CS TRACK
Investigating Citizen Science

Horizon 2020 / Science with and for Society Programme
Grant agreement number: 872522

D2.3: Documentation of datasets produced from selected CS activities in Europe

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 872522

Title of project	CS Track
Full title of project	Expanding our knowledge on Citizen Science through analytics and analysis
Title of this document	Documentation of datasets produced from selected CS activities in Europe
Number of this document	D2.3
Dissemination level	Public
Due date	30th November 2022
Actual delivery	30th November 2022
Versioning history	
Author	Coordination: Patricia Santos (UPF) Authors (*ordered alphabetically by institution)*: FORTH: Marinos Anastasakis, Katherina Kikis-Papadakis JYU: Kai Nils Weeber, Aaron J. Peltoniemi, Ohto Sabel, Raija Hämäläinen, Paavo Räty MOFET: Yaela Golumbic, Reuma De Groot, Tslil Farchi, Anne Turbe RIAS: Ulrich Hoppe, Cleo Schulten UPF: Ishari Amarasinghe, Miriam Calvera Isabal, Patricia Santos, Nicolas Gutierrez Paez, Davinia Hernández-Leo URJC: David Roldán-Álvarez, Fernando Martínez-Martínez WiD: Marius Oesterheld, Vincent Schmid-Loertzer WLW: Christine Urban, Michael Strähle
Executive summary	This deliverable is aimed to compile the datasets that have been generated in the context of the work done by WP2. In addition to this, the source code to replicate the CS Track database is reported as well in this deliverable, mainly because this database has been used as the main source to generate the corresponding datasets. The details regarding the research conducted in the associated studies are fully documented in D2.2. This deliverable only contains a summary of the corresponding study in order to frame the purpose and relatedness of the dataset. In addition to this, for each dataset it is briefly described the content and its grouping (how data is organised). Each dataset corresponds to different nature of research, this means its content will differ among them.

Table of Contents

1. Orientation and general approach

This deliverable is focused on presenting the datasets resulted from the studies presented on D2.2 "Final documentation of initiatives selected for analysis". For further information about the studies please refer to D2.2.

According to the literature review done by Renear et al (2010): "*The concept of a dataset is common to almost every scientific discipline where data provide the empirical basis for research activities*." These authors propose that most definitions of a dataset have four features: grouping, content, relatedness, purpose. See for more detail the following figure 1.

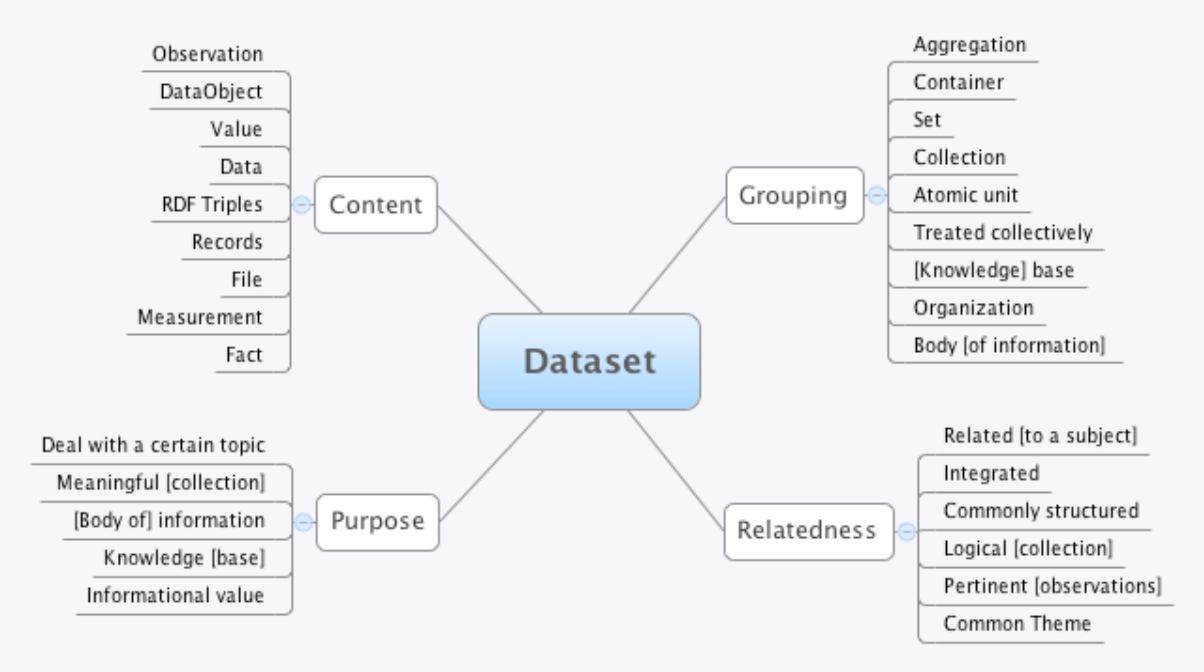

Figure 1. Conceptual map of dataset, figure extracted from Renear et al (2010)

This deliverable is aimed to present different types of datasets that have been produced depending on the nature of the research as a result of the research conducted in WP2 and presented in D2.2. This deliverable contains the final list of datasets produced, it is not aimed to present the detail of the corresponding studies/research lines (this is already presented in D2.2). Here we present a summary that includes information on the four main features of each dataset.

Before presenting the corresponding datasets, section 2 is included in this deliverable with the aim to present the main algorithms used to build the CS Track database and other algorithms used to classify data (and therefore to produce most of the presented datasets). The section contains the corresponding links with the documentation and code uploaded on Github.

Sections 3, 4, 5 and 6 provide detail of research data associated with the empirical studies presented in D2.2. We have followed the same number of sections in D2.2 and D2.3 to relate the studies and the datasets. Only the studies that have produced a dataset are documented in this deliverable.
In order to facilitate the reading of each collection the following structure has been followed:

- Title
- Authors
- Purpose of the study and relatedness
- Content of the dataset and grouping
- Type of access:

In Zenodo the access to the dataset can be open by default or with restricted access (this option applies to studies/research that are pending in progress to be submitted for a publication, once accepted the type of access will be modified to open).

The content of each study depends on the type of research and process carried out. As defined by Figure 1, examples of content can include: Anonymized Data Sheet: participant responses to survey questions; Instruments used for analysis (interview protocol, survey….); Anonymized collection of data from CS Track Database; Raw data; Measurements; Other files.

Following the CS Track data management strategy GitHub and Zenodo have been selected as the main repositories for this purpose.

Reference: Renear, A. H., Sacchi, S., & Wickett, K. M. (2010). Definitions of dataset in the scientific and technical literature. Proceedings of the American Society for Information Science and Technology, 47(1), 1-4.

2. CS Track database source code and main dataset

The source code and documentation developed to build the CS Track database is uploaded on GitHub: https://github.com/orgs/CS-Track-Code/repositories. The information is structured in the following repositories:

- **projectdatabase**: contains all the code and information to deploy the database. The wiki created in the repository contains information about the crawler documentation, the database selection, database structure, information about how to check robots.txt file and how to consume the data from the Metabase tool.
- **text-analysis-tools_applied_database**: contains all the code developed to:
    - Extract data automatically from the CS projects descriptions using the workbench (CSTrack_anonymizer_NER_SDG.py)
    - Classify CS projects by SDGs using the workbench (CSTrack_anonymizer_NER_SDG.py)
    - Anonymize the data in the database (CSTrack_anonymizer_NER_SDG.py)

- Fill descriptors information from NER results (CSTrack_New_categories.py)

Furthermore, it has been developed a proof of concept of how other descriptors could be filled with similar automatic text analysis techniques. You will find more information in this repository (https://github.com/maramartiez00/TFG_MaraMartinez/blob/main/InformationExtracction.ipynb) and their related publication (https://repositori.upf.edu/handle/10230/54792). Check other CS-Track-Code repositories if needed

- **project-categorization:** contains files related to the data analysis provided in deliverable D2.1 and section 7 of the deliverables D2.2 and D2.3.

All the code contained in the repositories has been developed to carry out the different studies and generate the associated datasets described below.

In addition to the code used to build the CS Track database, and the classification algorithms. This is the main dataset which consists of a list containing 4849 records of projects (useful for analysis) associated with relevant descriptors of the CS Track database.

Purpose and relatedness: This dataset contains the results of classification of CS projects (the total number of projects collected in the CS Track database) with research areas and SDGs. This list can be used for further research purposes, each project is related to its corresponding CS platform and therefore researchers could have access to its description or further details. Details about the research conducted with this dataset can be found in D2.,2 section 7.

Content and grouping: This dataset contains the following information from the CS Track database:

- Citizen Science (CS) projects title
- the data extracted date
- the language of the CS projects informations
- the URL(s) of the website(s) from where the CS projects information was extracted. For other studies developed in CS Track consortium it might be useful to consult this data
- full list of assignments for research areas and SDGs for each CS project.

Zenodo URL: https://zenodo.org/record/7356627#.Y39bEnaZNPY

Finally, the following file contains the resulting list of descriptors (extended version of the PPSR) derived from the analysis of CS platforms. For more information, see D2.2 section 6.1.

Zenodo URL: https://zenodo.org/record/7310445#.Y2zph3aZNPY

3. People involved in CS

3.1 An explorative study on the effects of the recent pandemic on online Citizen Science: lessons learnt for improving project management and implementation

Marinos Anastasakis[a] and Kathy Kikis-Papadakis[a]

[a] Foundation for Research and Technology - Hellas (FORTH)

Purpose and relatedness: This mixed methods study intends to shed light to research concerns pertaining to the pandemic's effects on participation, project coordinators' perceptions and experiences of the situation and, how the pandemic's effects were counterbalanced. Results indicate that although the pandemic was an opportunity for online CS, it also highlighted a number of challenges such as the need for recognizing citizens' contributions and the need for sustaining engagement with citizen scientists.

More information on this research can be found in D2.2 section 3.1.

Content and grouping:

The following dataset contains the instrument used for collecting data for this study i.e., the interview protocol. The other file in Zenodo does include excerpts from the interviews. For each project we have included the question asked (denoted as "Q") and below the answer provided by the project coordinator (denoted as "A").

- Interview protocol: https://doi.org/10.5281/zenodo.7299312
- Interview excerpts: https://doi.org/10.5281/zenodo.7307600

3.2 Report on a Survey among Organisers of Citizen Science Projects

Michael Strähle[a] & Christine Urban[a] *(alphabetical order)*

[a] Wissenschaftsladen Wien - Science Shop Vienna

Purpose and relatedness: The purpose of this study was focused on collecting data on some project characteristics which cannot be answered by visiting project websites. This was mainly conducted with a survey among organisers of citizen science projects. This dataset is composed of the research data which contain the responses obtained from the very short questionnaire "VSQ" and the responses used for the report.

More information on this research can be found in D2.2 section 3.2.

Content and grouping:

- Collection of responses obtained from the very short questionnaire "VSQ"
- Collection of responses used for the report
- Full report

Zenodo URL: https://zenodo.org/record/7310071#.Y2zOXHaZNPY

3.3 Availability of information on citizen science activities, checked against the Activities & Dimensions Grid of Citizen Science on the basis of some projects

Michael Strähle[a] & Christine Urban[a] *(alphabetical order)*

[a] Wissenschaftsladen Wien - Science Shop Vienna

Purpose and relatedness: The research resulting in this report aimed at answering the following questions:

- Which information on citizen science activities is online available that matches the Activity & Dimension Grid of Citizen Science or goes beyond it?
- Is there any contradictory information?
- What can be the reason for the availability or non-availability of information about citizen science activities?
- How does/could this impact on the CS Track's recommendations?

The corresponding dataset consists of the results of a keyword-based search in the WP2 project database. The information retrieval resulted in 3318 projects on which information is available in German or English.

More information on this research can be found in D2.2 section 3.2.

Content and grouping:

- List of projects analysed (check dataset published in Zenodo)

Zenodo URL: https://zenodo.org/record/7376970#.Y4X87XaZOUk

3.4. Is it a match? Motivations on citizen science volunteers and recruitment arguments in project descriptions

Kai Nils Weeber[a], Nicolas Felipe Gutierrez Paez[b], Ohto Sabel[a], Raija Hämäläinen[a]

[a] University of Jyväskylä

[b] TIDE Research Group, Universitat Pompeu Fabra

Purpose and relatedness: Data has been collected by quantitative triangulation. 1076 participants in citizen science projects answered a survey about the 12 motivational factors for participating. They had access to the survey by social media posts or email invitations sent to people in charge of projects. Data regarding motivational arguments in recruitment come from quantitative content analysis of 367 project descriptions of the website Zooniverse. The content analysis of the project descriptions was done manually by two coders independently. Then, both coders analysed their codings and reached consensus.

More information on this research can be found in D2.2 section 3.4.

Content and grouping:

- Survey about 12 motivational factors for participating in CS projects.
- Content analysis of 367 project descriptions from Zooniverse platform. With the content analysis, 5 factors were extracted, namely desire of contributing, joining a community, social interaction, enjoyment and project reputation.

Zenodo URL: https://doi.org/10.5281/zenodo.7310080

4. CS and its relation with the Science System

4.1 Investigating the potential of citizen science to respond to emerging challenges - The case of COVID-19

Yaela Golumbic [a, b], Reuma De - Groot [a], Tslil Farchi [a] & Anne Turbe [a]

[a] The MOFET Institute
[d] The Steinhardt Museum of Natural History

Purpose and relatedness: This study investigated the power of citizen science to respond to emerging challenges, using the case study of the COVID-19 pandemic. It employed a two-phase research approach using website content analysis followed by interviews and detailed case studies of seven citizen science initiatives.

More information on this research can be found in D2.2 section 4.1.

Content and grouping:

- List of attributes of citizen science projects which were extracted during the content analysis phase
- Final list of projects serving as case studies, and their characteristics
- Interview template and schedule
- Full report of the study

All the above data are available in the final study report.

Zenodo URL: https://zenodo.org/record/7310477#.Y4SQa3aZOUk

4.2 Mapping Sustainable Development Goals to Citizen Science projects

Patricia Santos[a], Ishari Amarashinghe[a], Miriam Calvera-Isabal[a], Cleo Schulten[b], H.Ulrich Hoppe[b], David Roldán-Álvarez[c], Fernando Martínez-Martínez[c]

[a] TIDE Research Group, Universitat Pompeu Fabra

[b] RIAS Institute

[c] Universidad Rey Juan Carlos

Purpose and relatedness: This work presents opportunities, achievements, and future challenges in using computational analytics to better understand the connection between CS and the SDGs. The work in its status does not fully cover SDGs in CS, but it evaluates and shows the potential of the text-classification techniques for identifying SDGs in CS project descriptions and for assessing trends in connection of CS and SDGs based on available data.
A total of 56 websites have been used to extract randomly 208 project descriptions from the CS Track database collected during 2019-2022.

The following resources have been considered for elaboration of the compiled list of keywords:

- Fraisl, D., Campbell, J., See, L., Wehn, U., Wardlaw, J., Gold, M., ... & Fritz, S. (2020). Mapping citizen science contributions to the UN sustainable development goals. Sustainability Science, 15(6), 1735-1751. https://doi.org/10.1007/s11625-020-00833-7
- Schleicher, K., & Schmidt, C. (2020). Citizen science in Germany as research and sustainability education: Analysis of the main forms and foci and its relation to the sustainable development goals. Sustainability, 12(15), 6044.. https://doi.org/10.3390/su12156044
- Körfgen, A., Förster, K., Glatz, I., Maier, S., Becsi, B., Meyer, A., ... & Stötter, J. (2018). It's A Hit! Mapping Austrian research contributions to the sustainable development goals. Sustainability, 10(9), 3295.

More information on this research can be found in D2.2 section 4.2.

Content and grouping:

- Compiled list of keywords for SDG mapping.
- Spreadsheet containing training data from SciStarter (set of projects associated with SDGs)
- Spreadsheet containing the list of projects associated with SDGs in relation to the different automatic classifiers (nCoder, ESA and OSDG) and manual coding.
- Presentation in CS Track ECSA event (8th October 2022, Berlin): CS-Track database: a central database of CS projects in Europe that can be key to understand the connection of CS and SDGs "Understanding the nature of Citizen Science in a rapidly changing world"
- The associated full report resulting from this study is in progress. Please contact the corresponding authors for further information.

Zenodo URL: https://zenodo.org/record/7310477#.Y2zornaZNPY

4.3 Tasks in Citizen Science: proposing a hierarchical framework for categorising citizen scientists' activities in CS projects

Authors and Research Affiliation

Marinos Anastasakis[a] and Kathy Kikis-Papadakis[a]

[a] Foundation for Research and Technology - Hellas (FORTH)

Purpose and relatedness: This study intends to propose a way of categorising in a hierarchical manner the tasks undertaken by citizen scientists in CS projects, and based on that to suggest a set of new descriptors for the CS Track database. The data used in this study are extracted from CS Track's database.

More information on this research can be found in D2.2 section 4.3.

Content and grouping:

This dataset (derived from the CS Track database) contains a list of projects with the descriptors "activity type", "development space", "development time", "investment", "language", "member age", "country", "topics" and "research areas". The above descriptors were selected because they provide useful context and a sufficient description of the tasks citizen scientists are engaged in each project.

Zenodo URL: https://doi.org/10.5281/zenodo.7307761

5. CS and Education

5.1 Identifying learning dimensions in CS project descriptions

Marius Oesterheld,[a] Vincent Schmid-Loertzer,[a] Miriam Calvera-Isabal,[b] Ishari Amarasinghe,[b] Patricia Santos,[b] & Yaela N Golumbic[c,d]

[a] Wissenschaft im Dialog
[b] TIDE Research Group, Universitat Pompeu Fabra
[c] The MOFET Institute
[d] The Steinhardt Museum of Natural History

Purpose and relatedness: For this study, we conducted a qualitative content analysis of a random sample of 94 English-language project descriptions stored in the CS Track database with the goal of determining which dimensions of learning are reflected most prominently in these texts. Using a slightly modified version of the model of individual learning outcomes developed by Phillips et al. in 2018 as a coding rubric, two members of the research team independently coded all project descriptions by manually assigning phrases, sentences and short paragraphs to eight main categories and 21 subcategories. From these text snippets, distinctive and frequently occurring keywords were extracted, which have since been used in follow-up studies.

Our analysis revealed that some learning dimensions (such as data collection or using technology) are very prominently discussed in the project descriptions we studied, while others (e.g. experimenting, study design, community action) are clearly underrepresented. In other words, the project descriptions analysed only partially reflect the educational potential of participation in CS. Based on these findings, we suggested possible explanations and ways in which the issue could be addressed on the level of both project design and project communication.

This study profited immensely from the kind support of Tina Phillips and her colleagues, who agreed to share parts of their coded dataset with us.

Details related to the analysis procedure are provided in a paper which is currently under review (on the date of submission of this deliverable - November 2022), no link to a repository is available. Contact the main authors if you have interest to receive further information).

References: Phillips, T., Porticella, N., Constas, M., & Bonney, R. (2018). A framework for articulating and measuring individual learning outcomes from participation in citizen science. Citizen Science: Theory and Practice, 3(2).

More information on this research can be found in D2.2 section 5.1.

Content and grouping:

- coding rubric used (Main category, Subcategory, Definition (Draft) and Inclusion/exclusion criteria) columns)
- complete coding (i.e. including CS Track projects title)
- calculation of rate of agreement between coders
- list of keywords extracted

- list of projects titles and the website URL from where we extracted the information

Zenodo URL: https://zenodo.org/record/7374000#.Y4UGjnaZNPY

5.2 Learning in citizen science: a triangulation approach

Miriam Calvera-Isabal [a], Marius Oesterheld [b], Fernando Martínez-Martínez [c], Aaron J. Peltoniemi [d], Patricia Santos [a], and Yaela N Golumbic [e, f]
[a]TIDE Research Group, Universitat Pompeu Fabra
[b]Wissenschaft im Dialog
[c]Universidad Rey Juan Carlos
[d]Jyväskylän yliopisto
[e]The MOFET Institute
[f]The Steinhardt Museum of Natural History

Purpose and relatedness: This multi-perspective study investigated the alignment between (1) learning opportunities mentioned by CS project initiators in CS project descriptions available online, (2) learning opportunities shared in tweets from project or platform accounts and, (3) the participants' perceived learning experiences as reflected in survey responses and tweets from individual user accounts. The aim was to identify overlaps and discrepancies between learning opportunities envisioned by project coordinators' and citizen scientists' perspectives on learning in CS projects. To this end, we triangulated three datasets - project descriptions (N=94), tweets (N = 216,786) and survey responses (N = 610). The first dataset - a qualitative content analysis of project descriptions stored in the CS Track database – was created in the context of a previous study (cf. section 5.1). The keywords derived from this manual coding of project descriptions were used to conduct an automated analysis of tweets and thus form the basis of the second dataset used in this triangulation study. The third dataset consists of citizen scientists' responses to the CS Track online survey, which focused primarily on Europe and was distributed for a period of seven months (January-July 2021) through multiple channels. As a second step, we conducted a case study of 11 projects, which allowed us to narrow down the three datasets (project descriptions: N=11, tweets: N=118, survey responses: N=139) and draw conclusions on the level of individual projects.

The results of both the general comparison and the project-level case study reveal that there is a significant discrepancy between the learning opportunities described by project coordinators and the learning experiences reported by project participants. This gap is particularly evident with regard to skills related to communication and project or research design, but also when it comes to scientific literacy and critical thinking. What our findings also show is that responses vary considerably even among volunteers who participated in the same CS project, which suggests that the citizen scientists' individual backgrounds, interests and motivations play an important role in shaping their learning experiences.

References: M. Oesterheld, V. Schmid-Loertzer, M. Calvera-Isabal, I. Amarasinghe, P. Santos, & Y. Golumbic (2022). Identifying learning dimensions in citizen science projects. In proceedings of Engaging Citizen Science Conference 2022, PoS(CitSci2022) 070. [forthcoming]

More information on this research can be found in D2.2 section 5.2.

Content and grouping:

- Datasets and documentation related to Oesterheld et al. (2022) study (see section 5.1)
- CS Track survey questions (The CS Track survey data is not currently publicly available. Specific questions regarding survey data (e.g. questions) can be sent to Raija Hämäläinen (raija.h.hamalainen@jyu.fi).
- Triangulation roadmap (juxtaposition of the categories/survey questions we compared)

Zenodo URL: https://zenodo.org/record/7371616#.Y4SVC3aZOUk

5.3 Educational uses of CS data

Miriam Calvera-Isabal [a], Marius Oesterheld [b], Fernando Martínez-Martínez [c], Aaron J. Peltoniemi [d], Patricia Santos [a], and Yaela N Golumbic [e, f]
[a]TIDE Research Group, Universitat Pompeu Fabra
[b]Wissenschaft im Dialog
[c]Universidad Rey Juan Carlos
[d]Jyväskylän yliopisto
[e]The MOFET Institute
[f]The Steinhardt Museum of Natural History

Purpose and relatedness: This section contains a collection of datasets because they correspond to three different (but interrelated) case studies:

1. An initial analysis of the CS projects information available on CS platforms
2. The analysis websites and online platforms which share CS projects information to understand how CS is communicated online and exploring how automatic methods can be applied with educational purposes
3. A co-design of a tool to show the CS data gathered for teachers to support their practice

A total of 59 platforms were analysed and a total of 4949 CS projects were extracted from them. A list of descriptors were defined to create the database. From the workshops conducted with teachers, a prototype was created by analysing the card sorting and paper prototypes results.

More information on this research can be found in D2.2 section 5.3.

Content and grouping:

- Descriptors definition: https://zenodo.org/record/7310445#.Y2zmbXaZNPY
- List of websites: https://zenodo.org/record/7310295#.Y2zmwXaZNPY
- Questionnaires: https://zenodo.org/record/6655987#.Y2z-yHaZNPY
- Cards designed for card sorting and paper prototyping activity: https://zenodo.org/record/6655972#.Y2z_LnaZNPY
- Prototype 1st version - CS projects dashboard: https://zenodo.org/record/6655902#.Y2z_-naZNPY
- Prototype 2nd version - CS projects dashboard: https://zenodo.org/record/6655910#.Y20AAXaZNPY
- Learning activity design canvas: https://zenodo.org/record/6655958#.Y20AAXaZNPY

6. Visibility of CS

6.1 How to automate the extraction and analysis of information for educational purposes

Miriam Calvera-Isabal [a], Patricia Santos [a], H.Ulrich Hoppe[b] & Cleo Schulten[b]
[a] TIDE Research Group, Universitat Pompeu Fabra
[b] RIAS Institute

Purpose and relatedness: The objective of this study was to analyse how data about CS is shared and communicated online. Furthermore, explore how it could be extracted and stored in a central database. Finally, we studied how this data can be used in formal education contexts. We analysed both websites and CS projects information extracted from 72 websites selected.

More information on this research can be found in D2.2 section 6.1.

Content and grouping:

- Database: https://zenodo.org/record/7356627#.Y39bEnaZNPY
- List of descriptors: https://zenodo.org/record/7310445#.Y2zph3aZNPY

6.2 Citizen science project descriptions as science communication texts - the good, the bad, and the ugly

Yaela N Golumbic [a b], Marius Oesterheld [c] & Nicolas Felipe Gutierrez [d]

[a] The MOFET Institute
[b] The Steinhardt Museum of Natural History
[c] Wissenschaft im Dialog
[d] TIDE Research Group, Universitat Pompeu Fabra

Purpose and relatedness: This study aimed to determine to what extent do CS project descriptions actually contain the kinds of information relevant to prospective participants and whether this information is conveyed in a comprehensible and attractive manner. To this end, we conducted a qualitative content analysis of a random sample of 120 English-language project descriptions stored in the CS Track database. The coding rubric used for this study is based on the ten-step template for writing engaging project descriptions we recently designed and published (cf. section 7.1). The sample was produced as follows: After creating a dataset containing only English-language project descriptions, we excluded all descriptions which consist of less than 100 or more than 500 words. Texts of less than 100 words cannot be expected to contain a significant amount of information. Project descriptions of more than 500 words are less likely to be read in their entirety than shorter texts and thus ill-suited to the task of capturing the readers' interest and prompting them to join the project in question. Finally, we applied the 'random' function of RStudio to randomly select 120 texts from the resulting dataset of 1283 descriptions. The qualitative content analysis was performed in two consecutive steps. First, in order to ensure that the coding rubric is fit for purpose and all categories within it well-defined and demarcated, all three members of the research team independently coded 40 project descriptions. After discussing the results and making slight modifications to the coding rubric, each team member coded roughly one third of the remaining 80 descriptions.

Preliminary results suggest that the majority of project descriptions in our sample fail to mention how citizen scientists will benefit from participating, what kind of training they will receive, how their contributions will be acknowledged, and whether they will have access to project results. Furthermore, the project's goals, its target audience, and the tasks volunteers will be expected to complete are very often not described explicitly and clearly enough. For instance, very few project descriptions contain concrete information on required skills and equipment or on the time commitment associated with participation.

More information on this research can be found in D2.2 section 6.2.

Content and grouping:

- coding rubric
- visualisation of preliminary results

ZENODO URL: https://zenodo.org/record/7381002#.Y4c9CcuZM2w

7. The CS Track database: contribution of the empirical studies to enhancing understandings on/for CS

7.1 Research Areas in Citizen Science

Ishari Amarasinghe [a], Miriam Calvera [a] & Patricia Santos [a]

[a] TIDE Research Group, Universitat Pompeu Fabra

Purpose and relatedness: The assignment results of **research areas to CS project descriptions** are provided in the following dataset. The analysis was conducted using data retrieved from the CSTrack database on 2022/09/15. At this point the database consisted of 4949 CS Project records (this includes English and non-English descriptions).

See further detail about the study in D2.2 section 7.2.

Content and grouping:

- The dataset contains the following details: Platform ID (from which platform the CS project descriptions were retrieved), Project Title (Name of the CS project), Research area assignment results (More details about the assignment technique can be found in D3.2 'Web Analytics Toolset and Workbench' - ESA backend), Research Area assignment reported in section 7.2 of D2.2 (which only considered the research area assignment with the highest similarity), Sub research area assignment ((which only considered the sub research area assignment with the highest similarity).

Zenodo URL: https://zenodo.org/record/7310341#.Y2zhgXaZNPY

7.2 Sustainable Development Goals (SDGs) covered by CS projects

Ishari Amarasinghe [a], Miriam Calvera [a] & Patricia Santos [a]

[a] TIDE Research Group, Universitat Pompeu Fabra

Purpose and relatedness: The assignment results of SDGs to CS project descriptions are provided in the following dataset. The analysis was conducted based on data retrieved from the CSTRack database on 2022/09/15.

See further detail about the study in D2.2 section 7.3.

Content and grouping:

- The dataset contains the following details: Platform ID (from which platform the CS project descriptions were retrieved), Project Title (Name of the CS project), SDG assignment results (More details about the assignment technique can be found in D3.2 'Web Analytics Toolset and Workbench' - ESA backend), SDG assignment reported in section 7.2 of D2.2 (which only considered the SDG assignment with the highest similarity).

Zenodo URL: https://zenodo.org/record/7310353#.Y4STY3aZOUk

7.3 Skills of science inquiry in projects' descriptions

Ishari Amarasinghe [a], Miriam Calvera [a] & Patricia Santos [a]

[a] TIDE Research Group, Universitat Pompeu Fabra

Purpose and relatedness: The dataset provides results of a prediction task aimed at predicting the presence of **science inquiry skills in CS project descriptions.** Only 2939 English project descriptions from the database were used for prediction and the results indicated 438 projects (around 15%) consist of one or more skills of science inquiry that we were interested in. In total 20 different types of science inquiry skills were considered for this study.

See further detail in D2.2 section 7.4.

Content and grouping:

- The dataset contains the following details: Platform ID (from which platform the CS project descriptions were retrieved), Project Title (Name of the CS project), How many times each science inquiry skill (keywords) was mentioned in each project description.

ZENODO URL: https://zenodo.org/record/7332112#.Y4ST93aZOUk