	1
Pollution	1
Smart cities	1
Urban planning	1

In cases where environmental science has been selected, we asked for more details about the exact themes that an individual project addresses. Noticing that projects had the opportunity to select multiple themes, the majority of replies (68% of the 102 projects that selected "environment") covered the theme of biodiversity, followed by water quality (25%), land cover (21%), land use (18%), air quality (17%) and noise (9%), see also Figure 5. The "Other" category allowed respondents to provide further details about the topics of the project (see also Table 2).

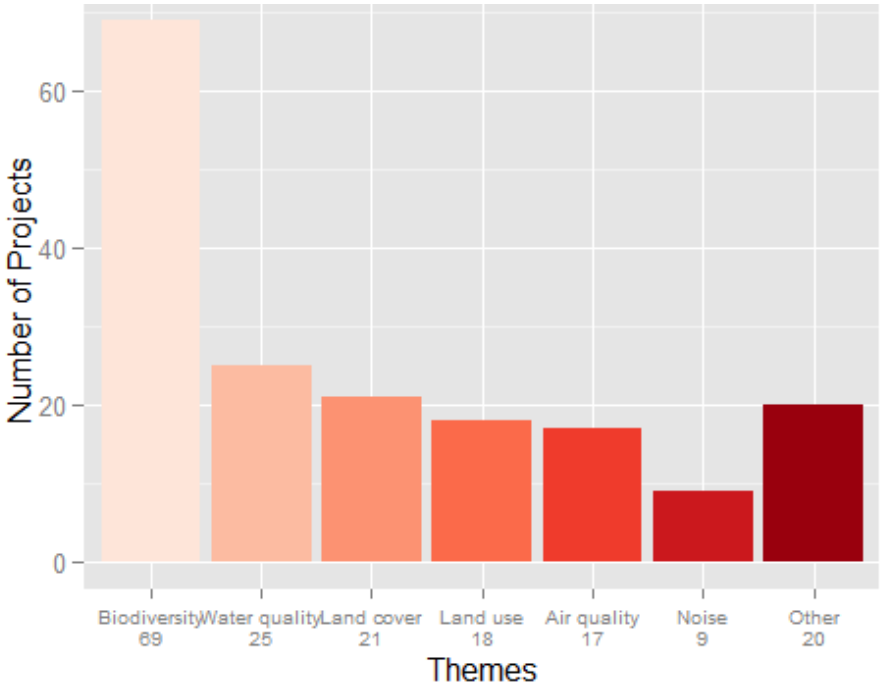

Figure 5: Answers to Question 1.2 - "Which environmental theme(s) does it cover?"

Table 2: Answers to Question 1.3 – "Which 'other' environmental theme(s) does the project cover?" (themes related to ecology highlighted in bold, those on pollution in light grey)

Other environmental theme listed	Frequency
Endangered species	**2**
Light pollution	2
Plant phenology	**2**
Biological observations	1
Carbon footprint	1
Contaminated sites	1
Environmental burdens	1
Environmental change	1
Fungi	**1**
Human Impacts to Ecological Systems	**1**
Invasive species	**1**
Light	1
Marine biology	**1**
Marine litter	1
Mobility	1
Phenology	**1**
Plastic pollution	1
Species and habitats	**1**

Traffic	1
Urban litter	1
Weather	1

5.1.2 Geographic Extent

Considering the spatial extent of the projects, projects were asked to select between various geographical scales and to identify if the project was active within the European Union (EU) or outside. Again, projects could select more than one option. The results are presented in Table 3, below. Although the amounts of replies per category are not likely to be statistically representative, they indicate a good geographic coverage of the topic.

Table 3: Answers to Question 2 – "Which geographic extent does the project cover?"

Extent	Inside EU	Outside EU
Neighborhood	22	24
City level	29	30
Regional	31	39
Country	37	30
Continental	27	26
Total coverage	**146**	**149**

5.1.3 Duration

The responses to the question about project duration revealed that more than half of the participating projects (55%) intend to last for more than 4 years, 28% are planned for one to four years and 5% for less than a year. 12% of the respondents did not know about the project duration (see also Figure 6).

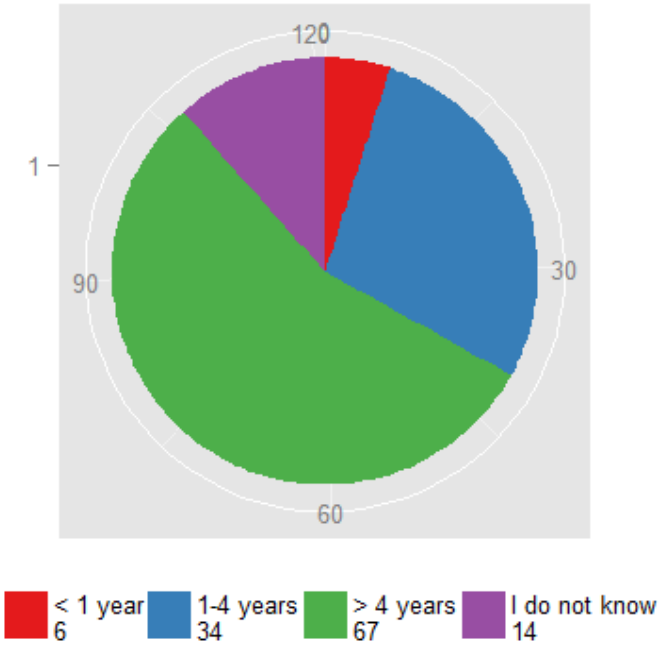

Figure 6: Answers to Question 3 - "What is the planned duration of the project?"

5.1.4 Status

Asked about the status of their project at the time of filling the survey (*Question 4 - "What is the status of this particular project?" - multiple choice, optional*), 90 participants (74%) were of the opinion that their activities take place in an operational setting, i.e. currently in use or ready for use; whereas 37 participants (21%) indicated that they were at an experimental stage, i.e. currently testing ideas and new techniques. Out of these, 7 participants (6%) chose both stages.

5.1.5 Funding

In terms of funding (see also Figure 6), national funding schemes dominated (36% of all participants), closely followed by in-kind contributions (34%) and then donations (22%). 23 participants (19%) indicated that at least part of their funding comes from EU-funding schemes. 10 participants (8%) said that they include participant or membership fees within their funding model. 34 replies (28%) pointed to other funding sources (*Question 5.1 – "Which 'other' funding mechanism is used?"*). However, the provided details were highly heterogeneous. Amongst other elements, they included crowdfunding, sales, local grants and institutional support.

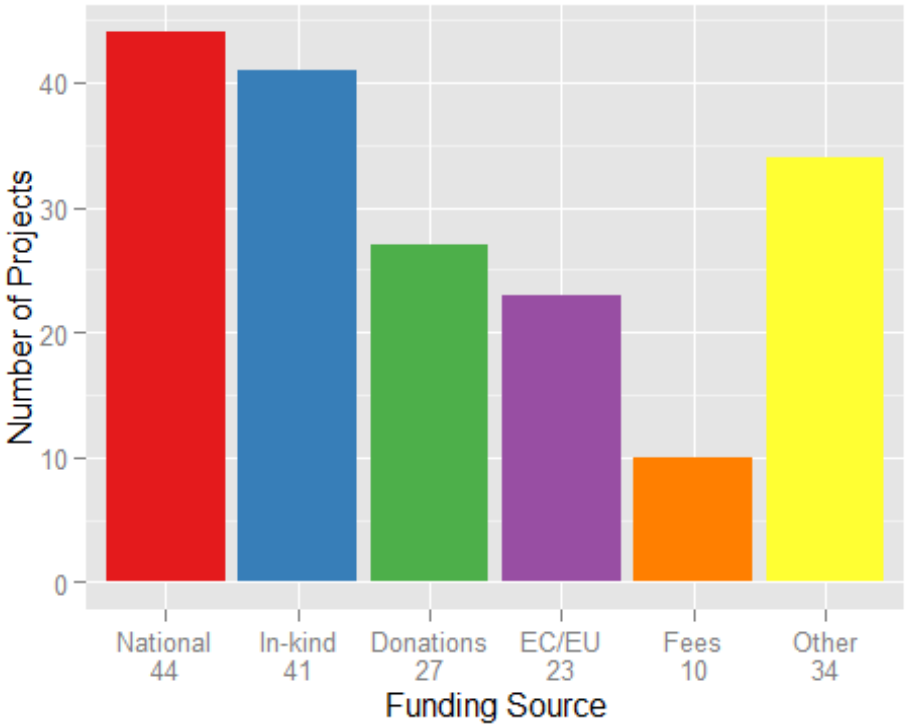

Figure 7: Answers to Question 5 - "How is this citizen science project financed?"

5.1.6 Data Management Plans

Asked for the availability of a data management plan (Question 6 - "Does your project include an explicit data management plan?" – single choice, mandatory), 72 of the 121 participants (almost 60%) indicated the availability of such a plan; 49 (40%) answered negatively.

5.2 Discoverability of Citizen Science data

As we are interested in data-sharing from such projects the next section of the survey focussed on discoverability, i.e. the ease of finding the data collected by the different Citizen Science projects.

5.2.1 Discoverability through catalogues and search engines

Asked for the discoverability of Citizen Science data (*Question 7 - "Is the data and all associated metadata from your project discoverable through catalogues and search engines?"*), 58 of the 120 respondents to this optional question (almost 47%) answered positively; 62 (51%) answered with "No".

5.2.2 Persistent Identifiers

Asked about details of the inclusion of identifiers in their data (Question 8 - "Does the data contain persistent, unique, and resolvable identifiers?" - single choice, optional), 82 of the 116 respondents to this optional question (almost 68% of the total participants) answered with "Yes"; 34 (28%) answered with "No".

5.3 Accessibility of Citizen Science data

After overall discoverability, we addressed issues of data accessibility, i.e. the ability of a person to access or retrieve the data from the projects.

5.3.1 Level of Data Access and producer information

Considering that collected data might be provided raw, i.e. as collected or in some form of spatial, temporal or thematic aggregation, 46 respondents (38%) indicated that their projects provide access to raw data, 45 (37%) to aggregates, and 30 (25%) to both (see also Figure 8).

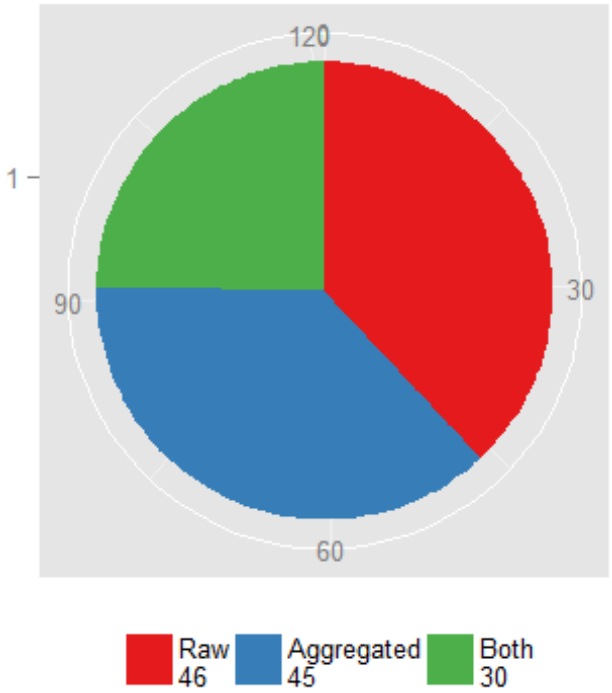

Figure 8: Answers to Question 9 - "Do you provide access to raw data sets or aggregated values?"

If the projects provide access to raw data (i.e. those 76 that have selected "Raw" or "Both" as reply to the previous question), we followed this up with another question about the inclusion of personal information (*Question 9.1 - "Do the raw data sets include information about the data producer (e.g. user names, individual identifiers, or locations)?" – single choice, optional*).In their replies:

- 23 participants claimed that this is not the case. This number corresponds to 30% of those project that provide access to raw data.

- 52 respondents indicated that their data included information about the data producer. This number corresponds to 68% of those projects that provide access to raw data and 43% of all projects that replied to the survey.

5.3.2 Management of Informed Consent

We were interested about the approaches used by the projects to get permission for data storage, collection and processing before conducting activities with people, i.e. informed consent. 118 projects informed us about their solutions; where 30 (25% of all participants to our survey) do not ask participants to provide informed consent, 64 (53%) said that they use a generic Term of Use (ToU) agreement, 10 (8%) ask participants to sign an explicit informed consent form, and 14 (12%) use other solutions (see also Figure 9).

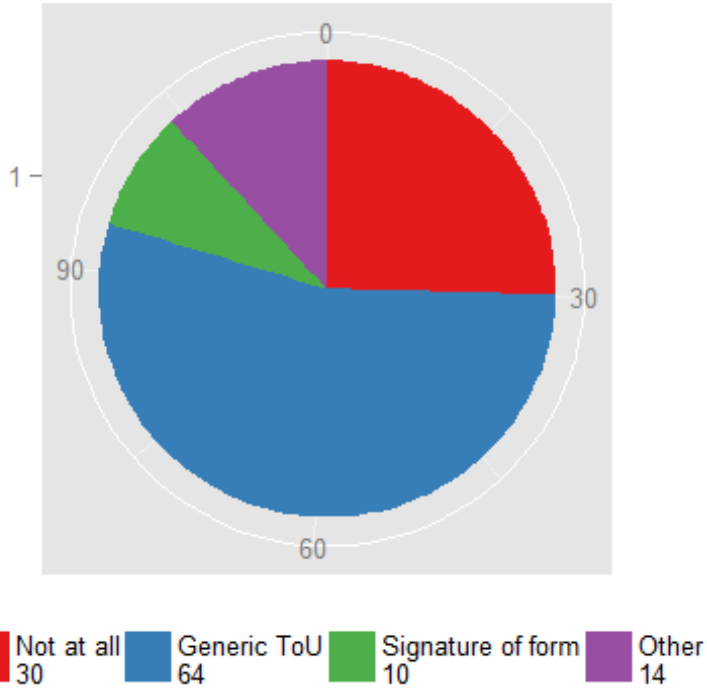

Figure 9: Answers to Question 10 - "How do you manage informed consent?"

Having asked what other mechanisms do manage informed consent (*Question 10.1 "Which 'other' mechanism do you use to manage informed consent?"*), survey participants provided the following answers:

- "Both generic terms of use and specific consent form when requested. Hence the many different IPR rules it is not possible to take a generic approach solely."
- "Data requests forwarded to data owner for consent."
- "University contracts granting consent to host researchers' data on website with different levels of privacy as required. Some data is open and some is not depending on the researchers involved for each project so some questions are a bit difficult to accurately answer. The citizen science aspect is new and being developed currently."
- "We have privacy policy but assume consent to share data as records are submitted to us."
- "Not applicable. Informed consent is not necessary for my project."
- "The goal is clear for all participants. We do not ask for specific consent about data use."

5.3.3 Means for Data Access

As far as data access is concerned, we first asked the projects about the availability of view services, i.e. possibilities to display their Citizen Science data, for example, in the case of geographically referenced data using a Web Map Service (WMS) possibly in compliance with the related standard of the Open Geospatial Consortium (OGC)[21]. *Question 11 - "Is the data collected by the project accessible via online services for visualization (e.g. by OGC Web Map Service - WMS)?" (single choice, optional*) was answered by 119 participants. 76 (63% of all participants) selected "Yes", 43 (36%) selected "No". Notably, these answers do not allow any conclusion on the use of the OGC WMS.

In a second question, we asked participants for details about how someone can download the data from the project. We separated bulk downloads (obtaining all data at once, e.g. via File Transfer Protocol - FTP) from user-customizable downloads of selected parts of the data (e.g. via the OGC Web Feature Service[22] - with Filter Encoding[23]) and the option for user-customizable services for computation (e.g. by providing a data processing Application Programming Inteface - API, so that data can be processed before it is delivered to the user). 106 participants answered this question; where 29 of them (24% of all participants) indicated the availability of bulk download of ther data; 43 (36%) said that they provide user-customizable downloads; and 34 (28%) state that they even provide computational capacities.

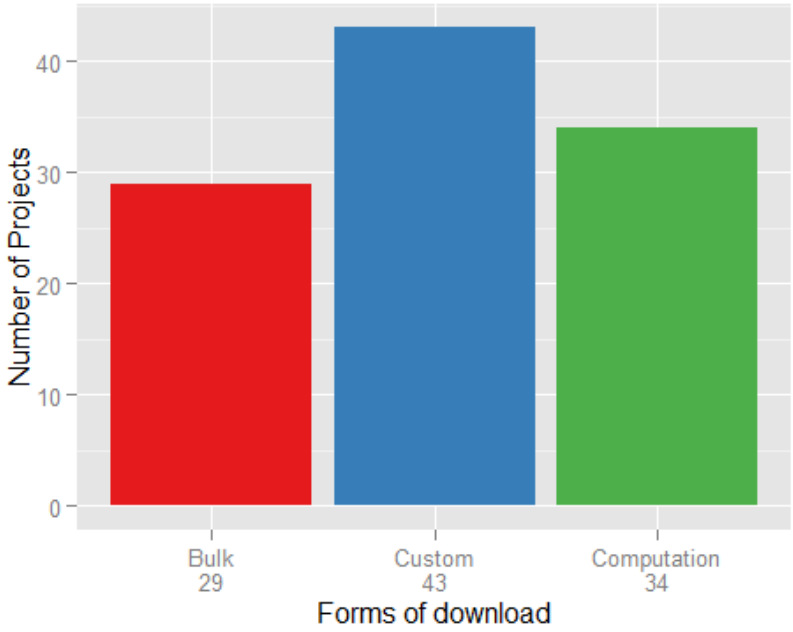

Figure 10: Answers to Question 12 - "If the data collected by the project is accessible via an online download service, how can it be accessed?"

5.4 Re-use conditions of Citizen Science data

Having explored discoverability and accessibility of Citizen Science data, we asked for details about that actual conditions for the re-use of these datasets, with a focus on the licensing approaches within the projects.

[21] The official web page of this standard is available from http://www.opengeospatial.org/standards/wms

[22] The official web page of this standard is available from http://www.opengeospatial.org/standards/wfs

[23] The official web page of this standard is available from http://www.opengeospatial.org/standards/filter

5.4.1 Data availability, re-use conditions and licensing

First of all, 90 of all (74%) indicate that they do make their Citizen Science data available. Only 20 (17%) indicated that the data should be directly available, whereas 70 (58%) use an embargo period, i.e. the use of the data remains restricted for a certain amount of time, usually to the project members, before it is released for wider re-use. 31 respondents (26%) said that the data from their project is not made available for re-use. Figure 11 provides an overview of these answers.

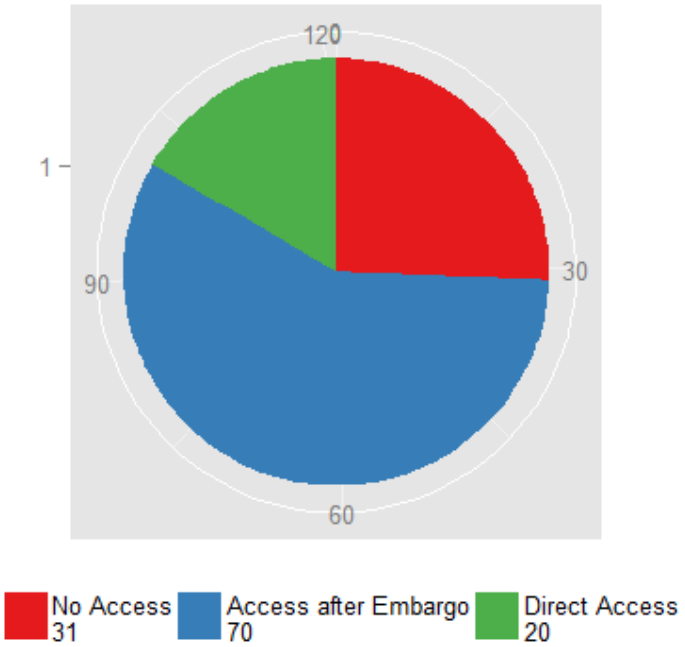

Figure 11: Answers to Question 13 - "Do you make the data available for re-use?"

For the two answers involving access, a further two more questions were asked in order to get a better understanding of the desired conditions for data re-use and the licensing approaches. As far as re-use conditions are concerned, we asked the participants to indicate re-use conditions followed, for example, those used by Creative Commons (CC)[24]. We distinguished re-use into the following:

- public domain, i.e. completely free from any restriction of intellectual property;
- with attribution, i.e. giving credit to the original creator;
- as share-alike, i.e. licensing derivatives under identical terms);
- non-commercial, i.e. allowing any re-use that is not of a commercial nature); and
- no derivatives, i.e. preventing any way of changing the original source or building upon it.

A total of 92 projects answered this question, several times selecting more than one option. The most popular condition was the public domain (46 selections, being equal to 38% of all participants to the survey). This was followed by re-use with attribution (33, 27%), non-commercial re-use (29, %), re-use as share-alike (23, 24%), and the restriction to no-derivatives (7, 6%) – see also Figure 12.

[24] A reference to the CC licensing scheme can be found here https://creativecommons.org/licenses/

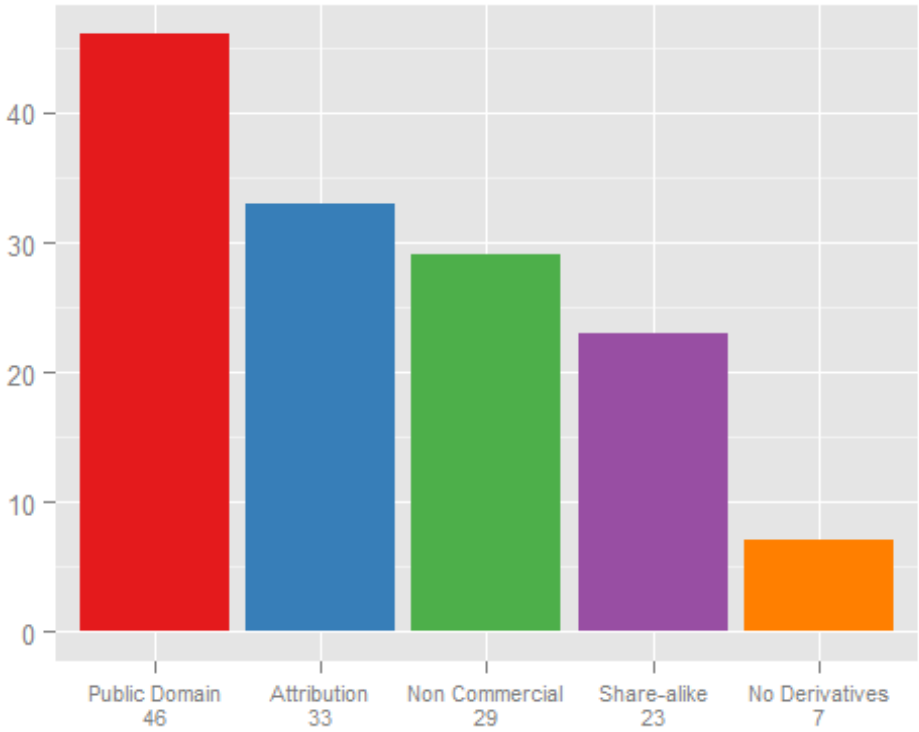

Figure 12: Answers to Question 13.1 - "Which are the conditions for re-use?"

An additional question about licensing (*Question 13.2 – "Which license do you use?"*) was answered by 56 participants. The free text answers can be summarized as follows:

- 2 said that they did not know how to answer this question.
- 5 participants said that this question is not applicable to their respective project.
- 7 participants said "None".
- 5 participants said that they were not yet decided.
- 4 participants indicated that they are in the process to establish a license.
- 4 respondents said Open Data Commons Open Database License (ODbL)[25].
- 1 respondent indicated Open Data Commons Public Domain Dedication and License (PDDL)[26].
- 1 respondent referred to the Open Software License (OSL)[27].
- The remaining 26 participants indicated various CC licenses.

5.4.2 Reasons for allowing or restricting re-use

Asked for the reasons why the participants of the survey decided to make their data available (or not available) under these conditions and licenses (*Question 14 – "Why did you take this decision?"*), we received the following replies (examples):

- Because of institutional policy (*this reply was given 5 times*).
- "Promoting the idea of Open Science."
- "Data is only useful if it is available - we want to attract scientists to use the data we collect."
- "We believe that scientific data should be available to all, not just scientific communities."
- "Motivate users to produce public data."
- "To give back to the community."

[25] See more at http://opendatacommons.org/licenses/odbl/

[26] See more at http://opendatacommons.org/licenses/pddl/#sthash.8SCe6WOC.dpuf

[27] See more at http://opensource.org/licenses/osl-3.0.php

- "We promised to the users we wouldn't share this data outside the research project."
- "We consider the volunteer, observer of biodiversity as owner of their records."
- We make the data available based on autonomous decisions by researcher-participants on a case-by-case basis.

5.5 Usability of Citizen Science data

Having explored the re-use options, we then moved to the potential usability of the Citizen Science data, where we considered aspects of quality control, the use of standards and the availability of metadata that helps others to better understand the nature of the available data sets.

5.5.1 Quality Control

Initially, we were interested if the data produced by the projects had some form of quality controlled. 97 participants (80%) replied positively about this question, while 24 (20%) indicated that their projects do not involve any quality control (see also Figure 13).

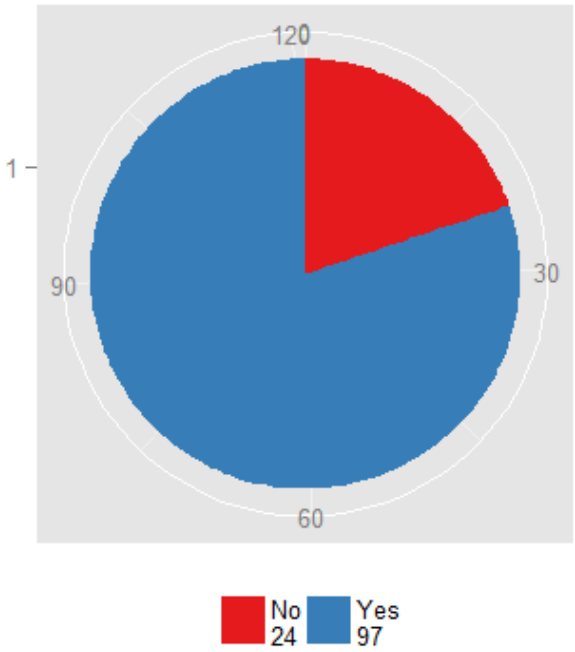

Figure 13: Answers to Question 15 - "Is the data collected by the project quality-controlled?"

In the following, 85 respondents (70% of all projects that replied the survey) provided details about their quality assurance procedures by answering *Question 15.1 "How do you control the quality of the data?" – semi-structured free text, multiple choice*. 58 (48%) provided details about quality control measures **before** data collection (e.g. by use of code lists, standards, metadata), 45 (37%) provided details about quality control measures **during** data collection (e.g. by use of atomized data, minimizing required entries, change documentation) and 75 (62%) provided details about quality control measures **after** the data has been collected (e.g. by checking for null values, checking for value assignments, statistic summaries). Several projects indicated that they have multiple measures in place.

Considering the measures for controlling quality **before** collection, the most prominent answers can be summarized as follows:

- Standards in equipment/instruments, data format, metadata, protocols/procedures/methods, identifiers (*replies like this were given 28 times*).

- Code lists, controlled vocabularies and dictionaries - enumerations, specifically designed field data sheets (*replies like this were given 20 times*).
- Training, tutorials, user guides, accreditation (*replies like this were given 12 times*).
- Automatic check before measurement is taken into account, e.g. check if values are off range (*replies like this were given twice*).

Considering the measures for controlling quality **during** collection, the most prominent answers can be summarized as follows:

- Automatic check, e.g. for duplicates or outliers, or redundant collections (*replies like this were given 10 times*).
- Continue training and supervision/counselling, FAQs (*replies like this were given 5 times*).
- Control questions and code lists (*replies like this were given 5 times*).
- Experts controlling required entries (*replies like this were given 5 times*).
- Multiple observations, self-management (*replies like this were given 3 times*).

Considering the measures for controlling quality **after** collection, the most prominent answers can be summarized as follows:

- Automatic validity and consistency checking, e.g. for outliers (*replies like this were given 28 times*).
- Expert review and validation - all or samples, e.g. when entering data for final system (*replies like this were given 28 times*).
- Checking for null values (*replies like this were given 6 times*).
- Peer/volunteer review (*replies like this were given 6 times*).
- Sample validation against reference data (*replies like this were given 5 times*).

5.5.2 Data standards

Having been asked about the use of data standards (*Question 16 – "Is the data collected by the project structured based on international or community-approved (non-proprietary) standards?"*), 67 (55% of all survey respondents) replied positively, 41 (34%) negatively, 13 (11%) did not reply at all.

5.5.3 Metadata and metadata standards

Considering the availability of data documentation, i.e. metadata, (*Question 17 – "Is the data collected by the project comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata?"*), 63 (52% of all survey respondents) replied positively, 49 (41%) negatively, 9 (7%) did not reply at all.

For those who replied positively, we followed-up with a question on the use of standards for their metadata records (*Question 17.1 – "Are these metadata based on international or community-approved (non-proprietary) standards?"*). 50 participants (41% of all survey respondents) replied positively, 11 (9%) negatively, 60 (50%) did not reply at all.

Independently from the questions about metadata, we directly asked about their provision of a contact point for data related matters (*Question 18 – "Are you providing a dedicated contact point for the data collected by the project?"*). 89 respondents (73% of all survey participants) replied positively, 25 (21%) negatively, 7 (6%) did not reply at all.

The metadata related questions were concluded by a question on the documentation of use conditions and licenses (*Question 19 – "Are the data access and use conditions, including licenses, clearly indicated?"*). Here, 57 respondents (47% of all participants) replied positively, 52 (43%) negatively, 12 (10%) did not reply at all.

5.6 Preservation and curation of Citizen Science data

Apart from current data access, (re-)usability and documentation, we were interested in long-term data provision from Citizen Science projects, including how access would be provided and for how long.

5.6.1 Data Access Guarantee

Having been asked about the relative time of ensured data access, 70 participants (58% of all survey respondents) replied that they intend to provide access beyond the life-time of the project; whereas 51 (42%) indicated that access will only be guaranteed within the project's life-time (see also Figure 14).

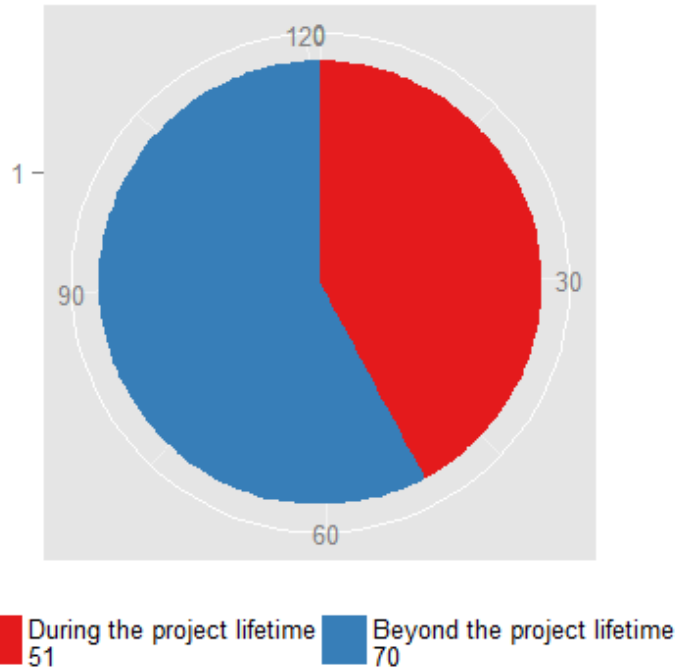

Figure 14: Answers to Question 20 - "For how long do you ensure the access to the data from you project?"

In the follow-up question to those who should have data access beyond the project ending (*Question 20.1 – "For how long (at least)?" – free text, optional*), 8 respondents (7% of all participants to the survey) intend to provide data access for one to three more years, 26 (21%) for up to ten years, 14 (12%) for ten to fifty years and 6 (5%) for more than a hundred years.

Asking about details on the applied storage facilities, we obtained 112 replies in total (Figure 15). 46 respondents (38% of all survey participants) indicated that they store their Citizen Science data on a remote server that is hosted by a project member ("PM" in Figure 15), 28 (23%) referred to a public repository, 19 (16%) said that they use a local machine of a PM.

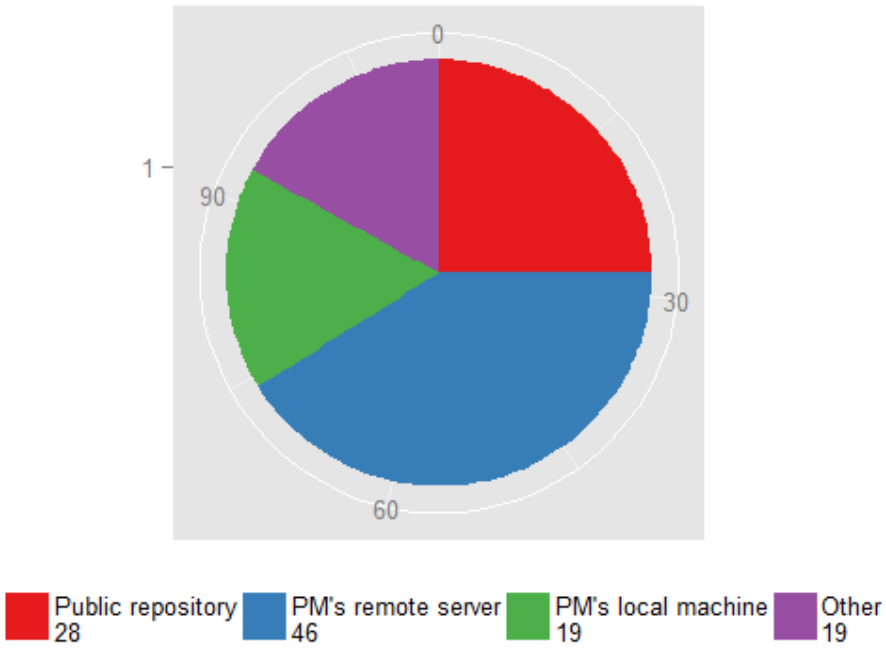

Figure 15: Answers to Question 20 - "For how long do you ensure the access to the data from you project?" - single choice, mandatory.

The 19 respondents (16% of the total) that selected other means of data storage were asked to provide more details (*Question 21.1 – "Which 'other' mechanism do you use?" – free text, optional*). Answers included (but were not restricted to):

- Data Centre (*this reply was given 4 times*).
- Cloud storage (*this reply was given 3 times*).
- Multiple levels, e.g. local, network and third party repository (*this reply was given twice*).
- Curated art exhibits (*this reply was given once*).

Last but not least, and to help understand some possible legal constraints, we asked about the actual country the data were stored in, where participants could provide more than one response. The replies are listed in Table 4 (below), indicating that 70 of the participating projects (58%) store at least parts of their data in Member States of the EU or in countries of the European Free Trade Association (EFTA), 23 (19%) in the United States of America, 20 (17%) in Australia, and 8 (7%) elsewhere . Figure 16 gives a visual impression about the answers.

Table 4: Answers to Question 22 - "In which country/countries do you store the data?" EU Member States and EFTE countries highlighted in grey.

Country	Frequency
USA	23
Australia	20
Spain	12
UK	10
Germany	8
Ireland	8
Canada	4
Denmark	4
Italy	4
Netherlands	4
Switzerland	4
Slovakia	3
Austria	2
Belgium	2
Poland	2
Argentina	1
Chile	1
Czech Republic	1
Finland	1
France	1
Greece	1
Hungary	1
India	1
Montenegro	1
Norway	1
Portugal	1

Figure 16: Word cloud summary of answers to Question 22 - "In which country/countries do you store the data?"

5.7 Additional Information

Towards the end of the questionnaire, participants were offered the possibility to provide further information about themselves and their project. As we did not ask for any consent to publish the related information, we only include generic or anonymized statements about the replies in this section.

5.7.1 Project Name

93 of all participants (77%) replied to the offer "*If you would like to share the name of your Citizen Science project for future reference, please use the box below.*"

5.7.2 Project Web Page

83 of all participants (69%) replied to the offer "*If you would be so kind to share the web page of your project, please use the box below.*"

5.7.3 Come-back email Contact

97 of all participants (80%) replied to the offer "*If you allow us to come back to you after the survey, please leave your e-mail address below.*"

5.7.4 Final remark(s)

46 of all participants (38%) replied to the offer "*If you have any additional information to share with us, please feel free to use the text box below.*"

Final comments from the participants were manifold. Apart from repeating issues that they…

- just initiated their project (commented 10 times),
- had challenges in answering some of the questions, e.g. difficult to understand or non-partial yes/no option (commented 6 times), and
- were not sure if their project is citizen science really (commented twice),

…replies included (but were not restricted to) the following statements:

- *"Getting data online and available is a costly venture with the time required a major limitation. We've had great difficulty working with IT professionals to ensure the data is made available according to our needs."*
- *"The survey points into a very clear direction, which is public access and highly documented information also beyond the lifetime of the project. This is a mandate that is made more and more often, but funding agencies and other donors are unwilling to pay for the tremendous effort that comes along with doing that."*
- *"Our project has been in operation for about 6 years. We are currently in the process of formalizing many of the data storage, documentation and access arrangements. The [project] is itself only a new organization and is still evolving its administrative structures. In the past 12 months these have progressed significantly with respect to data archiving and access. Our project will soon be formally incorporated into the government system, bringing with it large scale and long term archiving facilities, as well as permanent web access points. Thus some of the answers to the survey appear uneven with strange gaps in coherence. We are aware of these and should have them resolved within the coming year. For example, we are searching for a suitable template for metadata for our kind of project."*
- *"We are in the process of improving access to our data and metadata for this project. Some pieces are more accessible than others but nearly all pieces are accessible upon request."*
- *"We are particularly concerned about the quality and re-usability of data collected by means of citizen science projects."*

- *"I am in the process now of finding a place to host the data and I am having a very hard time finding a free online data portal that accepts biotic and abiotic data. Thank you for doing this survey, common databases are truly needed."*

6 Cross-question analysis

The separate investigation of questions provides only a starting point for the analysis but gives some first insights into the topic. A deeper understanding of the current data management practices in Citizen Science projects can emerge only when we examine the interplay between the responses to multiple questions in more detail. While there are many possible combinations, we restrict our analysis to a few topics that are of our major interest.

6.1 Investigations related to data granularity

Notably, a majority of the projects intended to make the collected data available for re-use, whether as raw data or in some form of aggregated data; more than half of the projects, however, apply an embargo period (Figure 17).

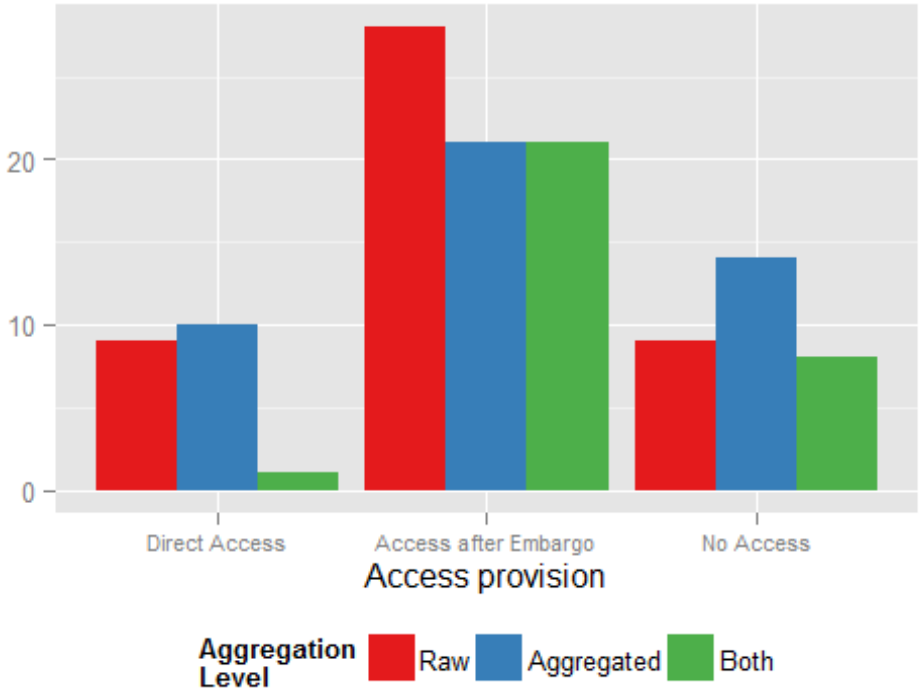

Figure 17: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 13 - "Do you make the data available for re-use?".

In general, the ability to download the data appears to be more frequently provided than view or discovery functionalities (Figure 18). This observation holds equally for projects that provide only access to raw data sets, aggregates or to both. The download of aggregated data appears less prominent compared to the other two forms.

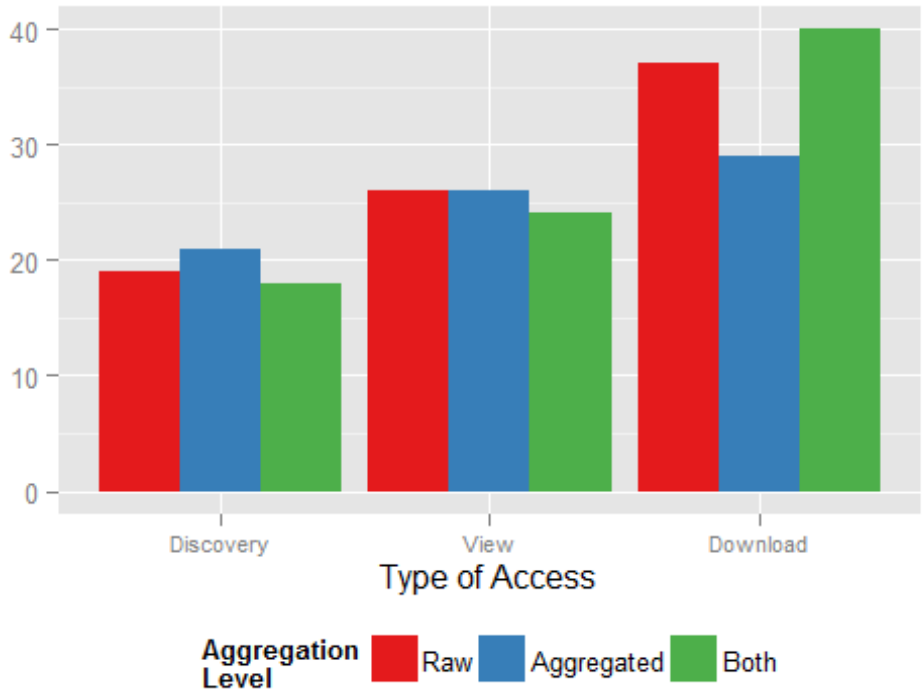

Figure 18: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and questions 7 (on discovery service), 11 (on view service) and 12 (on download service), where answers have been combined.

More than half of the projects intend to provide access to data beyond a project's life-time (Figure 19), favouring raw data access. Many projects already considering to offer access for more than four years after the project has ended (Figure 20).

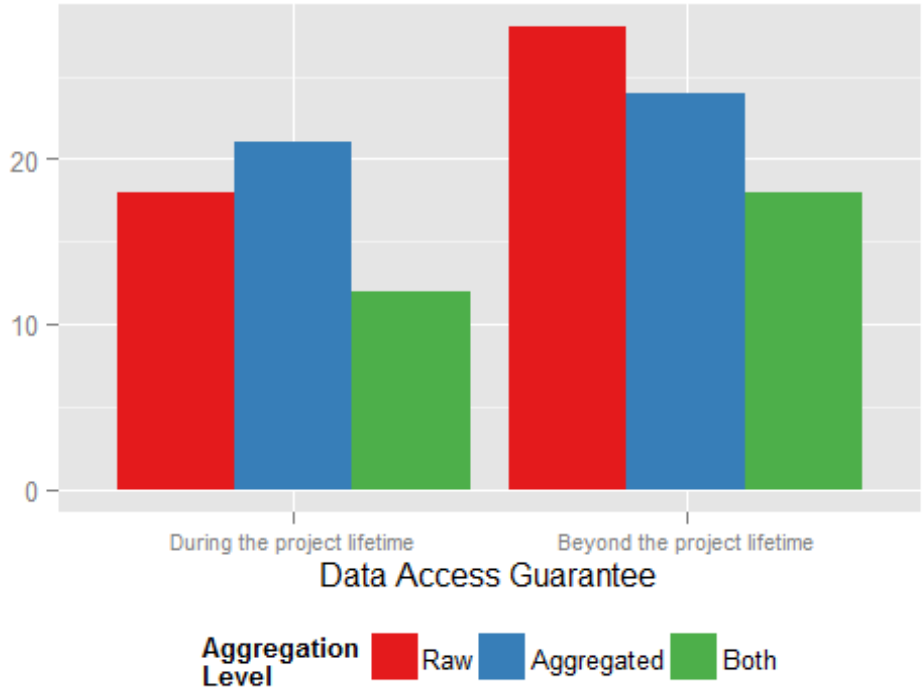

Figure 19: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 20 - " For how long do you ensure the access to the data from you project?".

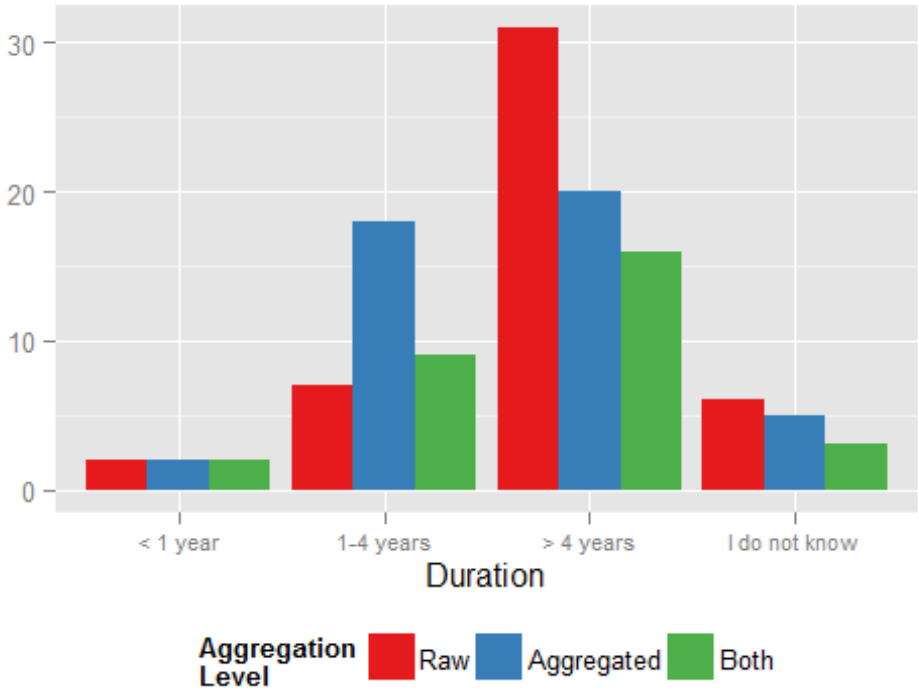

Figure 20: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 3 - " What is the planned duration of the project?".

However, considering the storage facilities of these projects, remote servers of a member of the project consortium (often the project manager) is a predominant solution. Notably, it is predominantly aggregated data that is stored on local machines, whereas public repositories are much more used for storing raw data.

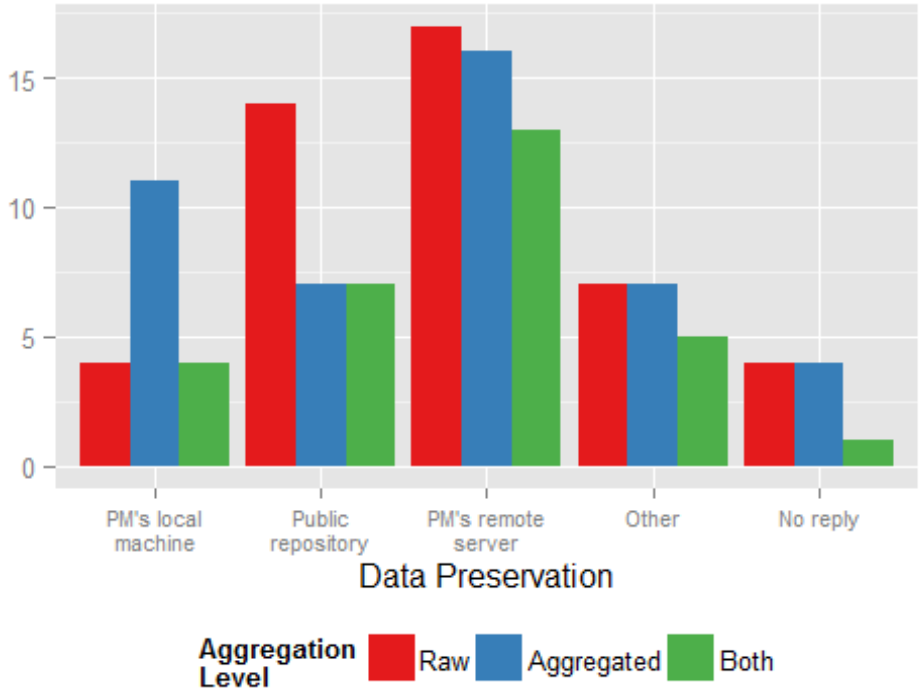

Figure 21: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 21 - " How do you preserve the data from your project?", where PM means Project Member.

When considered in relation to the duration of the project, these approaches to preserve the data sets seem to differ (Figure 22). Most notably, projects with a longer duration, especially those lasting more then 4 years, lean towards public repositories and remote servers. Projects that are set-up for a total duration of less than one year primarily indicated the use of local machines of a project member, where only one used a public repository. One participant selected "Other", but did not specify the type of storage facility used.

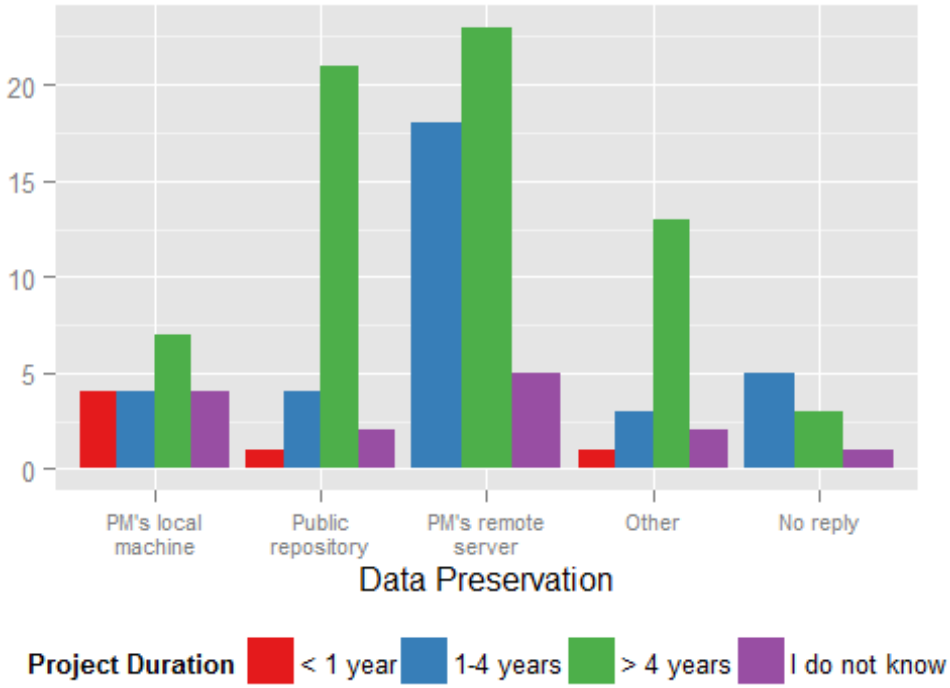

Figure 22: Combined analysis of question 3 - " What is the planned duration of the project?" - and question 21 - " How do you preserve the data from your project?", where PM means Project Member.

6.2 Findings related to Open Access

Following the earlier observation that a majority of projects favour open access – mostly after an embargo period – more than half of these projects provide access after the project ends (Figure 23). This option is favoured especially by those who provide access after an embargo period also prolong accessibility beyond the project life-time. 23 projects (19% of all respondends) preserve their data beyond the life-tine of the proejct, but to not provide (public) access.

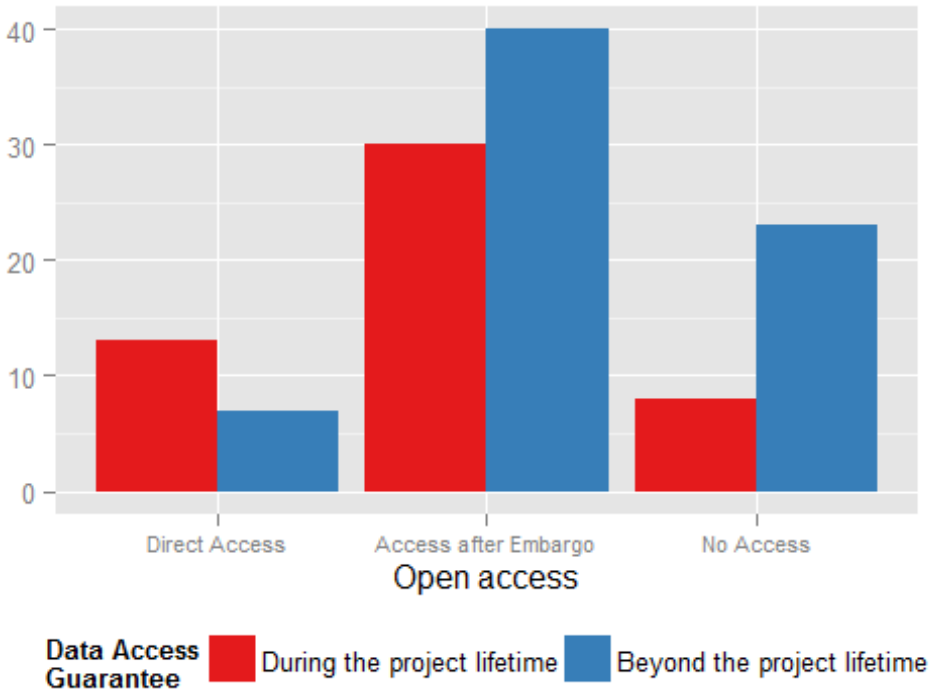

Figure 23: Combined analysis of question 20 - "For how long do you ensure the access to the data from you project?" - and question 13 - "Do you make the data available for re-use?".

Notably, not all projects that allow open access to their data also provide related documentation/metadata (Figure 24). 40% of the projects that do offer data access directly, or after an embargo period, replied that they do not provide comprehensive documentation.

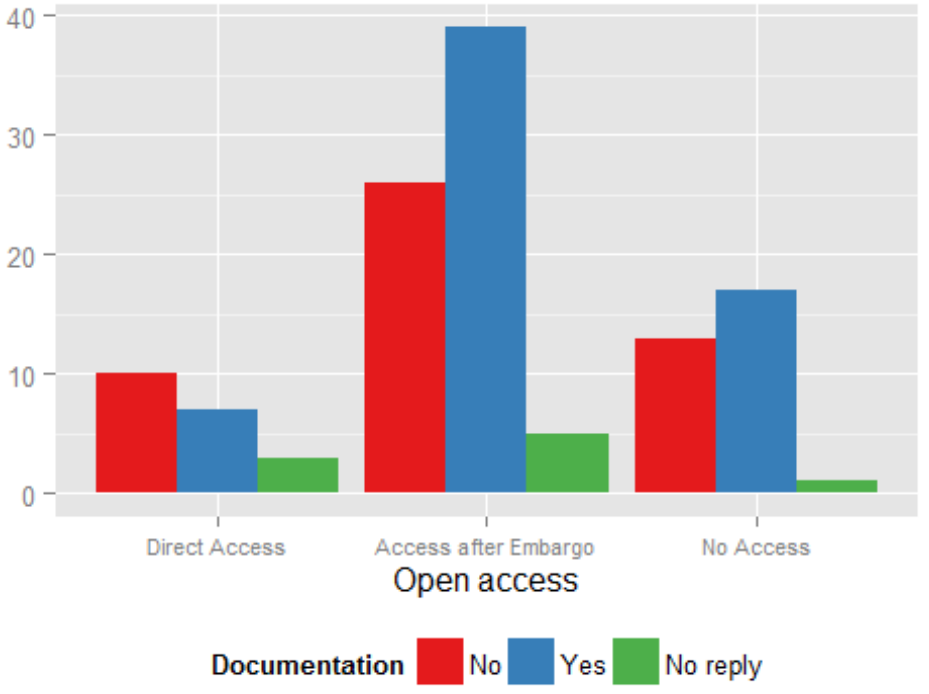

Figure 24: Combined analysis of question 17 - "Is the data collected by the project comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata?" - and question 13 - "Do you make the data available for re-use?".

A similar picture emerges when considering the provision of information about access and use conditions (Figure 25).). A little more than 40% of the projects that do offer data access directly, or after an embargo period, replied that they do not detail access and use conditions (including licenses) for their data.

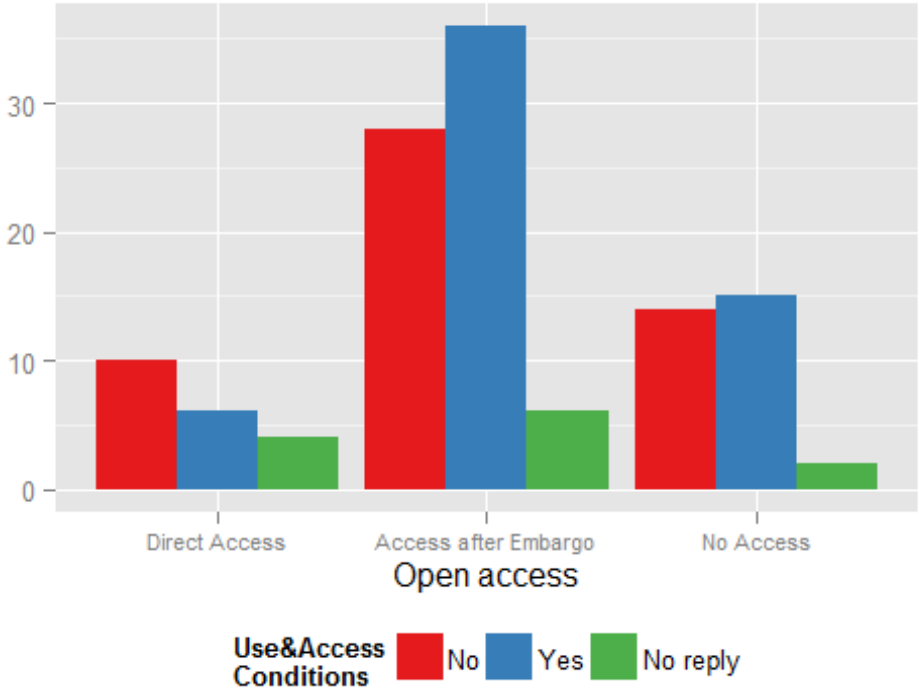

Figure 25: Combined analysis of question 19 - "Are the data access and use conditions, including licenses, clearly indicated?" - and question 13 - "Do you make the data available for re-use?".

Another important finding from the survey relates to the re-use conditions of the project outcomes. It should be underlined that a vast majority of projects apply an open approach, offering results in the public domain, with attributions or under share-alike conditions. Notably, few projects (less than a third) have already selected a license that supports this approach. Following further investigation we found that 14 projects even selected a license that is not in line with the intended re-use conditions.

6.3 Funding-related analysis

Funding seems to be another determining factor for data mamagement practices, where we noticed that the majority of respondents (two thirds) indicated a single funding source, whereas 33% are supported by multiple sources (Figure 26).

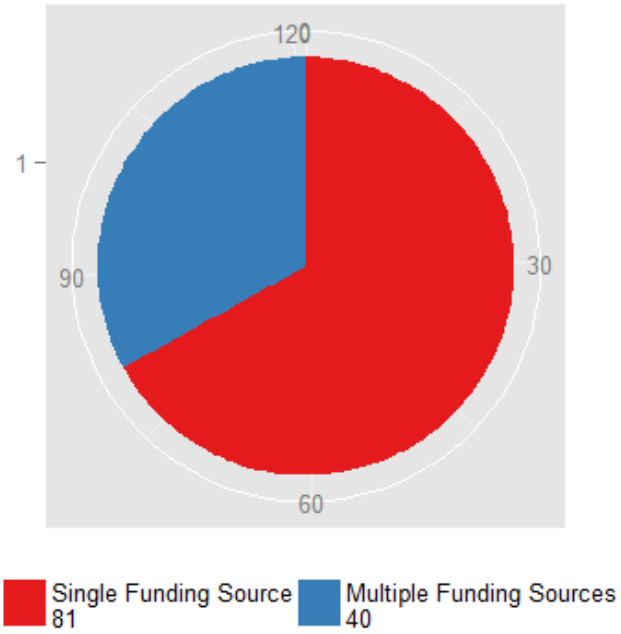

Figure 26: Analysis of question 5 - "How is this citizen science project financed?", where projects with a single source of funding have been distinguished from those receiving funding from multiple sources.

Separating the 23 projects that contain at least a partial contribution form the EU from those (98) that rely on other funding mechanisms, we find an almost equal order in the overall separation across topics (Figure 27). Only the funding of other topics than Earth science, environmental science, social sciences and space science appears more prominent for non-EU bodies. Aside being the most popular topic, environmental science receives relatively more funding from other sources than the EU. Given the low numbers of (at least) EU-funded projects in the survey this and the following statements can only be considered indicative.

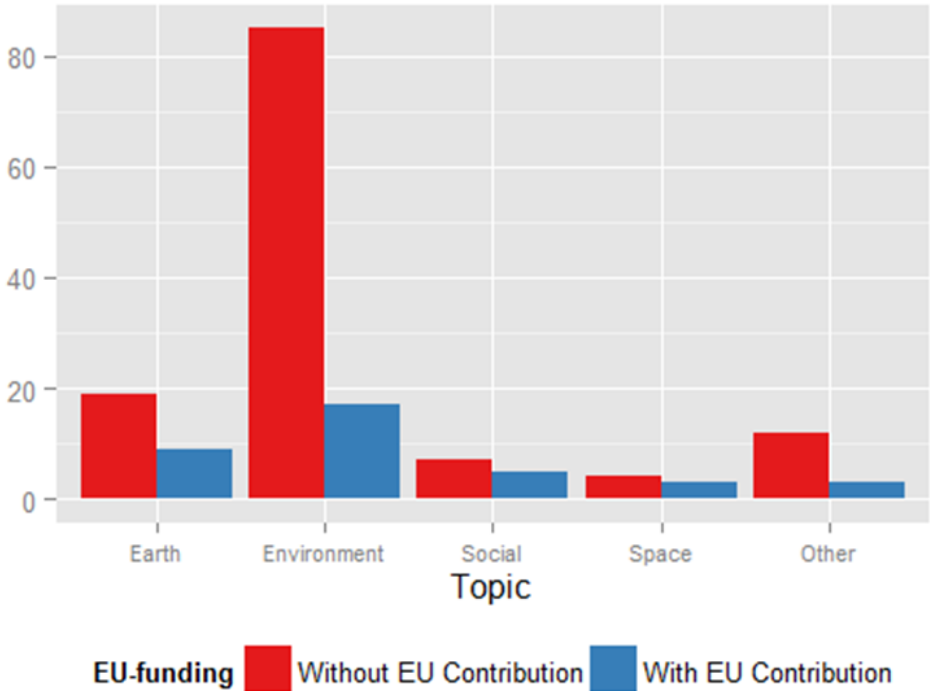

Figure 27: Combined analysis of question 1 - "Which topic areas does your project cover?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution.

The separation along different thematic areas within the environmental topic provides details (Figure 28). Compared to projects without any EU contribution, air quality appears to be a particularly prominent topic for EU funding (35% compared to 9%), whereas the contrary holds for biodiversity (22% compared to 63%). The defined themes covered all projects with the involvement of EU-funding, while 20 projects without EU contribution address other themes.

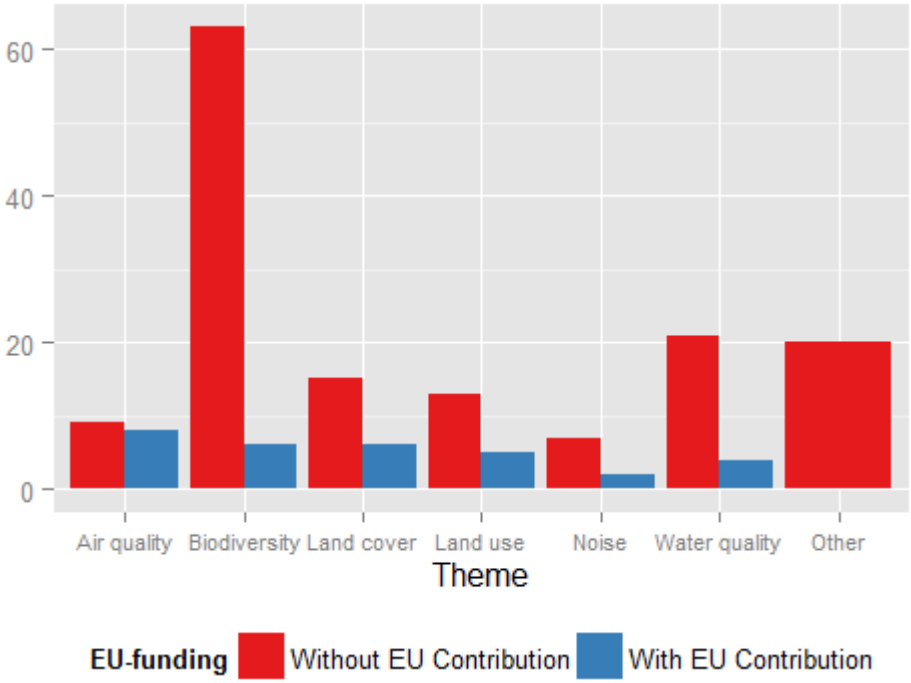

Figure 28: Combined analysis of question 13 - "Do you make the data available for re-use?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution.

Shifting the attention to the relationship with data provision under open access, we find the same order of categories, but considering the relative distributions less direct access and more closed data sets when it comes to funding of the EU. (Figure 29).

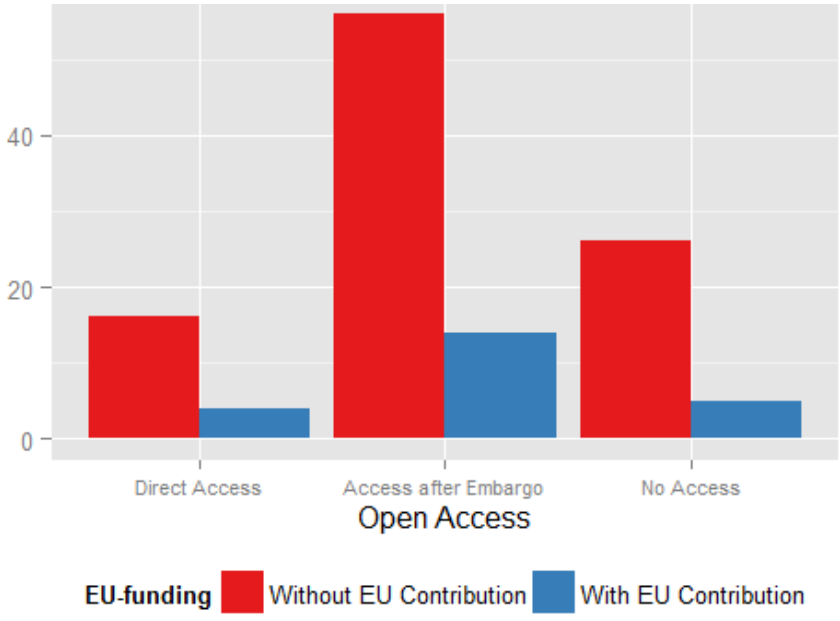

Figure 29: Combined analysis of question 13 - "Do you make the data available for re-use?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution.

Considering those participating Citizen Science projects who have explicit data management plans compared to those who have not explicitly include such plans, we do not see a slight tendency in the relationship to EU finding (Figure 30). 65% of projects receiving some funding of the EU do have a data management plan, as compared to less than 60% of the projects without any EU contribution.

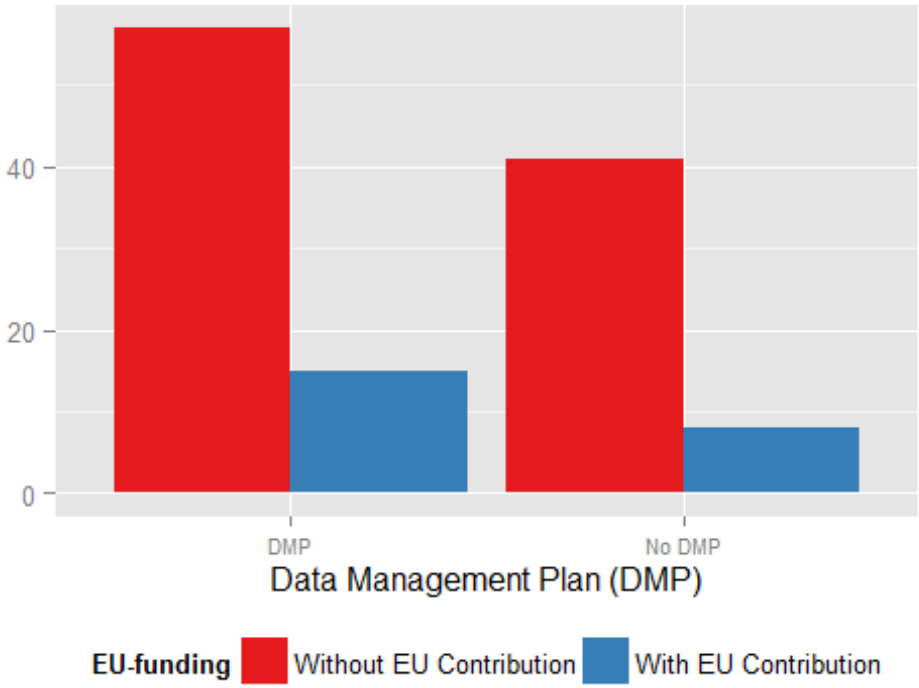

Figure 30: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution.

6.4 Relations to Data Management Plans

Examining the availability of a data management plan further, we tested the independence between the plans and the provision of data as open access (using the chi-square tests). The assumed independence could not be rejected. The decision of data access provision appears to be independent from the availability of a data management plan (Figure 31).

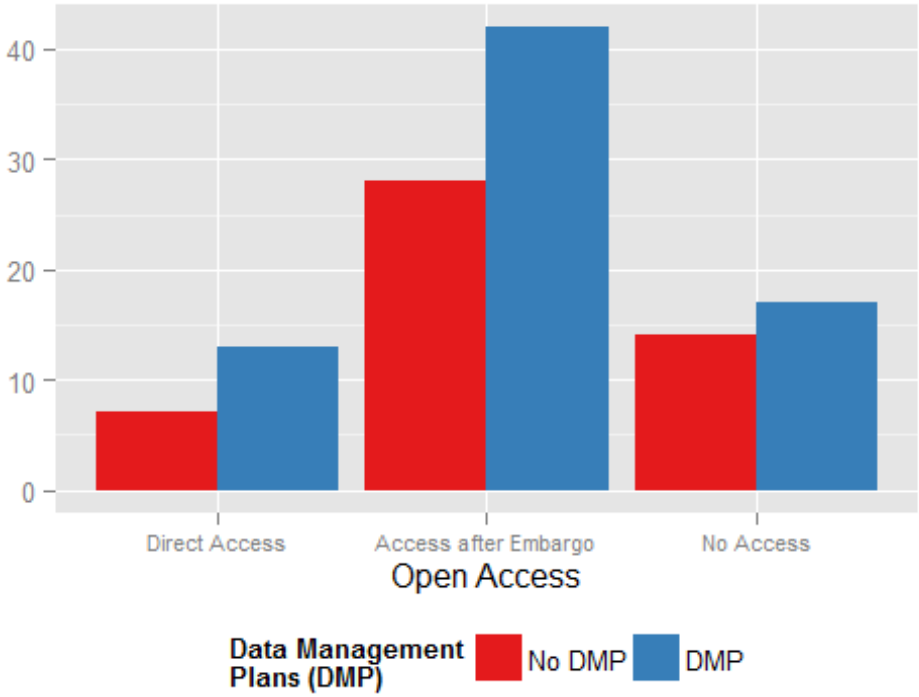

Figure 31: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 13 - "Do you make the data available for re-use?".

57% of the participating projects that said they had a data management plan also provide metadata, while it is only 45% for the projects without such a plan (Figure 32). Still, more than 20 projects that have a data management plan in place do not document their data comprehensively.

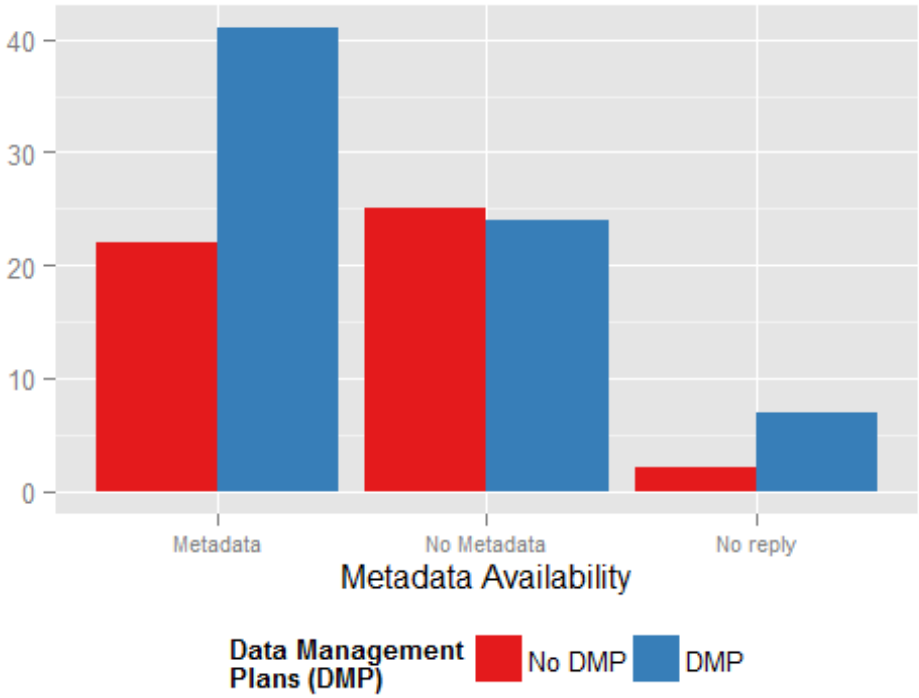

Figure 32: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 17 - " Is the data collected by the project comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata?".

A similar picture appears considering the use of standards for these metadata records (Figure 33). Although there was a considerably lower response rate, the 36 projects having a data management plan also indicated their application of metadata standards (this corresponds to 50% of all projects with a data management plan), where 14 projects not having a data management plan indicated the application of metadata standards (this corresponds to 29% of all projects without a data management plan). Only 4 projects (less than 6% of all projects with a data management plan) indicated to have a data management plan but do not use any metadata standard. Apart from addressing the noise or air quality thematic, we could not find any commonality between these four projects.

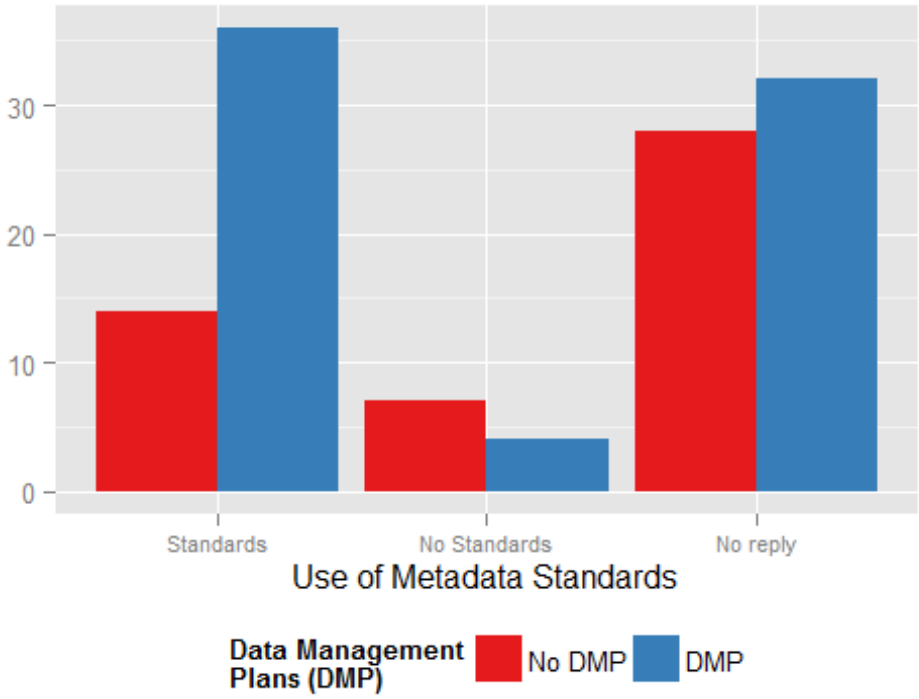

Figure 33: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 17.1 - "Are these metadata based on international or community-approved (non-proprietary) standards?".

The situation for metadata standards can also be projected to the use of standards for data production data (Figure 34) - this time with higher response rates where standards are being used (85% or all projects with a data management plan and also 85% of projects without). 44 projects having a data management plan also indicate the application of data standards (this corresponds to 61% of all projects with a data management plan) and 23 projects not having a data management plan indicate the application of metadata standards (this corresponds to 46% of all projects without a data management plan).

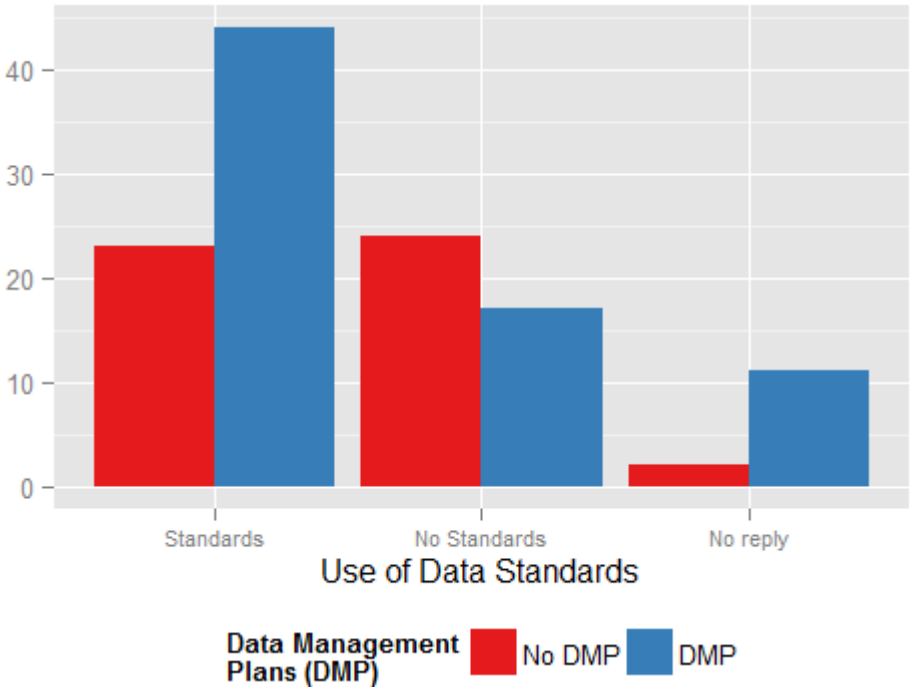

Figure 34: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 16 - "Is the data collected by the project structured based on international or community-approved (non-proprietary) standards?".

Figure 35 illustrates the combination of these last two observations. It clearly indicates that a large amount of projects applying both data and metadata standards (point cloud on the lower-left part of the diagram) and that most of these do have a data management plan in place (green dots). This illustration also indicates that there are very few projects that appear to apply either data or metadata standards, but not the other.

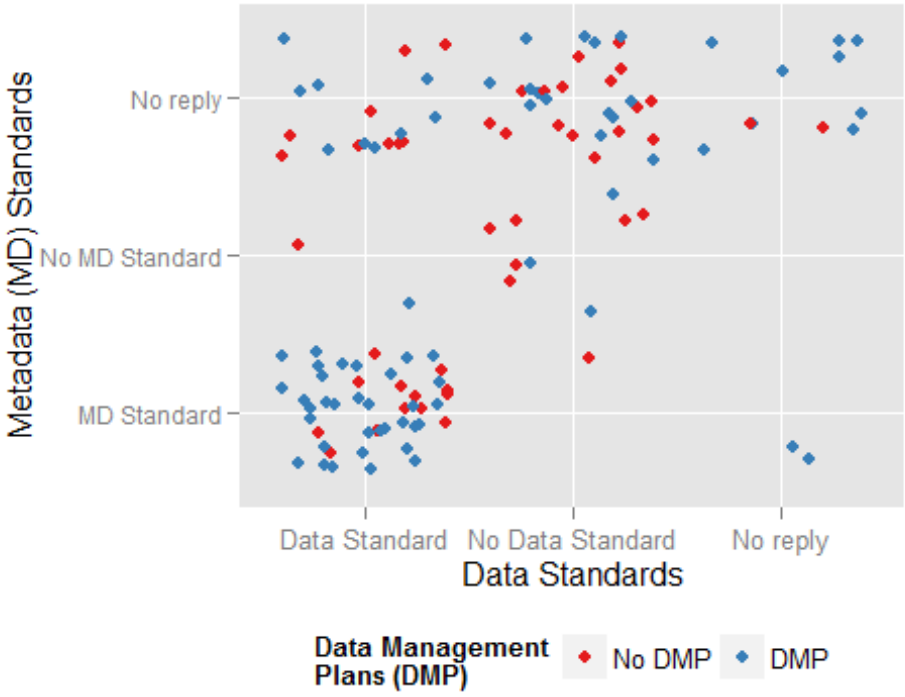

Figure 35: Combined analysis of question 6 - "Does your project include an explicit data management plan?", question 17.1 – "Are these metadata based on international or community-

approved (non-proprietary) standards?" and question 16 - " Is the data collected by the project structured based on international or community-approved (non-proprietary) standards?".

6.5 Characteristics of projects running for more than four years

More than half of the responding projects are set up for a period of more than four years (see also Section 5.1.3 and Figure 6). Some of their main characteristics deserve further examination.

First of all, it can be observed that most (93%) of the projects that are set-up for more than four years run in an operational setting. As a comparison, only 74% of all responding projects indicated they were operational.

In particular, 61% of all projects in the environmental domain appear to be set-up to run for more than four years (Figure 36), followed by 46% of all participating projects that deal with Earth science.

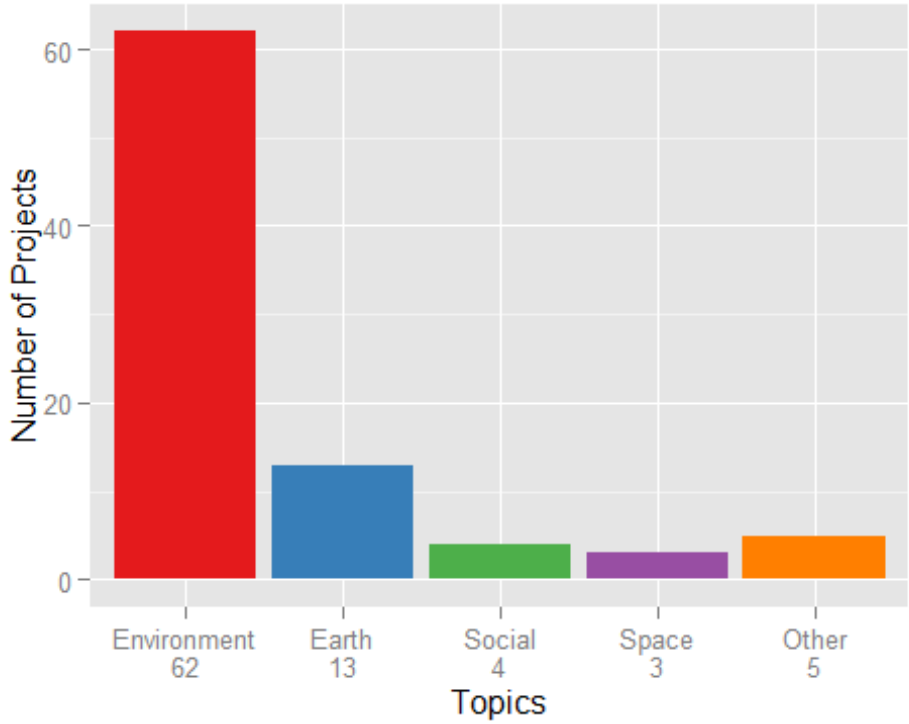

Figure 36: Analysis of Question 1 - "Which topic areas does your project cover?" restricted to projects that are set up for more than four years.

While the overall ratio between projects running less than four years and those running longer (see also Figure 37) resonates across most of the environmental themes, it is particularly low in the cases of air quality (24%) and higher for biodiversity (71%) and 'other' topics (80%).

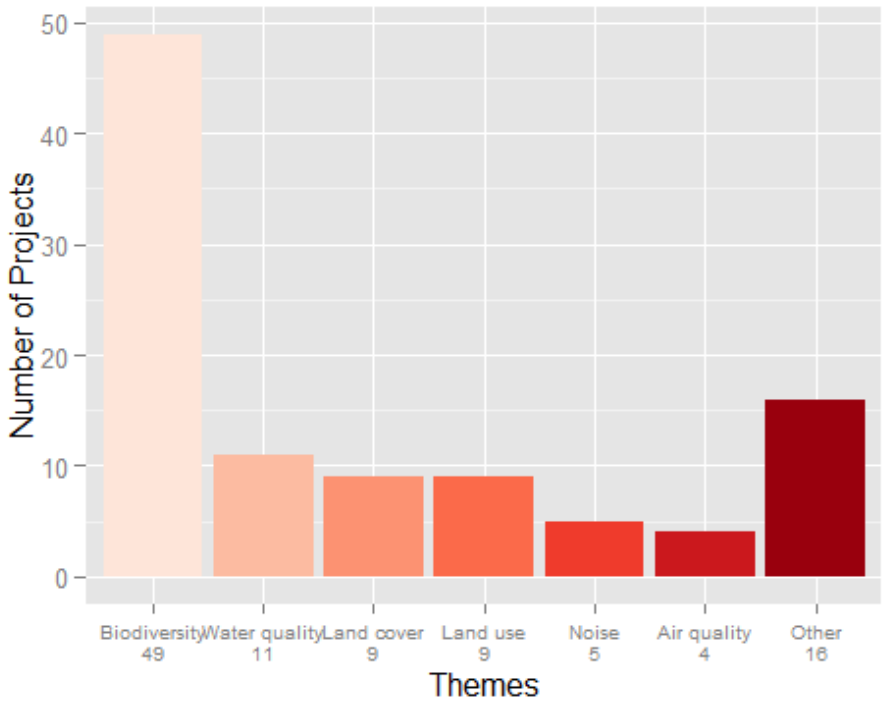

Figure 37: Analysis of Question 1.2 - "Which environmental theme(s) does it cover?" restricted to projects that are set up for more than four years.

Considering the funding model, about half (33 out of 67) of the projects running longer than four years use multiple funding sources. This is a larger number compared to all projects, where only 20% build on multiple sources (see also Figure 26). In terms of specific sources, fewer projects running for more than four years are funded from EU budget lines, as compared to the projects with a shorter time line (Figure 38). By comparison, donations and direct fees are more popular for longer projects, as compared to those during at most four years.

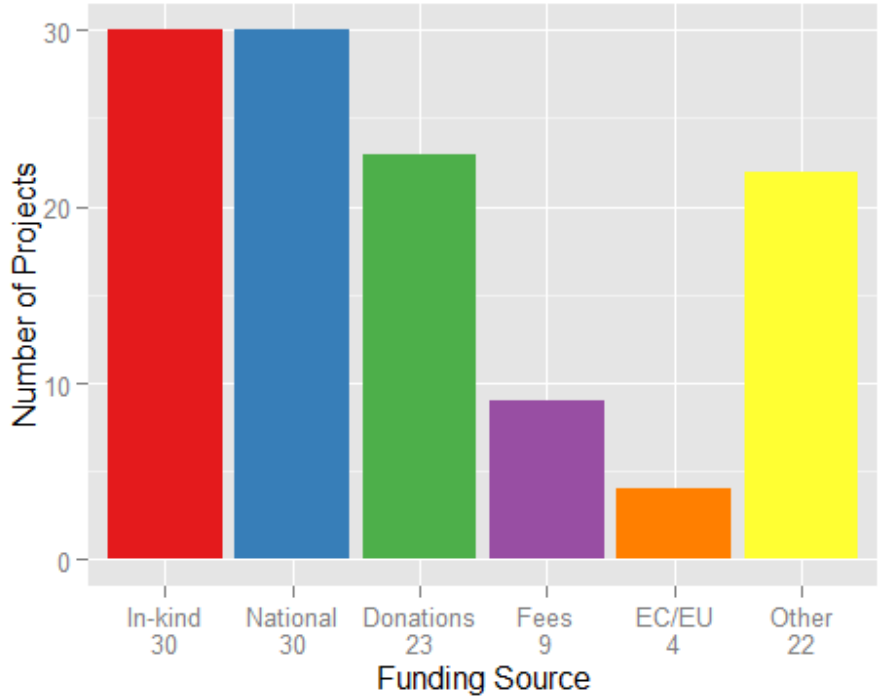

Figure 38: Analysis of Question 5 - "How is this citizen science project financed?" restricted to projects that are set up for more than four years.

In terms of geographical extent, we found that more long-term projects are outside the EU than inside (Table 5), with the percentage of projects increasing with scale from local to continental projects.

Table 5: Table 3: Answers to Question 2 – "Which geographic extent does the project cover?" including projects that are set up for more than four years (numbers on the left in each cell) and percentage.

Extent	Inside EU	Outside EU
Neighbourhood	8/22 (36%)	15/24 (62%)
City level	10/29 (34%)	19/30 (63%)
Regional	14/31 (45%)	24/39 (62%)
Country	19/37 (51%)	20/30 (67%)
Continental	16/27 (59%)	18/26 (69%)
Total coverage	**67/146 (46%)**	**96/149 (64%)**

6.6 Selected details about projects with long-term data curation

It would appear that there are more projects offering long-term data curation in the environmental sciences (Figure 39), as 59% of participating projects that deal with environmental sciences promise long-term data access, compared to 51% being associated to topics other than environment.

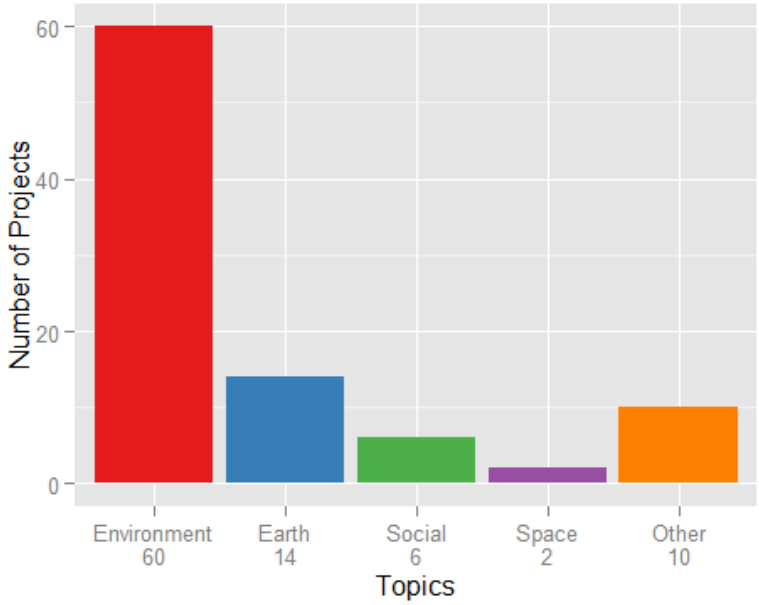

Figure 39: Analysis of Question 1 restricted to projects that store the data also beyond the life-time of the project.

Many environmental Citizen Science projects offering long-term data curation are coming from biodiversity related themes (Figure 40), with 65% of participating projects on biodiversity promise long-term data access, compared to 45% being associated to themes other than biodiversity.

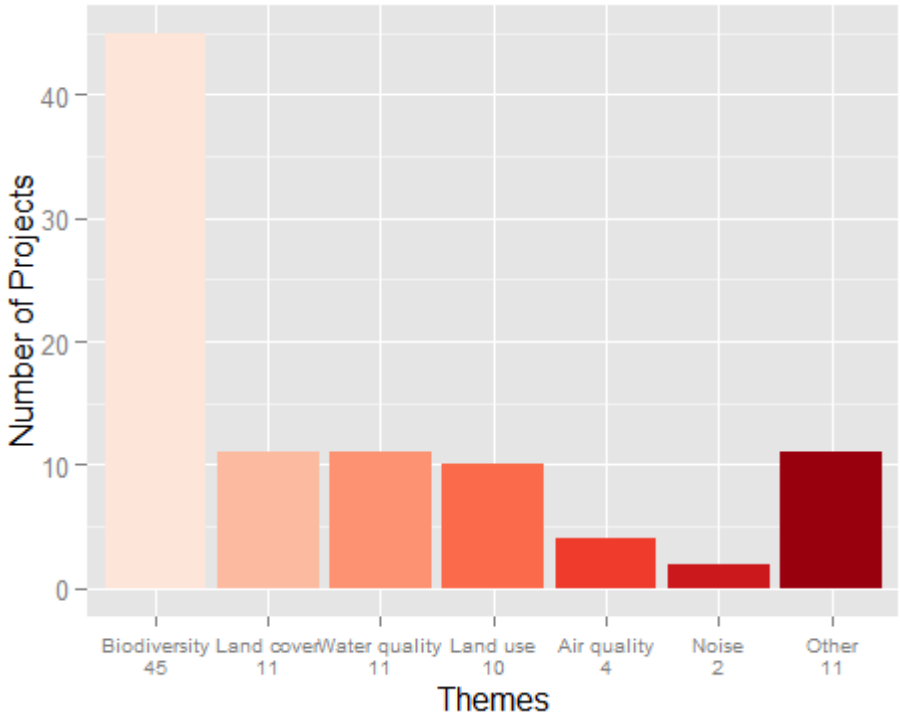

Figure 40: Analysis of Question 1.2 restricted to projects that store the data also beyond the life-time of the project.

7 Discussion of survey results

Before interpreting the results, we have to note that – as for many surveys – the responses may not be taken as statistically significant but they do offer some insight into the current situation, including the intentions of a range of different actors in this context thanks to the open nature of the survey. This may create some sample bias, as there were a high number of respondents from the environmental field (Figure 4) and especially those dealing with biodiversity-related matters (Figure 5). We can, however, also note that there is also a long tradition of citizen participation in data-gathering in this domain and would expect a large number of responses from them.

From the response, it would appear that relatively few biodiversity-related projects receive direct funding from the EU (Figure 28). Several other environmental themes – here especially including different aspects of environmental pollution, as well as, mobility and transport – appear only to be supported by other funding mechanisms (Table 2). Overall, independent of funding sources, the survey managed to reach many projects operating within the EU and at diverse geographical scales (Table 3).

We are (positively) surprised that more than half of the responding projects are set-up to run at least for the medium-term, i.e. for a duration of more than four years (Section 5.1.3). Not unexpectedly, almost all are in an operational (i.e. not in an experimental) stage (Section 6.5), where almost three quarters of all biodiversity projects are included. The high contribution of national funds is noteworthy. It would also appear that, in terms of percentages, projects outside the EU are leaning towards longer durations than those within. Within the EU, we can see that projects with larger geographic extents seem to be set-up for longer periods.

A second important finding is the extensive use of persistent identifiers (Section 5.2.2). However, our questions did not cover any details about concrete implementations. This needs to be investigated further, including what projects understand as persistent, unique, and resolvable identifiers, how they guarantee their availability, and which approaches are used inside individual projects.

Moreover, we saw that the related question on discoverability (Section 5.2.1) split the respondents in two almost equal parts. In fact, download and, then, view functionalities for data seem to be more favoured than data discovery capabilities. While this might be due to the intended audience of the data – project participants as opposed to third party (re-)users – it also indicates a need to promote possibilities to increase the discoverability of data sets by both mainstream search engines and more specialised catalogues. At the same time, these results raise a series of questions, including: "Should a core data management principle for Citizen Science projects be to have all project data discoverable?"; "Which catalogues should be used for Citizen Science projects?"; and "Which catalogues are currently available?". In respect to view services we did not investigate the provided functionalities (e.g. tables, maps, diagrams) or specific technologies used. It might be worth to investigate best practices depending on the kind of data gathered by the various activities.

In terms of accessibility, projects provide (or intend to provide) access to raw or aggregated data – if not both (Section 5.3.1). Also the means being used for data access show that projects are use existing technologies. This observation could be important, as not all available data sets will be as large as to require filtering capabilities, and as not all projects focus on the re-use of the collected data for further development. Those projects may not also be in need of (standards-based) processing APIs close to the data. Nonetheless, we did ask respondents only very little about any ethical or legal implications of data-sharing – e.g. the country in which the data is gathered or stored (Section 5.6.1) and the current handling of informed consent (Section 5.3.2). Responses indicate a need for further awareness-raising of these and related issues and to promote best practices that address them across Citizen Science projects, in turn potentially helping data to be more re-usable.

This could be important as three quarters of the projects indicated an interest in data re-use, mostly after an embargo period and especially when it comes to raw data (Section 6.1). It can be seen that a vast majority of projects apply an open approach, either following institutional policies or more generally endorsing the concept of Open Science (Section 5.4.2). This also illustrates an overall positive attitude towards openness inside the Citizen Science community. While particularly intending to provide results in the public domain, only few of the projects (less than a third) have already selected an appropriate license (Section 5.4.1). Some projects also selected a license that may not match their intended re-use conditions. In addition, the free text answers provide further evidence for such a weakness, e.g. by considering that the question related to licenses was not applicable to a respondent's particular project (Section 5.4.2). We, therefore, believe there is a need for further awareness-raising in two areas; firstly, on the overall issue of re-use conditions and, secondly, on the use of appropriate licenses and the fact that the lack of a licence does not necessarily mean people are 'free' to use the data. It may also be useful for further research to explore how open and re-usable results are then picked-up by third parties, especially in relation to the availability (or not) of data documentation (see also Figure 24). Here, the lack of documentation creates issues of long-term data access, as metadata can also help in-house storage and maintenance of a dataset, as well as providing details to users. Matters of liability and tractability of data might be further investigated to complement these findings.

Only one fifth of all the projects do not apply any form of quality control (Section 5.1.1). Unfortunately, we missed to include any follow-up question that would have helped to clarify if this is due to scoping issues (e.g. projects concentrating on awareness-raising) or if the participants have been informed about the absence of any quality assurance procedure. A project that, for example, intends to raise public awareness about domestic energy consumption might choose to measure consumptions without much concern for accuracy, but the participants should be properly informed that they should not challenge their local energy provider with the collected data. The additional information about quality assurance mechanisms that are used before, during, and after data collection (still in Section 5.5.1) underline the overall diversity of projects and the

dependency of quality measures from project motivations, scopes and methods. Still, a few common approaches can be identified. Before collection, those primarily include the use of standards, code lists and training; during collection they encompass automatic checks and continuous scientific and technical support; and after the collection they entail automatic validations and expert review. As a first indication, standards for data and metadata are supported by approximately half of the respondents. The strong positive relation to data management plans (Figure 35) is remarkable. This indicates room for improvement and experience-sharing, but the survey has not specifically looked at the kinds of standards that are being used. We still require a better picture that identifies the diversity and possible relationships between current approaches in terms of standards and technical infrastructure, where we expect variation between thematic communities. Moreover, the pure provision of contact point information and clear indication of use conditions and licenses has to be further promoted.

The replies to the question about data access guarantees (Section 5.6.1) also reveal a clear need for action. Almost 60% of the participating projects have the intention to provide access to the results beyond the end of the project, whereas we are (and will keep) losing the results of many others. While this might not appear as an issue for a particular project, the community risks losing the 'long tail' of Citizen Science data that could be the basis for new (also longitudinal) research. Already the opportunity to reflect upon the 'early days' of technology-supported Citizen Science and its evolution requests a solution.

Ultimately, the above interpretations suggest public access to highly documented information, also beyond the lifetime of a project (see also Section 5.7.4). This raises two important questions: who has the mandate to do this and which donors are willing to support such a long-term commitment? Future suggestions and decisions might be informed by the approaches of those 70 activities that plan to host their data beyond the life-time of their projects, especially those that favour access to raw data (Section 6.1). This phenomenon is not only likely to be due to the topics the projects are addressing. Biodiversity-related projects, which typically deal with individual occurrence records and build on a long tradition of citizen involvement, are leading the related statistics from the survey- both absolutely and relatively (Section 6.6). However, projects dealing with other environmental themes and with other research topics are not far behind. Some of these cases, however, do not foresee any long-term curation (Section 5.6.1) and not all of them promote open data-sharing approaches (see also Figure 25). Figure 21 and Figure 22 suggest that combinations of institutional remote services and public repositories might be part of a solution to more accessible data from Citizen Science projects but the relatively low numbers in this context may indicate that there could be some alternatives requiring further exploration. At the end of the day, the sustainability of data storage and access infrastructures will depend on the use of the data independent of its originating project and purpose.

8 Conclusions from the survey

If the survey respondents are found to be representative, then it would be possible to say that the environmental sector (especially biodiversity) identifies itself most with the 'Citizen Science', especially as notable proportion of replies from related fields cannot only be explained by the dissemination channels that we applied. Land use/cover, water and air quality could also be seen as prominent topics, whereas we found significantly fewer responses related to noise and different forms of pollution. Accordingly, any new activity in the above-mentioned areas might already directly benefit from the re-use of existing methods and tools in recent or ongoing projects, potentially helping to shape good practice related to data management within and across some themes. Although the survey did not explicitly seek details about existing collaborations between projects, we see these forms of knowledge exchange, especially within topic areas, as an area for future work. Knowledge exchange across different application areas will be more challenging but should be further exploited in parallel.

The responses to the survey clearly demonstrated the diversity of projects in terms of topicality, funding mechanisms and geographic coverage. They also provided valuable insights related to the access and re-use conditions of project results. While, for example, 60% of the participating projects follow a dedicated data management plan and a majority of projects provides access to raw or aggregated data, the exact use conditions are not always put into place and, in some cases, well-defined licenses are missing. This would appear, therefore, to be an area requiring some improvement. In order to complement and build on the findings from this survey, we require concrete cases that can help illustrate the way ahead, where we have identified the following needs:

- train and support Citizen Science projects on data management plans (especially including long-term access and sustainable data curation);
- help Citizen Science projects to explore data and metadata standards to make data more re-usable and to adopt tools that follow standards so time is not spent building custom-built solutions;
- raise awareness about the variation in ethical and legal issues related to data gathering, sharing and storage depending on countries and cultures and promote best practices for addressing them;
- educate projects about issue associated with liability and traceability, and with related re-use conditions and the use of appropriate licenses;
- raise awareness about the need for clear indications of the use conditions and licenses for project results, especially data;
- encourage the provision, where appropriate, of contact point information
- promote possibilities to increase the discoverability of data by mainstream search engines and more specialised catalogues; and
- exploit existing (and potentially establish new) possibilities for long-term data curation and sustained access provision.

In the final part of the survey, each participant was encouraged to provide a closing remark, one of which resonated many times "*Thank you for doing this survey, common databases are truly needed"*. Although the architecture that would be related to the required data storage and processing is not yet clear (as service oriented or centralised approaches may have different benefits and costs for projects), this comment and the survey as a whole do provide us with the impetus for more detailed investigations on how to prevent knowledge loss from EU-funded Citizen Science projects. Following this survey we will continue to explore related knowledge re-use, creation and sharing in Citizen Science Platforms, and to seek deeper collaborations with a broader Citizen Science community.

If this survey should be followed by another round or a complementary consultation activity, the following open questions might be considered:

- Persistent identifiers: What projects understand as persistent identifiers? How do they guarantee their availability? Which approaches are used inside the individual projects?
- Discoverability: Should it indeed be a core data management principle for Citizen Science projects to have all project data discoverable in the first place? Which catalogues should be used for Citizen Science projects? Which catalogues are available in the first place?
- Ethical and possible legal issues: What are the ethical and legal implications of data gathering and storage in relation to different countries and cultures? Do you (need to) store the data in the country in which it is collected?
- Long-term data access: Is any thing missing to guarantee long-term access to the data produced by your Citizen Science activities? Who should be responsible to provide such services? Which donors might support such a service?
- Re-use: Which and how are open and re-usable results picked-up by third parties? Is the availability of detailed documentation a determining factor?

- Quality assurance: For which reasons do projects exclude quality assurance? Do those projects that do not see any need for quality assurance inform the participants about the absence of any related procedure?
- Standards: Which data and metadata standards are used inside the Citizen Science projects? Assuming diversity, how are the applied standards related to each other?

If we were to repeat the survey, we would consider exploring different forms of citizen participation more widely and to first examine the use of vocabularies (Citizen Science, crowdsourcing, citizen engagement, public participation, volunteered geographic information, etc.) in relation to topic areas. This extension beyond the Citizen Science community might also address Smart Cities, in order to follow the commonalities that have been previously identified[28]. The phrasing of the questionnaire and promotional messages would then have to be carefully aligned.

[28] M. Craglia and C. Granell (2014). Citizen Science and Smart Cities. JRC Technical Report, at http://publications.jrc.ec.europa.eu/repository/bitstream/JRC90374/lbna26652enn.pdf

List of figures

Figure 1: Overview of methodology. 6

Figure 2: examples of the survey featuring on community web pages. 8

Figure 3 examples of the survey featuring on social media platforms. 8

Figure 4: Answers to Question 1 - "Which topic areas does your project cover?" 10

Figure 5: Answers to Question 1.2 - "Which environmental theme(s) does it cover?" ... 11

Figure 6: Answers to Question 3 - "What is the planned duration of the project?"..... 12

Figure 7: Answers to Question 5 - "How is this citizen science project financed?" 13

Figure 8: Answers to Question 9 - "Do you provide access to raw data sets or aggregated values?" 14

Figure 9: Answers to Question 10 - "How do you manage informed consent?" 15

Figure 10: Answers to Question 12 - "If the data collected by the project is accessible via an online download service, how can it be accessed?"..... 16

Figure 11: Answers to Question 13 - "Do you make the data available for re-use?" 17

Figure 12: Answers to Question 13.1 - "Which are the conditions for re-use?" 18

Figure 13: Answers to Question 15 - "Is the data collected by the project quality-controlled?" 19

Figure 14: Answers to Question 20 - "For how long do you ensure the access to the data from you project?" 21

Figure 15: Answers to Question 20 - "For how long do you ensure the access to the data from you project?" - single choice, mandatory. 22

Figure 16: Word cloud summary of answers to Question 22 - "In which country/countries do you store the data?"..... 23

Figure 17: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 13 - "Do you make the data available for re-use?". 25

Figure 18: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and questions 7 (on discovery service), 11 (on view service) and 12 (on download service), where answers have been combined. 26

Figure 19: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 20 - " For how long do you ensure the access to the data from you project?". 26

Figure 20: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 3 - " What is the planned duration of the project?". 27

Figure 21: Combined analysis of question 9 - "Do you provide access to raw datasets or aggregated values?" - and question 21 - " How do you preserve the data from your project?", where PM means Project Member. 27

Figure 22: Combined analysis of question 3 - " What is the planned duration of the project?" - and question 21 - " How do you preserve the data from your project?", where PM means Project Member. 28

Figure 23: Combined analysis of question 20 - "For how long do you ensure the access to the data from you project?" - and question 13 - "Do you make the data available for re-use?"..... 29

Figure 24: Combined analysis of question 17 - "Is the data collected by the project comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata?" - and question 13 - "Do you make the data available for re-use?". 29

Figure 25: Combined analysis of question 19 - "Are the data access and use conditions, including licenses, clearly indicated?" - and question 13 - "Do you make the data available for re-use?”..... 30

Figure 26: Analysis of question 5 - "How is this citizen science project financed?", where projects with a single source of funding have been distinguished from those receiving funding from multiple sources. 31

Figure 27: Combined analysis of question 1 - "Which topic areas does your project cover?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution. 31

Figure 28: Combined analysis of question 13 - "Do you make the data available for re-use?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution. 32

Figure 29: Combined analysis of question 13 - "Do you make the data available for re-use?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution. 32

Figure 30: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 5 - "How is this citizen science project financed?", where project funding including EU sources have been distinguished from those without a European contribution. 33

Figure 31: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 13 - "Do you make the data available for re-use?". 34

Figure 32: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 17 - " Is the data collected by the project comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata?". 34

Figure 33: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 17.1 - "Are these metadata based on international or community-approved (non-proprietary) standards?"..... 35

Figure 34: Combined analysis of question 6 - "Does your project include an explicit data management plan?" - and question 16 - "Is the data collected by the project structured based on international or community-approved (non-proprietary) standards?"..... 36

Figure 35: Combined analysis of question 6 - "Does your project include an explicit data management plan?", question 17.1 – "Are these metadata based on international or community-approved (non-proprietary) standards?" and question 16 - " Is the data collected by the project structured based on international or community-approved (non-proprietary) standards?". 36

Figure 36: Analysis of Question 1 - "Which topic areas does your project cover?" restricted to projects that are set up for more than four years. 37

Figure 37: Analysis of Question 1.2 - "Which environmental theme(s) does it cover?" restricted to projects that are set up for more than four years. 38

Figure 38: Analysis of Question 5 - "How is this citizen science project financed?" restricted to projects that are set up for more than four years. 38

Figure 39: Analysis of Question 1 restricted to projects that store the data also beyond the life-time of the project..... 39

Figure 40: Analysis of Question 1.2 restricted to projects that store the data also beyond the life-time of the project..... 40

List of tables

Table 1: Answers to Question 1.1 – "Which 'other' topic(s) does the project cover?".... 10

Table 2: Answers to Question 1.3 – "Which 'other' environmental theme(s) does the project cover?" (themes related to ecology highlighted in bold, those on pollution in light grey) 11

Table 3: Answers to Question 2 – "Which geographic extent does the project cover?".. 12

Table 4: Answers to Question 22 - "In which country/countries do you store the data?"23

Table 5: Table 3: Answers to Question 2 – "Which geographic extent does the project cover?" including projects that are set up for more than four years (numbers on the left in each cell) and percentage. 39

ANNEX A: Survey introduction and questions

Fields marked with * are mandatory.

This survey addresses the state of play considering data management in Citizen Science projects. It is carried out by the European Commission's Joint Research Centre (JRC), as part of a detailed study of interoperability arrangements, hosting and data management practices for the re-use of citizen-collected data. The questions are inspired by the recently proposed data management principles of the Group on Earth Observations and those of the Belmont Forum. Beyond the pure stocktaking and awareness raising, the results should establish a base line for prioritising follow-up activities and measuring progress.

After discussions with members of the European Citizen Science Association (ECSA) and the international Citizen Science Association (CSA), it was decided to open the scope of the questionnaire to the international community, so that also non-EU and globally acting organizations could benefit from the outcomes.

We currently analyse the first 121 entries and plan to share the results by the end of September. In the meantime, **feel free to share also your experience today!** The analysis will be updated...

Please note that all the answers will be anonymised and analysed in aggregations in order to ensure confidentiality of the data provided. Depending on certain answers you give to the 22 main questions, you might be asked to provide more details. In any case, **it should not take more than 25 minutes to complete this survey**.

General information about the project

1. Which topic areas does your project cover? *

- ☐ Space science
- ☐ Earth science
- ☐ Environmental science
- ☐ Social science
- ☐ Other

2. Which geographic extent does the project cover?

	inside EU	outside EU
Neighborhood	☐	☐
City level	☐	☐
Regional	☐	☐
Country	☐	☐
Continental	☐	☐

3. What is the planned duration of the project?

- ○ < 1 year
- ○ 1-4 years
- ○ > 4 years
- ○ I do not know

4. What is the status of this particular project?

- ☐ Experimental (i.e. currently testing ideas and new techniques)
- ☐ Operational (i.e. currently in use or ready for use)

5. How is this citizen science project financed? *

- ☐ In kind contributions
- ☐ Private donations
- ☐ Participant or membership fees
- ☐ National funding scheme
- ☐ European Commission (FP7 or Horizon 2020)
- ☐ Other

6. Does your project include an explicit data management plan? *

- ○ Yes
- ○ No

Discoverability of Citizen Science data

7. Is the data and all associated metadata from your project discoverable through catalogues and search engines?

- ○ Yes
- ○ No

8. Does the data contain persistent, unique, and resolvable identifiers?

- ○ Yes
- ○ No

Accessibility of Citizen Science data

9. Do you provide access to raw data sets or aggregated values? *

- [ ] Raw
- [ ] Aggregated

9.1 Do the raw data sets include information about the data producer (e.g. user names, individual identifiers, or locations)?

- ( ) Yes
- ( ) No

10. How do you manage informed consent?

- ( ) Not at all
- ( ) Through a generic terms of use agreement
- ( ) Through signature of an informed consent form
- ( ) Other

11. Is the data collected by the project accessible via an online services for visualization (e.g. by OGC Web Map Service - WMS)?

- ( ) Yes
- ( ) No

12. If the data collected by the project is accessible via an online download service, how can it be accessed?

- [ ] Bulk download of all data at once (e.g. by FTP download)
- [ ] User-customizable download of selected parts (e.g. OGC Web Feature Service - with Filter Encoding)
- [ ] User-customizable services for computation (e.g. by a data processing API)

Re-use conditions of Citizen Science data

13. Do you make the data available for re-use? *

- ( ) Yes (directly after data acquisition/production)
- ( ) Yes (in a staged approach, i.e. after an embargo period)
- ( ) No

14. Why did you take this decision?

Usability of Citizen Science data

15. Is the data collected by the project quality-controlled? *

- Yes
- No

16. Is the data collected by the project structured based on international or community-approved (non-proprietary) standards?

- Yes
- No

17. Is the data collected by the project comprehensively documented, including all elements necessary to access, use, understand, and process, preferably via formal structured metadata?

- Yes
- No

18. Are you providing a dedicated contact point for the data collected by the project?

- Yes
- No

19. Are the data access and use conditions, including licenses, clearly indicated?

- Yes
- No

Preservation and curation of Citizen Science data

20. For how long do you ensure the access to the data from you project? *

- During the lifetime of project
- Beyond the life-lime of the project

21. How do you preserve the data from your project?

- Local machine of a project member
- Remote server hosted by a project member
- Research library
- Public repository
- Other

22. In which country/countries do you store the data?

Additional information

If you would like to share the name of your Citizen Science project for future reference, please use the box below.

If you would be so kind to share the web page of your project, please use the box below.

If you allow us to come back to you after the survey, please leave your e-mail address below. This information will not be further distributed, used as part of the analysis or made available with the survey results.

@

If you have any additional information to share with us, please feel free to use the text box below.

Europe Direct is a service to help you find answers to your questions about the European Union
Free phone number (*): 00 800 6 7 8 9 10 11
(*) Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed.

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server http://europa.eu

How to obtain EU publications

Our publications are available from EU Bookshop (http://bookshop.europa.eu),
where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents.
You can obtain their contact details by sending a fax to (352) 29 29-42758.

LB-NA-27920-EN-N

JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation

doi:10.2788/539115

ISBN 978-92-79-58387-2

Publications Office