



This project is co-financed by the European Union

Grant Agreement No.: 824603

Call: H2020-SwafS-2018-1

Type of action: RIA

Starting date: 1/02/2019



D4.1 Conceptual architecture and delivery plan

Coordinator: SINTEF
Quality reviewer: SOTON



Deliverable nature	Report
Dissemination level	Public
Work package and Task	WP1
Contractual delivery date	M12
Actual delivery date	M12

Authors

Author name	Organization	E-Mail
Dumitru Roman, Shady Abd El Kader	SINTEF	dumitru.roman@sintef.no
Gefion Thuermer	SOTON	gefion.thuermer@soton.ac.uk
Esteban Gonzalez	UPM	egonzalez@fi.upm.es
Irene Celino Gloria Re Calegari Damiano Scandolari	CEFRIEL	irene.celino@cefriel.com gloria.re@cefriel.com damiano.scandolari@cefriel.com



Abstract	In this deliverable we provide a conceptual architecture for data-driven citizen science projects. In order to do that, we analyzed the current practices of citizen science projects related to pollution from a data science perspective. For the citizen science projects that aim to be more data-driven, we outline a number of recommendations in that direction. Based on that, we devise the architecture of the ACTION toolkit with a number of areas to be addressed in the realization of the toolkit, together with a delivery plan of the toolkit in the project. The results of this deliverable contribute to providing a range of technology and assistance services relevant to the digital infrastructure citizen science initiatives.
Keywords	data science, citizen science, toolkit

Disclaimer

The information, documentation and figures available in this deliverable, is written by the ACTION project consortium under EC grant agreement 824603 and does not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.



This deliverable is licensed under a Creative Commons Attribution 4.0 International License

How to quote this document

Roman, D.; El Kader, S. A.; Thuermer, G.; Gonzalez, E.; Celino, I.; Calegari, G.; Scandolari, D. (2020), ACTION Deliverable D4.1 Conceptual architecture and delivery plan.





TABLE OF CONTENTS

Authors	1
EXECUTIVE SUMMARY	4
1 INTRODUCTION	5
2 DATA SCIENCE DRIVEN ANALYSIS OF POLLUTION CITIZEN SCIENCE PROJECTS	5
2.1 A Typical Data Science Pipeline	6
2.2 Analysis of Pollution Citizen Science Projects from a Data Science Perspective	8
2.2.1 Source of Citizen Science Projects	8
2.2.2 Selection of Projects of Interest (Pollution-related)	8
2.2.3 Metrics of Analysis	10
2.2.4 Problem Definition	11
2.2.5 Data Management	11
2.2.5 Data Analysis	13
2.2.6 Result Publication	13
2.2.7 Example of Citizen Science Projects with Implemented Data Science Tasks	15
2.3 Summary of Analysis	17
2.4 General Recommendations	17
2.5 Final Consideration	18
3 CONCEPTUAL ARCHITECTURE OF THE ACTION CITIZEN SCIENCE TOOLKIT	18
4 DELIVERY PLAN OF THE ACTION TOOLKIT	21
4.1 CONEY Toolkit	23
4.1.1 CONEY Create	24
4.1.2 CONEY Chat	25
4.1.3 CONEY Inspect	26
4.2 Templates	26
5 SUMMARY AND FUTURE WORK	27
6 REFERENCES	27



EXECUTIVE SUMMARY

In this deliverable we provide a conceptual architecture for data-driven citizen science projects. In order to do that, we analyzed the current practices of citizen science projects related to pollution from a data science perspective. For the citizen science projects that aim to be more data-driven, we outline a number of recommendations in that direction. Based on that, we devise the architecture of the ACTION toolkit with a number of areas to be addressed in the realization of the toolkit, together with a delivery plan of the toolkit in the project. The results of this deliverable contribute to providing a range of technology and assistance services relevant to the digital infrastructure citizen science initiatives.



1 INTRODUCTION

WP4 aims to provide a range of technology and assistance services relevant to the digital infrastructure citizen science initiatives. In conjunction with WP5, the project will deliver a toolkit for citizen science, which consists of methodologies, methods and other resources that respond to a wide range of citizen science characteristics. The toolkit aims to address some of greatest challenges citizen science teams encounter - from the choice of optimum data science methodologies and the design of citizen contributions and engagements to quality assurance, rewards and incentivisation, as well as monitoring, impact and sustainability.

In this context, this deliverable aims to define the overall conceptual architecture of the toolkit that will allow pollution-centered citizen science projects to be implemented and managed in a more effective and efficient way. In addition, the deliverable aims to outline a delivery plan for the realization of the toolkit as part of WP4 and WP5.

The approach taken in the definition of the conceptual architecture is based on a thorough analysis of existing citizen science projects related to pollution. We take a data science centric approach in analyzing the projects, with the primary focus to identify the degree to which pollution-related citizen science projects follow data science principles. Based on this analysis, for citizen science projects that aim to be more data-driven, we provide a set of recommendations to be followed to implement a more data science driven process. Based on these recommendations we outline the architecture of the toolkit in terms of a number of areas that need to be taken into account and outline the realization of the toolkit in the project.

The rest of this document is organized as follows. Section 2 provides a brief introduction to a typical data science pipeline, used to analyze a number of pollution-related citizen science projects, together with the results and recommendations based on the analysis. Section 3 outlines the architecture of the toolkit. Section 4 provides an overview of the delivery plan for the toolkit. Finally, Section 5 summarizes this deliverable and outlines possible future directions to be taken into consideration.

2 DATA SCIENCE DRIVEN ANALYSIS OF POLLUTION CITIZEN SCIENCE PROJECTS

This section starts with an overview of a typical Data Science pipeline (Section 2.1). We then use the components of the Data Science pipeline as a mechanism to analyze pollution citizen science projects (Section 2.2), report the results of the analysis (Section 2.3), provide recommendations and final remarks (Sections 2.4 and 2.5).



2.1 A Typical Data Science Pipeline

Data Science is a multidisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data Science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data¹.

In a typical Data Science pipeline, we have some common and general elements, like a phase of **problem definition**, a phase of **data management**, a phase of **hypothesis (HP) testing** and a phase of **result evaluation** as shown in Figure 1.

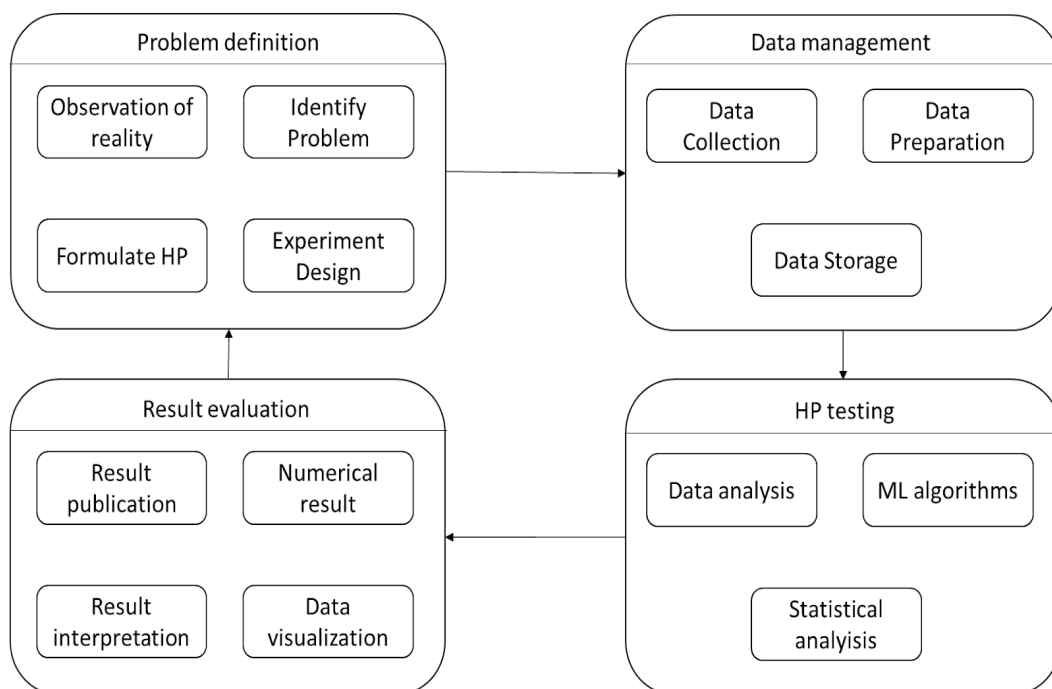


Figure 1. A typical Data Science pipeline²

With a deeper look to the pipeline we can identify some common activities:

Starting from the **problem definition** phase data scientists can **identify a problem** or a subject of interest, then analyzing this problem they **can formulate a hypothesis or a goal**. Data scientists will then plan the **experiment design** in order to solve the initial problem. In particular:

- A hypothesis states the predictions the research aims to find. It is a tentative answer to a research question that has not yet been tested.

¹ https://en.wikipedia.org/wiki/Data_science

² Based on information from

<https://michaelbrodie.com/wp-content/uploads/2015/04/Brodie-Onassis-Lecture-on-Big-Data-Lecture-2.pdf>



- A hypothesis must be testable, which means you can support or refute it through scientific methods (such as experiments, observations and statistical analysis of data). The hypothesis defined must be unequivocal.
- The formulation of the hypothesis is strictly connected to the **identification of the problem** and the **experiment design**. The hypothesis should consider the feasibility of its testing (defined in the experiment design) to avoid a not verifiable hypothesis.

Some examples of hypothesis are the following:

1. *Null hypothesis*: daily exposure to extremely high air pollution leads to cancer.
2. *Complex hypothesis*: there is no significant change in the health of a person during occasional exposure to extremely high air pollution.
3. *Empirical hypothesis*: roses watered with liquid Vitamin B grow faster than roses watered with liquid Vitamin E. (Here, trial and error is leading to a series of findings.)

In the **experiment design** phase, data scientists must define the next steps of the experiment and in which way it will be carried out, they must define which data are available for the analysis, which data must still be collected and where to store them.

In the **data management** phase data scientists focus on **data collection, preparation and storage** activities, with these activities influencing each other reciprocally. Choosing an unsuitable **data storage** tool based on the data available can lead to difficulties in future analysis (e.g., a good data management system for streaming data might not be good for static data). This must relate to the **data preparation** phase which is the process of cleaning and transforming raw data prior to analysis and often involves reformatting data, making corrections, and combining data sets to enrich data. The data management phase is the most time consuming one, in a typical data science project 80% of the time is spent on collecting datasets, cleaning and organizing data, as shown in Figure 2.

The **hypothesis testing** phase is the phase where data scientists apply **statistical evidence/data analysis/machine learning algorithms** in order to retrieve interesting information towards the initial hypothesis from the initial data. The possible techniques in this step are various and depend both on the dataset and on the hypothesis (e.g., if we want to understand what increases air pollution in metropolitan area a statistical model can be more useful than a predictive machine learning algorithm).

In the **result evaluation** phase, with the output of the previous step, data scientists can extract **numerical values** that support or reject the initial hypothesis. Scientists usually write a report on the project with their **interpretation of results**, create **data visualizations** to communicate results to stakeholders and they can **publish results** expanding the general level of knowledge on the matter.

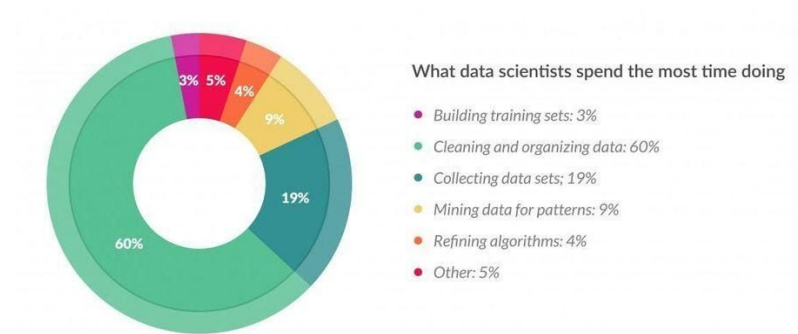


Figure 2. Time division in a data science project³

2.2 Analysis of Pollution Citizen Science Projects from a Data Science Perspective

2.2.1 Source of Citizen Science Projects

In order to find citizen science projects, we used two sources:

- SciStarter (<https://scistarter.org/>) is a platform that connects volunteers with citizen science projects and as of November 2019, it contains more than 1200 projects. The organization's primary goal is to break down barriers preventing non-scientists from fully engaging in scientific research.
- Wikipedia citizen science list (https://en.wikipedia.org/wiki/List_of_citizen_science_projects) lists approximately 300 projects, some of them are completed projects and there are also projects present in SciStarter.

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

2.2.2 Selection of Projects of Interest (Pollution-related)

The goal of our citizen science project exploration was to analyse citizen science projects that primarily deal with the pollution problem. For the selection of the projects we applied a filter on this topic. The projects of interest were the projects that are built around the problem of “pollution” which could be, e.g., water, air or ground pollution.

³

<https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#70cf126f6f63>

D4.1 Conceptual architecture and delivery plan



On the Wikipedia list we filtered the projects by the attribute “discipline”, looking for those projects that have pollution or air quality or water quality in that field as shown in Figure 3, in this way we reduced the projects from approximately 300 to 20 projects of interest.

Project name	Discipline(s)	Sponsoring organization(s)	Area	Began
Hubble Asteroid Hunter	Astronomy	ESA ESDC	Worldwide	2019
Muon Hunters	Astronomy	NSF, Asterics	Worldwide	2017
Zwicky's Quirky Transients	Astronomy	Zwicky Transient Facility	Worldwide	2019
COSMIC	Astronomy	JPL	Worldwide	2019
AgeGuess	Biology, Aging	Center for Research and Interdisciplinarity, University of Southern Denmark	Worldwide	2012
Agent Exoplanet	Astronomy	Las Cumbres Observatory Global Telescope Network	Goleta, CA, USA	
Air Quality Eggs	Air Quality, Pollution , Climate Change, Health	WickedDevices, LLC	Ithaca, NY, USA	
Air Quality				

Figure 3. Snapshot of Wikipedia citizen science list

On Scistarter we applied a filter on the projects keeping only those which had pollution in the title, in the description or in one of the keywords as shown in Figure 4, in this way starting from more than 1200 projects we reduced the projects of interest from SciStarter to 71.



Science we can do together.

Project Finder

Projects to do while...

Select a topic

only projects that...

Age group(s)

clear form

find projects

1-11 of 72

Industrial Smoke Hunting

Goal Train an Artificial Intelligence system to detect industrial smoke emissions automatically

Task Identify and label industrial smoke emissions from a collection of video clips

Where Online

Figure 4. Snapshot of SciStarter project finder feature

After removing the overlapping projects between Wikipedia and SciStarter we applied another filter, removing projects where pollution is not relevant in the analysis (e.g., <https://scistarter.org/platypuswatch-gold-coast-2>), projects expired with no data (e.g., <https://scistarter.org/cyber-citizen>) and duplicates. Also, for projects with the same domain but in different locations we kept only one of them (e.g., <https://www.curio.xyz/explore/missions/67> and <https://www.curio.xyz/explore/missions/66>). At the end of this process merging the two lists (Wikipedia and SciStarter) the projects selected were 48⁴.

2.2.3 Metrics of Analysis

In order to analyse citizen science projects from a data science prospect we aimed to compare which steps of a typical data science project (Figure 1) are used in a citizen science project.

The steps of interest are the following:

⁴ The list of projects together with the collected data can be downloaded from https://www.dropbox.com/s/4i5kzun42aff8j1/DS_analysis-of-pollution-projects.xlsx?dl=0.



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

2.2.4 Problem Definition

In the problem definition phase we can analyze two macro steps of a data science project pipeline, we omit the observation of reality activity since it should be implicit in any project and the experiment design since every citizen science project analysed, implicitly designed the project at least up to the data storage phase.

Table 1. Problem definition phase analysis

	Explicit problem identification/goal objective	HP formulation
Yes	48	2
No	0	46

Starting from the problem definition step the projects analysed show a clear preference to share with the citizen scientists a generic problem or a vision rather than a clear hypothesis to demonstrate, compared to a data science project that could lead to some advantages (e.g., it doesn't limit the research to an unique problem leaving flexibility to use of data) but also to some disadvantages, in fact data collection might be an end in itself, since there is no clear objective researchers need to adapt to present data.

2.2.5 Data Management

In the data management phase we analysed the cycle of the data from its acquisition to its storage to prepare it for further analysis.

Table 2. Data management phase analysis

	Data collection	Data preparation	Data storage
--	-----------------	------------------	--------------



Yes	45	8	25
No	3	3	1
No information	0	37	22

Almost all projects analyzed require an active participation of the user in the **data collection** step. The data collection is usually similar for certain projects and we found during the analysis that there are mainly four different types of data collection:

- “Citizen photographs” where volunteers are asked to take pictures of a certain object (litter, stars, water, etc.)
- “Citizen surveys” where volunteers take part in the citizen science project through surveys (e.g., report of chemical accident, report of biological analysis)
- “Citizen sensors” where volunteers help the project by buying or renting the sensors, doing their maintenance and using them during common activities (e.g., walking with a sensor on the wrist that measures pm10 levels)
- “Citizen sampling” where volunteers are asked to take a sample (e.g., an animal, a cup of water from the river, etc.) and send it to the organization in order to be analysed

These four categories are often preceded by a task or an activity or can be carried out only in particular situations (e.g., find a constellation, find litter pollution on the shore). Moreover categories are often mixed, some data collection tasks require more activities for a single data collection activity (e.g., take a photo and complete the survey) so it is difficult to categorize the activities but it is still possible to understand which type of tasks are often asked of volunteers.

Information on the **data preparation** step in citizen science projects is usually not shared with the public, so we were not able to identify the activities done by the organizations to prepare collected data. It would be interesting to analyse a bigger sample to understand why citizen science projects do not share this information, but also to understand at which level of quality data preparation tasks are performed, since in this way it would be possible to define which level of reliability citizen science data have. It must be mentioned that few citizen science projects, mainly the ones with sampling tasks, use a “Quality Assurance Project Plan” which is a written document outlining the procedures a monitoring project uses to ensure the data it collects and analyses meets project requirements. This could help to assess the quality of the data collected but still there is no information about data preparation in information systems.

In the **data storage** phase the project organization stores the data that the citizen science project has collected. As seen in the previous table, 50% of the projects don’t give any information on how or where the data are stored, but we can assume that almost all the projects examined use a type of data management method due to the fact that the citizen science projects examined rely on data collected by citizens.



Analysing more in depth the 25 projects that share information on how they store data we can divide more around their policy of publishing data.

Table 3. Data availability analysis

Project data		
Not downloadable	Downloadable by request	Downloadable
4	11	4

2.2.5 Data Analysis

The projects considered for our analysis started with a general goal instead of a clear hypothesis, project owners don't have a single way to analyse data nor a specific goal, moreover it often happens that the analysis of data is carried out by external institutions which leads to different types of analysis for the same citizen science project.

For these reasons in this preliminary study, it was not possible to look deeper in the activities done by researchers in the data analysis phase. It would be interesting to extrapolate this information from reports made by organizations, thus excluding results performed by external institutions in order to find at which level of analysis citizen science projects are carried out by their owners.

We haven't collected enough information on the data analysis phase but a superficial observation seems to show a prevalence of statistical evidence in self-created reports.

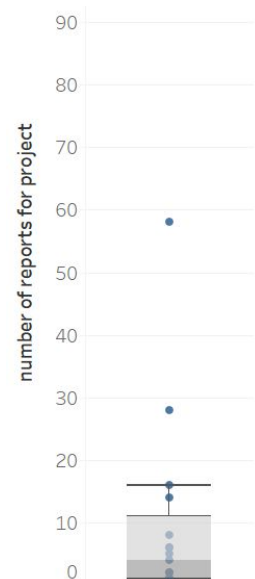
2.2.6 Result Publication

For the result publication phase we analysed whether the citizen science projects have published any results concerning the project. The results could be a report, a publication, an article that shares to the public the results of the citizen science project. In case there are no written reports we tried to find whether the project shared some sort of information concerning the project proceeding to the public in order to understand which kind of information is shared to volunteers. The results are the following.

Table 4. Result publication phase analysis

Publicly available reports	
Yes	No
19	29

Figure 5. Distribution of reports per project





For the 19 projects with publicly available reports we had a deeper look to understand how a user can find the reports and how many reports are usually available for citizen science projects concerning pollution.

Table 5. Reports positions on webpage

Publicly Available Results		
On site	Section that contains external link	no specific section
4	11	4

Considering the small sample of projects and the variability of the publications it doesn't seem possible to make a conclusion nor to find a pattern on the type or number of reports/publications/articles that a citizen science project shares with the public. It seems that most of the projects that share reports/publications/articles prefer to guide the user with a specific section where they can find them, also from the boxplot we have that the mode value (the most frequent value) is 1 report for project but the median value is 4 reports/publications/articles for a project, which shows, as anticipated in the problem definition section, that many citizen science projects don't focus only on one hypothesis but they challenge a general problem, which can be analysed in many ways. It must be also mentioned that projects have a great variability of what they define as a "publication", some projects publish an annual report, others share research done by independent institutions, others publish descriptive analysis about the data they have.

For projects that don't have reports we can still divide them by projects that share results of analysis with numerical values or data visualization of the results. During our evaluation it appeared clear that due to the characteristic of citizen science projects on pollution that often require geo-localization, project owners tend to share data with a view of the data on a map more than numerical results. In, fact starting from 29 citizen science project without reports, the results are the following:

Table 6. Data visualization analysis

Projects with no reports	
With map data viz	Without map data viz
20	9

This analysis could be interesting if carried out with a larger sample since it could show that many citizen science projects (in our case concerning pollution) would like to share more insight with their



user base, but they don't have the resources or the technical knowledge necessary to carry out experiments supported by scientific data.

2.2.7 Example of Citizen Science Projects with Implemented Data Science Tasks

Using the Data Science pipeline of Figure 1 we present an example of a citizen science project concerning pollution, exemplifying how the project addresses data science tasks. The project considered is <https://www.globeatnight.org/>, this project's goal is to raise public awareness of the impact of light pollution. The project owner asks volunteers to complete a survey <https://www.globeatnight.org/webapp/> after the observation of specific constellations during the night.

These data are available to the public in CSV format and can be downloaded from the citizen science project web page. The citizen science project then offers users an interactive infographic on site <https://www.globeatnight.org/infographic> with some statistical summary on the usage of the app (N.B. the goal of the project is to raise awareness so they are interested in these results).

Another citizen science project that could be represented by a data science pipeline is the Pieris project <http://www.pierisproject.org/>. The project doesn't deal directly with pollution, but it is one of the best examples of how a citizen science project can be carried out as a data science project.

The project starts with observations of nature and with the **identification of the problem**: "this butterfly (*Pieris rapae*) has invaded many parts of the world and is now one of the most successful and abundant butterflies on the planet".

Their **goal** is "to partner with the public to create the most comprehensive collection of a single species of butterfly that will act as a powerful tool for studying how organisms adapt to changes in their environment".

The project has **multiple hypothesis** to test:

- "...learn whether the cabbage white butterfly has invaded multiple times and if so, from what countries"
- "...understand how the genomes of these butterflies has been shaped by their environment"
- "...look at how the shape, size and color of these butterflies change depending on where they (you) live."
- "...how organisms respond to changes in their environment. This information will help us predict how other species might respond to similar changes, something we still don't know for most species"

In the **experiment design** they plan to study the genes of many butterflies in order to test different hypotheses, they also plan how to collect data (<http://www.pierisproject.org/participate.html>). Indeed they ask volunteers to capture a precise type of butterfly, freeze and send it to their laboratories with a detailed method. In the **data collection** phase citizens collect a sample, in this



case a butterfly, and send it to the laboratory where they perform chemical and biological tests in order to collect the data. For **the data storage** phase we know that data is available on request “We believe in open science, so all data will be made publicly available to anyone who wants it” (<http://www.pierisproject.org/faq.html>). There is also an explanation of the **data preparation** phase (<http://www.pierisproject.org/progress.html>) where they explain the use of many copies of genome mapping in order to retrieve errors in machine analysis. They then analyse the data through **descriptive analysis** (<http://www.pierisproject.org/resultsinvasionhistory.html>), a **machine learning model** “using an evolutionary model that tries to predict how many groups there should be”, group butterflies using a k-mean model, and **test alternative hypothesis** looking for alternative scenarios to strength their initial assumption. “For each scenario, we can simulate this process thousands or even millions of times”. This information is contained in a **public report** in a **specific section of their website**. The report also contains many **data visualizations** and the interpretation of **numerical results**.

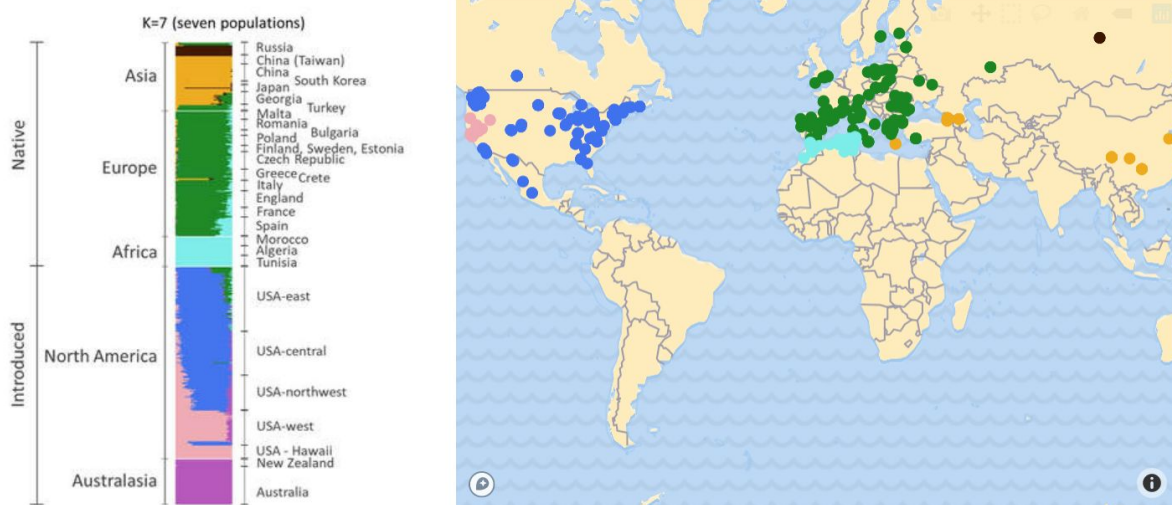


Figure 6. Snapshot of a Pieris project data visualization

2.3 Summary of Analysis

Problem definition:

- In all the analyzed project there is a definition of the problem or the final goal of the project
- Hypothesis formulation is mostly unclear or nonexistent in 96% of projects

Data management:

- In 93% of the cases examined the data collection is deployed to volunteers
- In 77% of projects there is no information about data preparation and only in 16% of projects the activity done to prepare data is made explicit
- In 46% of projects there is information about data storage



- Of these, 88% of projects share their data with the public

Data analysis:

- There is not enough information on data analysis, but a superficial observation shows an interesting component of descriptive statistical analysis in self-owned reports.

Result publication:

- In 60% of projects examined there are no publicly available reports/publications/articles
 - Of these, 69% of projects examined share at least a visualization of their data with the public
- The median number of reports for projects that have at least one report is 4
- The mode number of reports for projects that have at least one report is 1
- In 61% of projects with at least one report, reports are reachable through an external link in a specific section

2.4 General Recommendations

For citizen science projects that wish to aim for a more data-driven approach in implementing the projects, we outline below a set of recommendations for each phase in the data science pipeline applied in the context of citizen science.

Problem definition:

- Projects should state one or many explicit hypotheses or a definition of the practical use of data (e.g., study pollution geographical distribution in New York) in order to give a direction to the citizen science project
 - An alternative can be the possibility to state a general goal but offering the possibility to perform a punctual analysis to external researchers and institutions, deploying hypothesis definition to other researches.
- Good citizen science projects should always inform users what new science they are contributing to understanding
- In the experiment design it should be stated if data and results will be published in order to have a transparent relationship with volunteers

Data management:

- The public should be informed about the data pre-processing criteria and process, in order to evaluate the quality of the data, and consequently the quality of the research output
- In case of public data, it should be easy for the user to obtain data from the website or on request, or via an open data portal

Hypothesis testing:

- The correct method (Data analysis, ML algorithm, statistical evidence) should be applied depending on the project and the initial hypothesis, the only recommendation here is to



explore the dataset through preliminary statistical analysis in order to understand the characteristics of the data

Result publication:

- According to the type of public availability defined in the experiment design, projects should have a section that leads to results which can be on site or referring to external links
- Numerical results, data visualization and results interpretation should always be present in a written report in open access
- Software produced as part of the project should make the code available as open source software

2.5 Final Consideration

Citizen science projects that challenge pollution (light, air, water, litter) offer a new way to collect data exploiting economies of scale with the help of volunteers spread across the globe. The data often is not analysed by project owners but by other institutions. Due to the volume of information available but the lack of tools and knowledge to analyse it, delivering a support for citizen science projects in order to analyse their own data and retrieve meaningful information would be an important step to increase pollution scientific research and knowledge.

3 CONCEPTUAL ARCHITECTURE OF THE ACTION CITIZEN SCIENCE TOOLKIT

Based on the analysis and the recommendations from the previous section, we outline below the key aspects of the initial conceptual architecture of the ACTION toolkit. The citizen science toolkit includes methodologies, methods, tools, services and other resources that respond to a wide range of citizen science characteristics: online and offline activities, various and evolving goals and scopes, as well as different stages of development, from early ideas, to initiatives that have resulted in scientific publications and other forms of impacts.

The toolkit will support the lifecycle of the research process which is:

- Problem framing
- Research implementation, including
 - Research design
 - Data acquisition
 - Data analysis
 - Share and communicate results
 - Evaluation
- Conclusion and sustainability
- Policy impact

For each stage of the life cycle, we will offer methodologies, and tools to help users maximize and improve the results of their citizen science projects.





The objective of the first stage, **problem framing**, is to define and gather background information on the problem, as well as engage relevant stakeholders. This stage will be supported with documents and activities to outline the best way to define and narrow down the topic. This stage will be supported by tools such as fora and all our ideas platforms.

In the second stage, **Research implementation**, the citizen science experiment will be planned and implemented. This encompasses multiple substages, which will be supported by a variety of tools and resources. Through all substages, we will offer guidance on the best ways To engage citizens.

- During the **research design** stage, projects will create their research question, define their research design, and develop appropriate data gathering instruments. We will provide guidelines for task design, recommendations for quality assurance, as well as data gathering instruments.
- In the **data acquisition** and **data analysis** stages, projects will acquire, curate, process, analyse and interpret their data. We will offer tools to support these activities, including the CONEY toolkit described below, a data management tool, as well as guidelines and recommendations.
- To help projects to **share and communicate results**, we will offer a live dashboard and publishing portal, as well as data visualisation tool. We will create a methodology explaining best practices to share and publish results, and enable projects to pack all the material generated in the project as a shareable research object.



To support the **evaluation** of projects and help them measure success, we will develop an impact-self-assessment tool.

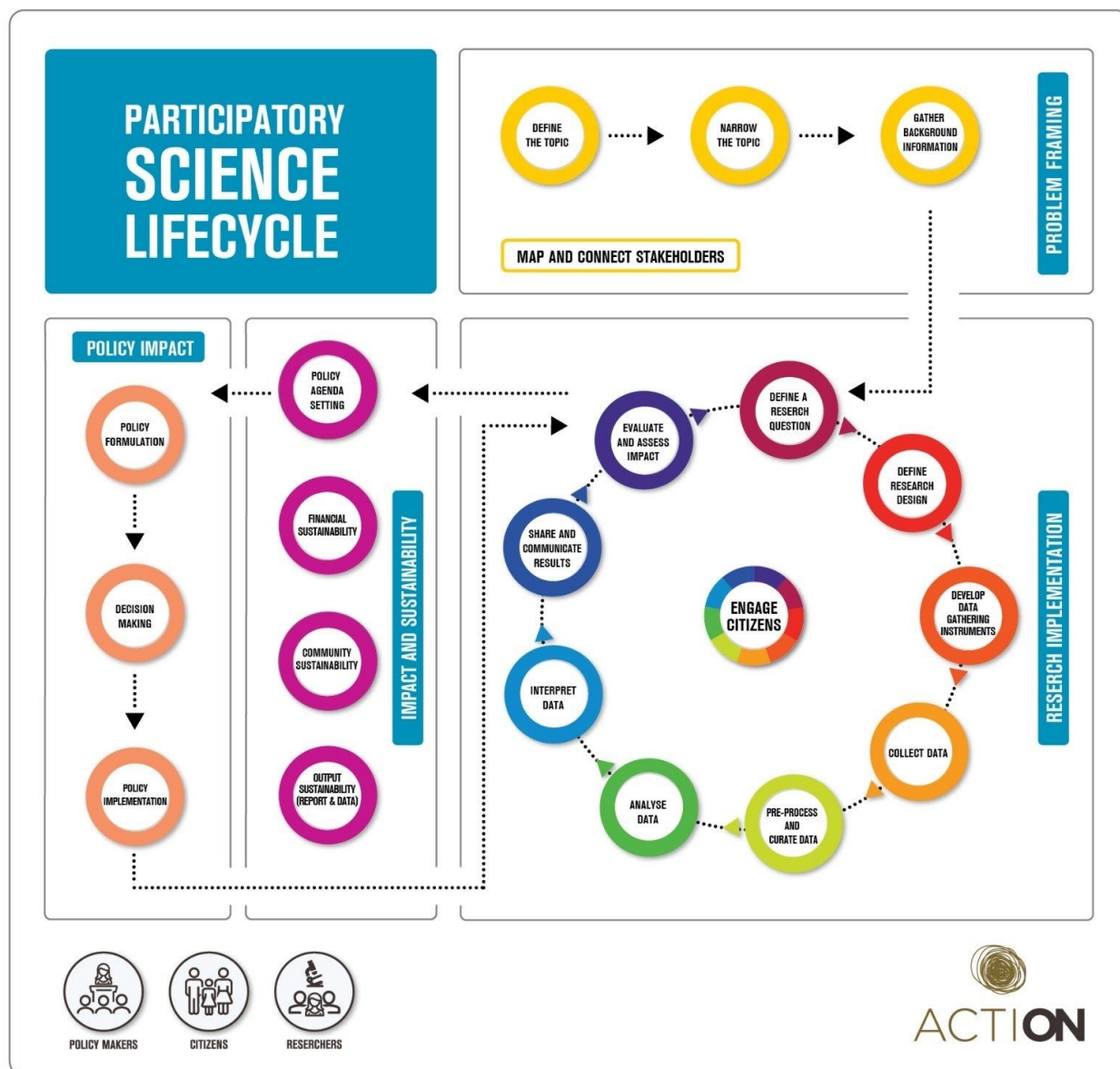


Figure 7. Conceptual architecture for the ACTION toolkit

In the third stage, **Conclusion and sustainability**, the goal is for projects to find and use routes for policy agenda setting, and achieve sustainability of their community and data, as well as finances. We will support these steps with guidelines, recommendations and webinars. We will also develop a data management plan creator, and offer an open data portal.

In the last stage, **Policy impact**, the goal is for projects to use their work to help formulate policies, influence decision-making, and the implementation of policies. We will support this with activities and recommendations.



4 DELIVERY PLAN OF THE ACTION TOOLKIT

ACTION will develop a series of tools and resources, case studies and activities to enable citizen science projects. Some of these have been part of the ACTION plan to begin with; others will be developed following our interactions with the pilot projects. Further tools and resources will be recommended, where sufficient offers are already available from external sources.

The toolkit will follow the citizen science process (Figure 7 above), and provide tools and resources, case studies and activities for each step in this process. Table 7 below indicates the schedule for currently planned resources. We outline some of the tools below. Further resources will be added following discussions with pilots after the accelerator launch in February 2020.

Table 7. Delivery plan overview

		Q1 2020	D	Q2 2020	D	Q3 2020	D	Q1 2021	D	Q4 2021	D
Problem framing	<i>Define the topic</i>										
	<i>Narrow the topic</i>										
	<i>Gather background information</i>										
Research implementation	<i>Create a research question</i>										
	<i>Define research design</i>					Guidelines for task design	5.1				
	<i>Develop data gathering instruments</i>					Data gathering instruments and guidelines	6.2	Guidelines, recommendations and tools for quality assurance	5.3		
	<i>Collect data</i>	Coney: Data collection	5.1								
	<i>Data pre-processing and curation</i>									Coney: Quality assurance CS Templates: Data Management executor	5.3 4.3
	<i>Data analysis</i>					Coney: Analysis motivation	5.6 6.2				
	<i>Data interpretation</i>										



	Share and communicate results				Live dashboard and media publishing portal Coney: data visualisation	4.8	
	Evaluate and assess impact			Impact self-assessment tool	6.1		
Conclusion and sustainability	Policy agenda setting						
	Achieve financial sustainability			Financial sustainability guidelines; Sustainability webinar	5.8		
	Achieve community sustainability	Guidelines for incentives and motivations in citizen science	5.6	Guidelines for inclusiveness and community sustainability; Engagement recommendations	5.4		
	Achieve output sustainability (reports & data)	CS Templates: Data Management Plan creator	4.2	Open data portal	4.4	Research Object Packer Coney: Research Objects	4.6
						Research Object Packer Open Data Portal Data Management Plan creator	4.7 4.5 4.3
Policy impact	Policy formulation						
	Decision-making						
	Policy implementation						

4.1 CONEY Toolkit

CONEY (CONversational survEY) is a toolkit to administer questionnaires in a chat-like form, so that the compiler experiences it as if it was a conversation with another person rather than a pure survey. The survey is designed as a pre-defined conversation flow (with the possibility of branches, as explained in the following) that is experienced by the compiler through a chat interface. This approach differs from those based on the adoption of chatbots and intelligent agents which imply some sort of natural language understanding (NLU) and artificial intelligence (AI).



CONEY aims to focus on the user experience, in the sense that the toolkit wants to provide a new and innovative interaction pattern from the compiler point of view; this of course implies a change in the survey design process, which not only is oriented to select and formulate the questions to collect relevant data, but should also take into account the different interaction pattern in the overall formulation of the questionnaire.

CONEY is composed of different components, as illustrated in Figure 8 and explained hereafter.

Briefly, CONEY Create and CONEY Inspect are the modules accessible by people designing, managing and analysing the survey, whereas the CONEY Chat component is the end point accessible by survey compilers.

CONEY Collect is the back-end component that gathers and stores the data related to both the survey design and the survey responses that acts behind the scene to ensure a reliable and efficient data collection process.

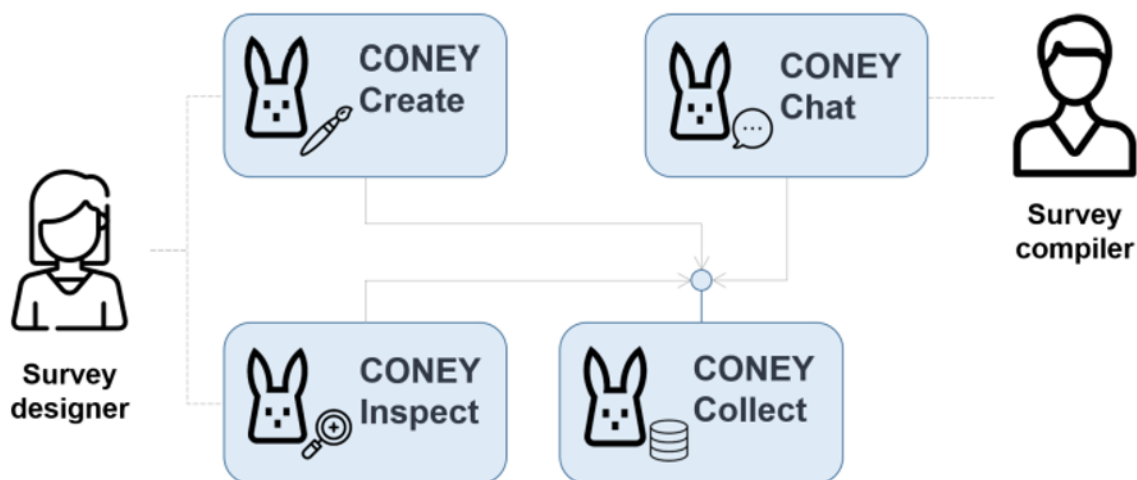


Figure 8. Components of the CONEY Toolkit

The tool is currently under development and it will be accessible at this link <https://survey.actionproject.eu/coney/>.

4.1.1 CONEY Create

CONEY Create is the graphical editor for the survey designer to create the questionnaire in the form of a conversation; inspired by Friedhoff (2013), the editor is a drag-and-drop tool which allows for the design of a questionnaire in a “hypertextual” fashion with text, question and answer blocks (respectively, blue, yellow and green boxes in the picture) and the possibility to create alternative branches depending on the compiler choice (i.e., depending on the chosen answer, the conversation flow continues in different ways, for example to ask clarification questions).



The editor offers plenty of question types: single choice questions with different answer visualizations (buttons, star-rating, emoticons, slider), multiple choice questions (with check-box answers) and open questions; as explained above, since there is no use of AI, the open questions allow for free-text answers, but no elaboration of the provided text is made: compilers' answers are only collected for post-hoc analysis. The “conversation flow” approach allows for storytelling, enhanced by the possibility to include colloquial and multimedia content.

Finally, question blocks can be annotated with a label to indicate the respective investigated latent variable, while answers can be annotated with the respective numerical coding: this kind of information is reused at answer analysis time, as explained in the CONEY Inspect component. The editor itself does not constrain the survey designer to the use of a specific language style, and they have the responsibility for the “storytelling” design. However, the tool allows for saving and reusing question/answer patterns across different surveys, as well as for cloning an existing survey to adapt it to a different usage scenario.

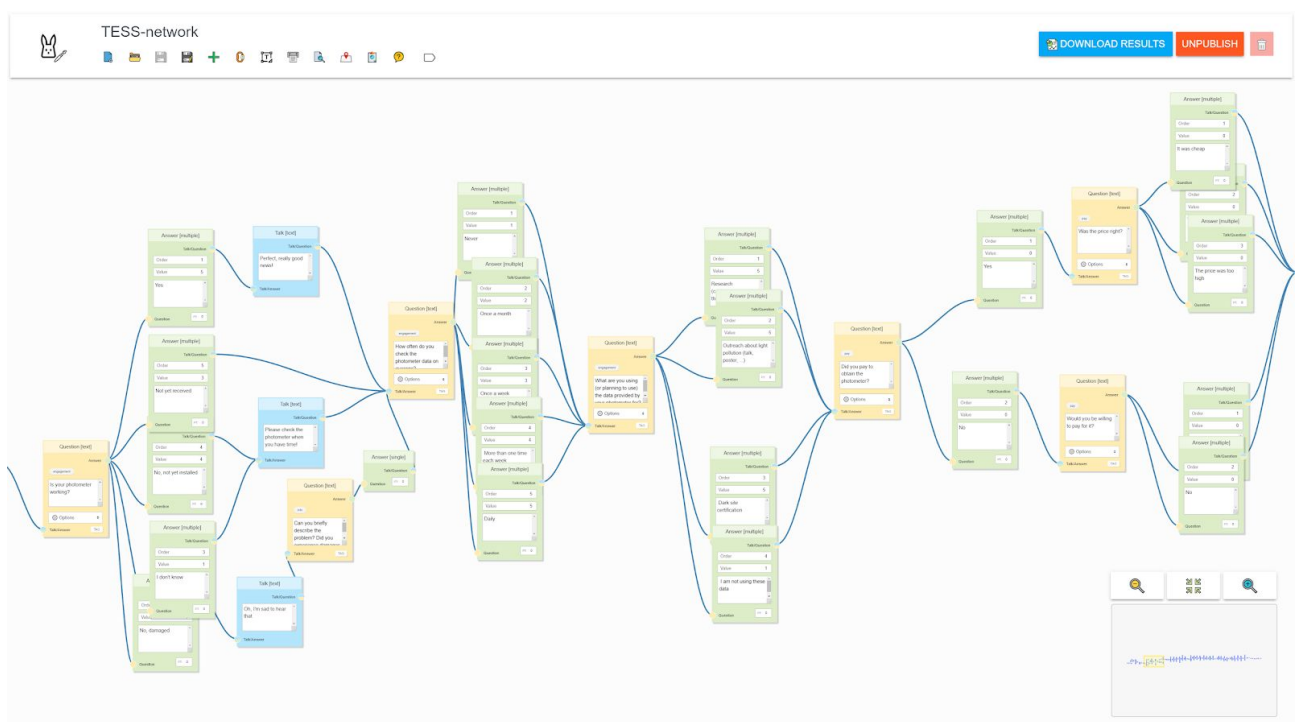


Figure 9. CONEY Create, component to design conversations with multiple flows

CONEY Create can be used in the phase of development data gathering instruments. Researchers and citizen scientists can design and create their own survey that best suits their current research questions by selecting the most effective interaction patterns and survey flow. More details on how to set up a survey can be found in Deliverable 5.1.

4.1.2 CONEY Chat



CONEY Chat is the Web-based user interface to administer the designed survey to compilers in the form of a chat; in this case, the inspiration comes from chat clients and popular mobile apps like Whatsapp, Messenger or Telegram. The user experiences a seamless flow thanks to the personalized path based on his/her answers. Furthermore, even when the survey is a purely quantitative research method (with closed questions with numerically-coded answers), the interaction style makes it resemble an interview, i.e. a qualitative research approach.

A demo survey can be experienced at this link <http://bit.ly/try-coney>.

Figure 10. CONEY Chat, interface for survey compilers

CONEY Chat can be very helpful in the "Collect data" phase. The link with the survey to be filled can be easily sent to all the survey compilers allowing a quick and reliable data collection. Survey compilers can experience the filling of a friendly and informal questionnaire.

4.1.3 CONEY Inspect

CONEY Inspect component is the dashboard application for the survey analyst to simplify the statistical analysis of the answers collected through the conversational survey.

CONEY Inspect can be used in the "Analyse data" phase. The dashboard offers basic indicators, like the number of started and completed surveys, the average values for latent variable, the



distribution of compilation time and the histograms of the answers per question. Survey designers and analysts can have an overall view on the survey campaign both to monitor trends and performances during the filling phase and to analyse data at the end of data collection process. For more detailed analysis, the dashboard allows for the download of all collected answers in the form of a CSV file, that can be deeper analysed by coding ad hoc scripts.

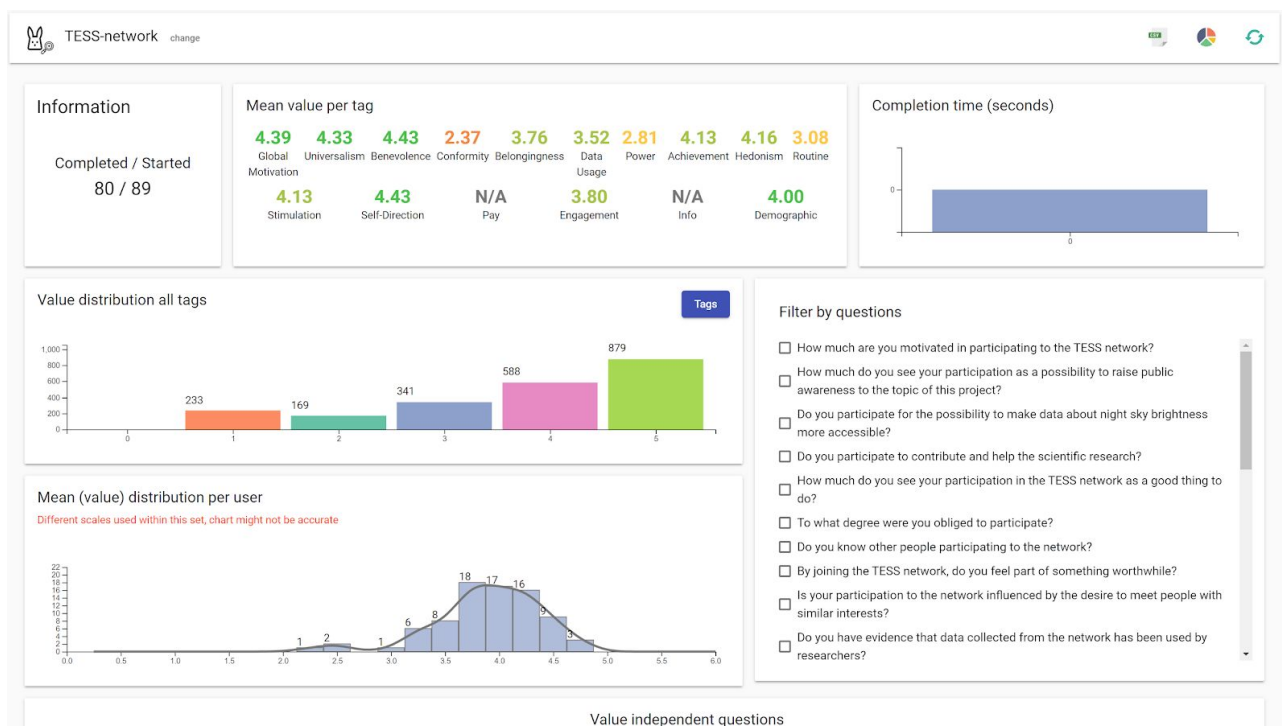


Figure 11. CONEY Inspect, dashboard showing survey results

4.2 Templates

In the context of the ACTION project, templates are mechanisms to build citizen science projects following the *good practices* that will be analyzed in ACTION, in particular in WP5. These templates can be either tools (software) or guidelines.

After the analysis of projects in Section 2 about citizen science projects, we discovered that many of the projects suffered a lack of data management. Due to this fact, we have decided to create two tools that will help users to deal with this problem.

The first tool, the Data Management Plan Creator, is a Web application to create Data Management Plan (DMP) documents based on a simplified questionnaire. In addition to the questions, the system will generate the content of the document based: i) on the data stored in our system and ii) following *good practices* of FAIR principles.

The second tool, the Data Management executor, is a simplified and configurable workflow to allow the projects present in ACTION to run data operations aligned with the information provided in the



DMP (defined in the previous tool). Based on a questionnaire, users will be able to configure the tool to cover the main steps of the data lifecycle from the extraction of the data to their publication in a repository.

More information about the ACTION template tools can be found in Deliverable 4.2.

5 SUMMARY AND FUTURE WORK

In this deliverable we provided a conceptual architecture for data-driven citizen science projects. In order to do that, we started by analyzing the current practices of citizen science projects related to pollution from a data science perspective. For the citizen science projects that aim to be more data-driven, we outlined a number of recommendations in that direction. Based on that, we devised the architecture of the ACTION toolkit with a number of areas to be addressed, together with a delivery plan of the toolkit. The results of this deliverable contribute to providing a range of technology and assistance services relevant to the digital infrastructure citizen science initiatives.

This deliverable analyzed projects specifically related to pollution. It would be interesting, as part of future work, to extend the analysis to other citizen science project from other domains.

6 REFERENCES

Friedhoff, J., 2013. Untangling twine: A platform study. In: DiGRA conference, pp. 1–10.