Analysis and comparison of local, national and European citizen science platforms

Varas Paneque, Núria

Curs 2020-2021

Directores: Patricia Santos i Miriam Calvera

GRAU EN ENGINYERIA MATEMÀTICA EN CIÈNCIA DE DADES

upf. Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Treball de Fi de Grau

ANALYSIS AND COMPARISON OF LOCAL, NATIONAL AND EUROPEAN CITIZEN SCIENCE PLATFORMS

BACHELOR'S THESIS

Núria Varas Paneque

Thesis directors:
Patricia Santos and Miriam Calvera

Bachelor's Degree in Data Science Engineering

Course 2020-2021

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Acknowledgements

I would like to thank my thesis mentors Patricia and Miriam, who have been available to provide their help at any moment. Also, to the CS Track project, and especially to the experts that have invested their time in checking the webpage to provide feedback that has been very meaningful. And of course, to the three platforms “EU-Citizen.Science”, “Observatorio de la Ciencia Ciudadana” and “Oficina de Ciència Ciutadana” for allowing to get the necessary information from their content to carry out this project and fulfil the proposed objectives.

Finally, to my family and friends for supporting me at any moment during the whole degree and specifically during the process of this project. It has been a complicated period with many doubts and concerns, but from which I have learned a lot, not only as an engineer but as a person.

Abstract

Citizen participation in science, known as citizen science, is emerging as support for formal science, so that citizens learn and contribute to the local policies that are carried out. This work aims to analyse and compare the current situation of citizen science in Barcelona, Spain and Europe through its web platforms, with special emphasis on its relation with education. The objective is that participants and other involved agents can learn about the current situation of this science at a local, national and international level. To do this, information will be extracted from the Barcelonian, Spanish and European citizen science platforms, analysing them through computational techniques that will permit to compare them, draw conclusions and indicators, showing the results in interactive dashboards.

Resumen

La participación ciudadana en la ciencia, conocida como ciencia ciudadana, está emergiendo como apoyo a la ciencia formal, de forma que los ciudadanos aprenden y contribuyen en las políticas locales que se llevan a cabo. Este trabajo tiene como objetivo analizar y comparar la situación actual de la ciencia ciudadana en Barcelona, España y Europa a través de sus plataformas web, haciendo un especial énfasis en su relación con la educación. El objetivo es que participantes y otros agentes involucrados puedan conocer la situación actual de esta ciencia a nivel local, nacional e internacional. Para ello, se extraerá la información de las plataformas barcelonesa, española y europea de ciencia ciudadana, analizándolas mediante técnicas computacionales que permitirán compararlas, extraer conclusiones e indicadores, mostrando los resultados en dashboards interactivos.

Resum

La participació ciutadana a la ciència, coneguda com ciència ciutadana, està emergint com a recolzament a la ciència formal, de manera que els ciutadans aprenen i contribueixen a les polítiques socials que es duen a terme. Aquest treball té com a objectiu analitzar i comparar la situació actual de la ciència ciutadana a Barcelona, Espanya i Europa a través de les seves plataformes web, fent un especial èmfasi en la seva relació amb l'educació. L'objectiu és que participants i altres agents involucrats puguin conèixer la situació actual d'aquesta ciència a nivell local, nacional i internacional. Per aconseguir això, s'extraurà informació de les plataformes barcelonesa, espanyola i europea, analitzant-les mitjançant tècniques computacionals que permetran comparar-les, extreure conclusions i indicadors, mostrant els resultats a dashboards interactius.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Context

"Citizen science is the use of scientific methods by the general public to ask and answer questions about the world and solve problems of concern" [1]. This practice is emerging as a support for formal science since scientists and citizens collaborate to produce new knowledge, which is done by and for the community. However, the role it plays in modern societies is not fully understood, and still has to be exploited.

As National Geographic [2] states: "Though citizen science is a relatively new term, people have been participating and contributing to scientific research for years". So, there is a long history of the community contributing to science, from programs about bird migration such as "Christmas Bird Count" [1], to the study and observation of the cosmos, as "AAVSO" [2] does. Nevertheless, modern advances in technology such as the Internet and smartphones make it really easy for people to share and contribute information. For example, the "Globe at Night" [3] project analyses light pollution by using citizens' observations, who only need a computer or smartphone. And the "CanAirIO" [4] project uses mobile and static sensors to measure air quality with cell phones and low-cost technology.

That is the reason why the number of citizen science programs has strongly increased, leading to the generation of several platforms such as "EU-Citizen.Science" [5] (from Europe), "Observatorio de la Ciencia Ciudadana" [6] (from Spain) or "Citizen Science Organization" [7] (from the United States). All of them work as repositories that pretend to make it easier for everyone to find out about projects they can join and also try to promote citizen science between citizens of the region.

There exists a European research project called "CS Track" [8], which focuses on obtaining conclusions on citizen science in all its aspects: activities, volunteers, motivations, etc... One of the main aims of the project is to build a database with data extracted from citizen science platforms and interviews to create datasets that will help to answer some defined questions and draw conclusions [3]. Therefore, the purpose of "CS Track" is to investigate how to apply computational techniques to extract data regarding CS and analyse it (see Appendix A for more information about "CS Track").

---

[1] The Christmas Bird Count is one of the longest-running community science projects, studying the population of birds since 1900. (2021, January 23) https://www.audubon.org/conservation/science/christmas-bird-count .

[2] The American Association of Variable Star Observers (AAVSO) studies the variable stars since 1911, with more than 100 years of citizen science. (2021, January 23) https://www.aavso.org/100-years-citizen-science .

[3] Globe at Night is an international citizen-science campaign to raise public awareness of the impact of light pollution by inviting citizen-scientists to measure & submit their night sky brightness observations. (2021, January 25) https://www.globeatnight.org/ .

[4] CanAirIO is a citizen science project using mobile and static sensors to measure air quality with cell phones and low-cost technology. (2021, January 25) https://canair.io/ .

[5] EU-Citizen.Science European website (2021, January 27) https://eu-citizen.science/ .

[6] Observatorio de la Ciencia Ciudadana en España website (2021, January 27) https://ciencia-ciudadana.es/ .

[7] Citizen Science U.S. government website (2021, January 27) https://www.citizenscience.gov/# .

[8] The aim of CS Track is to broaden our knowledge about Citizen Science and the impact Citizen Science activities can have. CS Track will do this by investigating a large and diverse set of Citizen Science activities, disseminating good practices, and formulating knowledge-based policy recommendations in order to maximise the potential benefit of Citizen Science activities on individual citizens, organisations, and society at large. (2021, January 27) https://cstrack.eu/about/.

Since this work is supervised by two members of the "TIDE-UPF" [9] group that belong to the "CS Track", it has been proposed to extract information from the Spanish and European platforms (finally also the Barcelonian one[10] was analysed) in order to analyse the current situation of citizen science and extract conclusions. Hence, it is closely related to what they do but extracting the data from both ES and EU citizen science platforms and comparing them. Also, it will put a special emphasis on the relation between citizen science and education, studying the offered resources and the opportunities CS brings to formal education. All the advances done will be integrated into the European "CS Track" project and will serve to extend the already existing analysis.

1.2 Motivation

This project intends to examine how citizen science is spread through its platforms, analysing which role CS currently has in society through what is exposed in the platforms. Therefore, it will study the diffusion that citizen science platforms do at an international, national, and local level, to see if there are different perspectives and methods to do this dissemination. It will compare the information presented in each case and its formats. Moreover, it will study what type of projects are posted and which data of them is available, plus what other materials are shown, making a special emphasis on educational resources. Thus, it is intended to achieve a comparison of both the way of disseminating content and the current state of citizen science, especially at the European and Spanish levels.

So, it expects to enrich the knowledge people have of this science and support its divulgation. By extracting the information of three platforms, visualizing and interpreting it, interesting conclusions and indicators related to CS will be extracted, giving a closer and more visual approach for all kinds of public. Specific attention will be given to educational aspects to evaluate how citizen science is currently being used in formal education.

The reason to analyse more than one platform lies in the fact that it will make it possible to compare the current situation in Europe and Spain, leading to a much wider analysis. Since the European platform has been recently updated, this study will be very novel. Moreover, these web pages have different ways to extract data from them, so two distinct methods will be used.

1.3 Objectives

The main objective of the project presented here is to analyse the situation of citizen science in Spain and Europe, obtaining meaningful conclusions addressed to a wide public that may not be familiarized with the technology, and focusing on aspects and resources related to education and learning. In a more systematic way, the objectives of this work are:

- Studying the three platforms: European, Spanish and Barcelonian.

[9] TIDE-UPF group website (2021, January 27) https://www.upf.edu/web/tide .
[10] Ciència Ciutadana website (2021, January 27) https://www.barcelona.cat/barcelonaciencia/es/ciencia-ciudadana .

- Extracting data from the European platform "EU-Citizen.Science" [11] by using the existing API.

- Designing a tool to extract information from the Spanish platform "Observatorio de la Ciencia Ciudadana en España" [12] and the "Barcelona Ciència" [13] Barcelonian platform by programming a crawler and creating the corresponding databases.

- Analysing the data of the three platforms individually.

- Comparing the information of the three platforms.

- Visualizing the data and generating an interesting output where the obtained results can be seen so that anyone interested in citizen science can consult it and get an overview of what this project has done and what is the citizen science current state through what the platforms show. So, the idea is to generate a useful product containing the most important insights, in this case, an easily understandable web page.

- Extracting meaningful conclusions regarding the situation of citizen science on a national and international level.

1.4 Planning

The planning of a project is a very important part, as well as dividing it into the corresponding stages so that deadlines can be established. Therefore, this section is necessary for the good development of the project, including the segmentation of tasks with the corresponding objectives and dates. Before defining the steps, it must be mentioned that this planning must be strict but at the same time flexible on the unexpected events that may occur. So, the delivery dates should be fulfilled in order to reach the project up to date, though they are approximated.

Considering that the delivery of the provisional report is on the 18th of June, the delivery of the final report on the 2nd of July and the oral presentations from the 5th to the 16th of July, the planning will follow the next defined steps.

	TASK DESCRIPTION	DELIVERY WEEK	PHASE
1.1	Identification of the opportunity, the context and motivation.	18 – 24 January	INITIAL PHASE
1.2	Determining the objectives.	18 – 24 January	
1.3	Specification of the tasks to be performed and deliverables.	25 – 31 January	
1.4	Initial tasks in relation to specific aspects of the project: consultation of bibliographic sources.	25 – 31 January	

[11] EU-Citizen.Science European website (2021, January 27) https://eu-citizen.science/

[12] Observatorio de la Ciencia Ciudadana en España website (2021, January 27) https://ciencia-ciudadana.es/

[13] Ciència Ciutadana website (2021, January 27) https://www.barcelona.cat/barcelonaciencia/es/ciencia-ciudadana

2.1	Elaboration of the state of the art: documenting the project, fundamental knowledge to develop the project.	15 –21 February	**EXECUTING PHASE**
2.2.1	Environment set-up, study, and selection of the tools.	22 –28 February	
2.2.2	Extracting data from the European platform.	1 – 7 March	
2.2.3	Building a crawler to extract data from the Spanish and Barcelonian platforms.	15 – 21 March	
2.2.6	Analysing the data from the three platforms individually.	5 – 11 April	
2.2.7	Generating visualizations of the three platforms content.	12 – 18 April	
2.2.8	Analysing the data from the three platforms together.	26 April – 2 May	
2.2.9	Generating global visualizations with the three platforms content.	3 – 9 May	
2.2.11	Comparing the format of the platforms.	10 – 16 May	
2.2.12	Comparing and analysing the content of the platforms.	17 – 23 May	
2.2.13	Web page creation.	17 – 23 May	
2.2.14	Reporting the results and interpreting them.	24 – 30 May	
3.1	Evaluation of the whole process.	31 May – 6 June	**ENDING PHASE**
3.2	Drawing general conclusions.	31 May – 6 June	
3.3	Preparation of the final report.	7 – 13 June	
3.4	Creating the final presentation.	21 – 27 June	

Figure 1. Project planning.

In the beginning, it was stipulated to build the crawler for the Spanish platform first and then use the API to extract data from the European one. But that was switched since extracting information first with the API will be useful to see the shape and how the data is structured so that it is easier to create the crawler for the Spanish later.

2. STATE OF THE ART

2.1 Citizen Science

2.1.1) CS definition and history

"Citizen Science refers to the general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources" [4]. The "European Citizen Science Association" [14] defined the ten principles of citizen science in one of its publications [15].

In citizen science, groups of volunteers work with professional scientists during the research process and perform research on their own, exerting a high degree of control over research. This important but understudied field frequently focuses on addressing citizens' preoccupations [1].

Even though the term citizen science is not known by everyone yet (the European community is working on its expansion [16]), this idea of citizens contributing to scientific research is not new at all. In fact, in the 1800s century, one member of the "American Ornithologists' Union" [17] called Wells Cooke started a program that can be considered one of the first citizen science projects. It consisted of looking at the patterns of bird migration. A big number of volunteers collected data about migratory bird patterns and population figures, recording the information on cards [2].

Historically, when professional researchers needed to collect a lot of information, they contacted groups of amateur people related to their research topic. However, projects have not always started by researchers, also citizens moved by curiosity organise themselves to do science [2]. After the invention of the Internet, more and more citizen science programs have appeared since it is much easier to contribute and share information. A good example is "FoldIt" [18], an online puzzle video game about protein folding, which is part of an experimental research project that analyses how humans approach the puzzles to solve the problem of protein structure prediction. Also, technology propelled the creation of citizen science websites and platforms about projects, being "Zooniverse" [19] the world's largest CS platform launched in 2009 [5].

In fact, smartphones have helped even more to a wider expansion of citizen science, allowing everyone to record and send meaningful data. Nowadays, with the GPS included in the phones participants can provide the location of species in real-time, and even more functionalities depending on each phone's capacities [2]. For example, in "Project Noah" [20] citizens can upload photos of organisms and identify species locations through the

[14] ECSA website (2021, February 1) https://ecsa.citizen-science.net/ .

[15] The mentioned document can be found in https://ecsa.citizen-science.net/wp-content/uploads/2021/05/ECSA_Ten_Principles_of_CS_English.pdf (2021, February 1).

[16] More information can be found in https://ec.europa.eu/digital-single-market/en/citizen-science (2021, February 3).

[17] American Ornithologists' Union website (2021, February 3) https://americanornithology.org/ .

[18] FoldIt website (2021, February 7) https://fold.it/ .

[19] Zooniverse website (2021, February 7) https://www.zooniverse.org/about .

[20] Project Noah website (2021, February 7) https://www.projectnoah.org/ .

GPS of their mobile devices. "EDDMaps" [21] also uses the phone's GPS to locate invasive plants and track the spread of these species.

This advance in technology represents a high thrust to citizen science, and that is why more and more networks, platforms and communities are being created to learn more about the world and how everyone can help to understand it better [2].

2.1.2) CS projects

Following the 10 principles of Citizen Science [22], CS projects must involve citizens in scientific processes that generate new knowledge, have an authentic science outcome, receive feedback from it and consider legal and ethical aspects [6]. They can be applied to many different scientific disciplines such as natural sciences, social sciences, or humanities [23]. Depending on the field of study, the understanding of citizen science can vary. However, these projects always involve methodologies, theories and techniques that in this context are developed by specialists and a contribution of the community.

Even though projects usually belong to scientific areas like ecological, meteorological and astronomical research, there exist initiatives of all types [24] [7]. Evidently, the way of working on a technological project will be different from the methodology of one related to arts and humanities or one investigating human health. Therefore, the activities of volunteers will differ depending on the methods of gathering data and interpreting it, as stated in the document "Volunteer participation in citizen science projects" [8]. "Identifying a research question; collecting or analysing data to support or refute a hypothesis; monitoring environmental or health conditions for management or policy outcomes; and the creation of generic data within a domain to support a wide range of research questions" [9].

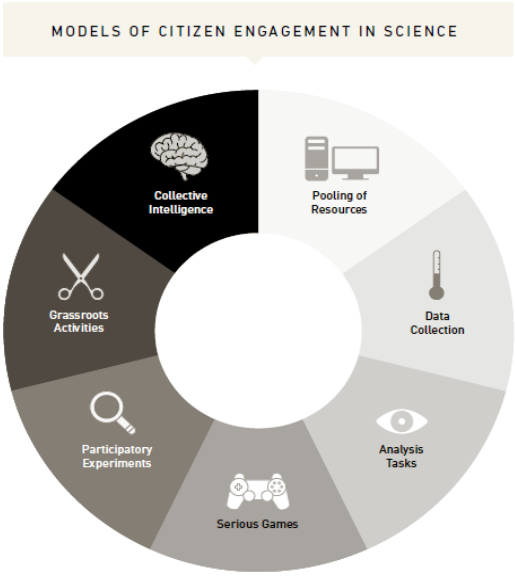


Figure 2. Models of citizen engagement in science [4].

Nevertheless, this is one of the key aspects of citizen science, that it can be used for many purposes and the citizens involved will always learn and gain knowledge, in one way or

[21] EDDMaps website (2021, February 7) https://www.eddmaps.org/ .

[22] The mentioned document can be found in https://ecsa.citizen-science.net/wp-content/uploads/2021/05/ECSA_Ten_Principles_of_CS_English.pdf (2021, February 9).

[23] The article "Citizen science in the social sciences and humanities: the power of interdisciplinarity": https://www.nature.com/articles/s41599-020-0471-y.pdf studies the impact of citizen science in social sciences and humanities (2021, February 9).

[24] The chapter "The European citizen science landscape – a snapshot" [7] verifies that there is a clear focus on projects within the life sciences (76 per cent) including ecology, environmental sciences and biology.

another [10]. The different models of citizen engagement in science are shown *in Figure 2*, and several formats and impacts of citizen science in *Figure 3*.

Citizen science projects can be done in a collaborative way either led by the community or by researchers. The degree of engagement of volunteers will also describe the project, depending on which phase the participation takes place and its conditions [8]. However, it must be taken into account that giving financial support to a project is not considered citizen science, since the citizens are not participating in any phase of the research [11].

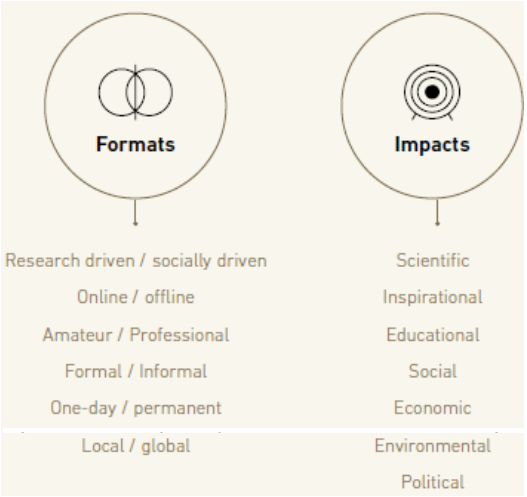


Figure 3. Citizen science formats and impacts [4].

2.1.3) CS organisations

Since citizen science is attracting more and more attention every year, several associations are being created to help in the development and management of CS. For example, the "Program Horizon 2020" is a Framework Program for financing R&D&I in the European Union corresponding to the period 2014-2020. The next (9th) PM is called "Horizon Europe", which includes the long-term budget for the years 2021-2027 with the aim of keeping the European Union at the forefront of global research and innovation [12].

Moreover, in order to boost the increasing citizen science communities, organisations such as the "Citizen Science Association" (CSA[25]), the "European Citizen Science Association" (ECSA[26]) and the "Australian Citizen Science Association" (ACSA[27]) have born [13]. The foundation of these associations converts citizen science in a field, and not just a type of coordinated commitment with science. They have been formed to share knowledge, opportunities, practices and challenges, driving citizen science as a research method and resource for education [14].

Therefore, citizen science organizations help to maximize the impact of this science, since they are key institutions that assist CS to solve global problems. By providing the institutional spine, these local or global institutions provide research agendas and infinity of opportunities to stimulate the growth of citizen science [13].

In Europe, there exist several organisations that can be governmental, non-governmental, academic, from the private sector, community-led or consortiums. In section 3 these organized bodies will be studied, as well as the ones existing in Spain.

[25] Citizen Science Association website (2021, February 12) https://citizenscience.org/.
[26] European Citizen Science Association website (2021, February 12) https://ecsa.citizen-science.net/.
[27] Australian Citizen Science Association website (2021, February 12) https://citizenscience.org.au/ .

2.1.3.1) CS organisations in Europe

In Europe, there is an old tradition of citizens participating in research, especially in the areas of birding, history, and astronomy [28] [15]. In order to give citizens a bigger role in science "science for and by the people", Europe funded the "European Citizen Science Association" [29] (ECSA) in 2013 [13]. It is a non-profit association that pretends "to connect citizens and science, to promote sustainable development through citizen science and to ensure that citizen science contributes to policy processes" [16]. It gathers researchers, professionals and civilians for the purpose of promoting citizen science in Europe and promoting the participation of society in investigation processes, through science, social science, humanities and arts [9].

Since Europe covers many different countries, there exist socio-geographical and cultural differences, which are also reflected in the citizen science state. The paper "The Science of Citizen Science: Theories, Methodologies and Platforms" [17] analysed the citizen science landscape in Europe and observed that there is a link between democracy and public participation in research, seeing that regions with bigger democracy indices have more engagement of citizens in science.

2.1.4) CS resources developed by CS projects

With the objective of getting everyone (individuals, families, schools, communities…) engaged in citizen science, numerous resources have been created. They can be of many types (educational resources, videos, guides, reports, books, interactive resources…) and are available on multiple websites, but especially on projects websites and CS platforms, which will be differentiated in section 2.2.1.

2.1.5) Citizen science as an educational tool

One of the most positive consequences of citizen science is the fact that "Citizen science offers the potential to increase student engagement through active and inquiry-based learning" [18]. However, for a project to belong to citizen science, educational outcomes are not the main goal. Thus, the project must have a scientific aim, but in the way to achieve this goal other multiple benefits as increasing volunteers' knowledge will take place.

Citizen science has the capacity to "develop connections between students' everyday lives and science so that they will have tangible reasons for continuing with the lifelong learning of science" [19]. Surveys of more than 1500 students participating in the citizen science program "ClimateWatch" [30] showed that their environmental engagement increased significantly after participating in data collection and data analysis [18].

As Shah and Martinez (2015) concluded in "Current Approaches in Implementing Citizen Science in the Classroom" [20], implementing citizen science projects in schools provides several benefits to students. Community awareness, critical thinking, and problem-solving added to practical experience are increased when learning through this recent

[28] For example "Galaxy Zoo" https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/ and "The Big Butterfly Count" https://bigbutterflycount.butterfly-conservation.org/ (2021, February 17).
[29] European Citizen Science Association website (2021, February 17) https://ecsa.citizen-science.net/ .
[30] ClimateWatch project website (2021, February 19) https://www.climatewatchdata.org/.

model of education. "Inquiry-based science education through citizen-science projects should be designed to nurture independence, persistence, creativity, and determination in the minds of young students to prepare them to solve community problems with a systematic approach" [20]. Independently of the method used or discipline explored, hands-on experience is a very important aspect to accentuate when exposing students to research.

2.2 Resources

2.2.1) CS platforms vs projects websites

> "Citizen science platforms are web-based infrastructures with one single entrance point that contain one or several of the following functionalities: (1) present active citizen science projects and activities; (2) display citizen science data and information; (3) provide overall guidelines and tools that can be used to support citizen science projects and activities in general; (4) present good practice examples and lessons learned; and (5) offer relevant scientific outcomes for people who are involved or interested in citizen science." [10]

Nowadays there are several citizen science platforms in Europe, since CS is an emerging field. According to "The Science of Citizen Science" book [10], they can be distinguished into five groups:

1. Commercial platforms for citizen science initiatives

Commercial platforms basically are created to take profit from their services, while the other kind of platforms do not gain money from their services. During the last year, several commercial platforms for citizen science, such as "SPOTTERON" [31] and "CitizenLab" [32], have been created. They mostly target project leaders like scientists and institutions, and they look for a bilateral market exchange, while the other types of platforms aim for a multilateral collaborative relationship [10].

2. Citizen science platforms for specific scientific topics

Citizen science platforms for specific scientific topics are platforms that have been created to address a particular field, like air pollution, water quality, biodiversity… They store different information and data that can be consulted by individuals, scientists and authorities. The "Luftdaten.info" [33] platform is an example, which contains information about air quality [10].

3. Citizen science platforms for specific projects

Several existing citizen science platforms have been created as a general outcome of specific projects, providing information about a particular initiative. They usually contain information on the topic, a toolbox on how to collect data, an option to upload data, visualization of the achieved data and links to similar initiatives [10]. The "Captor" [34] project is an example of a citizen science platform for a specific project.

[31] Spotteron website (2021, February 21) https://www.spotteron.net/.
[32] CitizenLab website (2021, February 21) https://www.citizenlab.co/es .
[33] Luftaden website (2021, February 21) https://sensor.community/en/ .
[34] Captor website (2021, February 21) https://www.captor-project.eu/en/ .

4. EU citizen science platforms

There exist two citizen science platforms at the European level, the "Joint Research Centre" [35] (JRC) citizen science platform (developed by the EC's JRC) and the "EU-Citizen.Science" [36], still in development. This second one was created due to the "EU Horizon 2020 Framework Programme for Research and Innovation" [37], and aims to build a central platform for CS in Europe, sharing all kinds of useful resources about citizen science. The "European Citizen Science Association" [38] (ECSA) is one partner of this project and will be the responsible member when finished [10].

5. National citizen science platforms

Multiple countries of Europe have generated citizen science platforms to show the variation of CS projects that they have going on. These platforms are usually done by citizen science networks, which aim to promote citizen science. They tend to use the national language to present the information and are targeted to everybody. These national citizen science platforms present initiatives and activities that can be interesting for the citizens, but they also offer other resources such as guides, metrics and events, which can be consulted by everyone [10]. The "Observatorio Nacional de Ciencia Ciudadana" [39] is the Spanish platform, which will be studied later on.

Therefore, the difference between CS platforms and individual project websites is that the first host a huge number of projects and other resources so that the public can discover initiatives and choose where to participate, while in the second case the sites just describe what that concrete project consists on and explains all the information regarding that program, but not others [10]. So, the platforms are repositories containing multidisciplinary aspects and resources related to citizen science whereas websites describe what a specific project does.

2.2.2) European level

"EU-Citizen.Science" [36] is an online platform to share knowledge, tools, resources and training of citizen science, done by and for the community [14]. Formed by partners from 14 European member states as well as other collaborators, they represent a variety of active agents in citizen science such as universities, non-governmental organizations, local authorities, community service organizations and museums [14].

The goal of the platform is to serve as a knowledge point that helps to popularize citizen science and increase the impact that society's participation is having in most areas of scientific research. This "EU-Citizen.Science" project is funded by the "Horizon 2020 program of the European Commission" [40], through its line of work "Science with and for Society" (SwafS). Their mission is to become a reference point for citizen science by creating a network to exchange knowledge among professional scientists, volunteer citizens, researchers, political agents and society in general at the European level [14].

[35] Joint Research Centre website (2021, February 23) https://digitalearthlab.jrc.ec.europa.eu/csp.
[36] EU-Citizen.Science European website (2021, February 23) https://eu-citizen.science/
[37] Details of its funding can be found at https://cordis.europa.eu/project/id/824580/es (2021, February 23).
[38] European Citizen Science Association website (2021, February 23) https://ecsa.citizen-science.net/ .
[39] Observatorio de la Ciencia Ciudadana en España website (2021, February 23) https://ciencia-ciudadana.es/ .
[40] Horizon 2020 program website (2021, February 25) https://ec.europa.eu/programmes/horizon2020/en/home .

The EU-funded project to develop the EU-Citizen.Science was founded in 2019 and released the first version of the platform in March 2020. It aims to be the central knowledge-sharing hub for CS in Europe by 2021, since the project ends in December 2021, after which the platform will be managed by ECSA following the same objectives [14].

It aims to be useful for any kind of public, showing distinct content such as:

- Resources that are useful for people interested in citizen science.
- Training resources and materials to put citizen science into practice.
- Projects that involve citizens in research through CS activities.
- Organisations that are involved in citizen science projects and research, which belong to a specific type.
- A Blog and Community Forums to share questions, conversations and collaborate with the rest of the community.
- An Events calendar to show the ongoing and upcoming events related to citizen science.

2.2.3) National level (Spain)

The "Ibercivis Foundation" [41] is a private non-profit organization that aims to develop citizen science projects and promote CS by making it more visible. Between 2012 and 2014 Ibercivis led a project called the "Socientize Project" [42], which was required by the European Commission to study the situation of citizen science in Europe and impulse it [12].

The "Observatorio de la Ciencia Ciudadana en España" [43] and the webpage https://ciencia-ciudadana.es/ were created in 2015 and launched in 2016, being the result of a project developed by the "Ibercivis Foundation" in collaboration with the "Fundación Española de Ciencia y Tecnología (FECYT) - Ministerio de Ciencia e Innovación" [44]. The Spanish CS platform aims to facilitate knowledge sharing about citizen science, analysing how the relationship between science and society evolves. By giving visibility to the projects, the goal is to increase participation in active citizen science projects. Anyone can share resources and results in this portal, as it is a tool for everyone [12].

Since the CS field is growing very quickly (projects, publications, resources…) there is a need to coordinate this development to make it more understandable. That is why the Spanish platform is born, being a "content repository and catalogue of practices, a platform to see and be seen, with scientific divulgation and education content" [12]. In addition, specific indicators will be implemented so that quantitative analysis can be also done. Moreover, several reports are posted by this platform summarising the situation of citizen science in Spain. The latest one reviews the activity of the "Observatorio de la Ciencia Ciudadana" and the agents involved in CS covering the period from April 2019 to November 2020 [12].

The platform has carried out the necessary visibility tasks, recognition and promotion, since its creation in 2015 and launch in 2016, with the collaboration of the FECYT [12].

[41] Ibercivis website (2021, February 27) https://ibercivis.es/ .
[42] Socientize Project website (2021, February 27) https://www.socientize.eu/ .
[43] Observatorio de la Ciencia Ciudadana en España website (2021, February 27) https://ciencia-ciudadana.es/ .
[44] (2021, February 27) https://www.ciencia.gob.es/site-web/;jsessionid=8ECA1241D0A7A8F59BF39BFE721D8430.

2.2.4) City level (Barcelona)

The Barcelona City Council launched a program to foster scientific knowledge and involve citizens in its progress. This program is called “Barcelona Ciència” [45], and wants that science and society get mixed to advance together, so that science is able to respond and find solutions to the needs of society [21]. Therefore, citizens are starting to play roles in scientific projects, leading to the boom of citizen science. In this context, the program created a webpage called "Oficina de la Ciència Ciutadana" [46] that acts as a multidisciplinary platform, including information on the different projects that are being carried out in the city, the ongoing activities related to citizen science, news regarding the topic of citizen science and a calendar of CS events [21].

The CS Office was created in 2012 to consolidate citizen science projects in Barcelona and help them in facing the social and environmental challenges of the city. Between 2012 and 2019, the Office has prepared different events to promote citizen science among the city’s residents, such as the Citizen Science Day [21]. Moreover, many research groups have shared their projects with the public. During these years different actions and programmes have brought citizen science closer to numerous districts, with the participation of 13.000 residents and schoolchildren, who have collected more than 10.000 valuable data for scientific research [21].

Moreover, the “Oficina de Ciència Ciutadana” and the “Consorci d'Educació de Barcelona” organize the “Ciència Ciutadana a les Escoles” [47] program to introduce citizen science in schools. Teachers and students from different schools in Barcelona participate in this 15-hour pilot program to collaborate with scientific research projects of the city [21]. The educational community is in contact with scientists to learn about the scientific method (sampling, collecting data with smartphones and applications, analysing the results…) and also school hours are dedicated so that students enlarge this knowledge [21].

2.3 Metadata

Citizen science (CS) and public participation in scientific research (PPSR) have increased a lot during the last few years. In 2013 the organized bodies related to citizen science created the PPSR_CORE Program Data Model Metadata Standard [22]. It is a sharing protocol to standardize the way citizen science projects are informed and share information among databases to maximize the utility of this data. Therefore, the information will be synchronized and updated so that citizen science is closer for everyone [23].

This PPSR_CORE initiative was founded by four organizations (SciStarter.org[48], Cornell Lab of Ornithology[49], The Wilson Center - Commons Lab[50], CitSci.org[51]) that are now

[45] Barcelona Ciència website (2021, March 9) https://www.barcelona.cat/barcelonaciencia/es
[46] Oficina de la Ciència Ciutadana website (2021, March 9) https://www.barcelona.cat/barcelonaciencia/es/ciencia-ciudadana.
[47] Ciència ciutadana a les escoles program website (2021, March 9) https://www.barcelona.cat/barcelonaciencia/ca/ciencia-ciutadana/ciencia-ciutadana-escoles .
[48] SciStarter website (2021, March 11) https://scistarter.org/ .
[49] Cornell Lab website (2021, March 11) https://www.birds.cornell.edu/home/.
[50] Wilson Center website (2021, March 11) https://www.wilsoncenter.org/publication-series/commons-lab.
[51] CitSci website (2021, March 11) https://citsci.org/.

working inside the Citizen Science Association to continue driving the use of citizen science [23].

Due to the fact that "EU-Citizen.Science" acts as an information centre to share data regarding CS (projects, experiences, tools, resources, training materials, resources…), it implemented an information structure that follows the PPSR Common Conceptual Model. This model is formed by the Project Metadata Model, the Dataset Metadata Model and the Observation Data Model, which were created in special for the citizen science initiatives [22]. By applying this metadata architecture, it will be easier to find the desired content, compare the information with other platforms and cover the user's needs.

The three important classes of content for "EU-Citizen.Science" are CS resources, CS training materials and CS projects. Its metadata models basically consist of attributes such as resources titles and authors, keywords that describe the content, projects' descriptions and status, research topics, audience and intended outcome… [14]. The creation of these ontologies is very helpful to organize the information in a way that makes it easier to find anything and understand all the content [14].

2.4 How to extract data from web pages

The World Wide Web contains information of all types. An enormous number of users use the internet every day, generating a very big of data. But information on the web is available in different formats and through different access interfaces [24], so the issue of how to handle such data overload and how the user will access the best information in the least efforts has appeared [25]. Web scraping is the technique that aims to address this problem. "Web scraping is widely acknowledged as an efficient and powerful technique for collecting big data" [26].

On one hand, an Application Programming Interface (API) gives access to the data of an application, website, operating system or other services through a set of procedures and communication protocols (See section 3.1). Usually, APIs belong to the owner of the dataset that is being used, who can provide it for free or charge a fee [27]. In this case, the European platform has generated its own API[52]. Therefore, it will be easier to extract data from this website.

On the other hand, with web scraping a user can extract data from any website, so it does not need to have an API. Web scraping is the technique or method that allows extracting data from websites, and then a crawler algorithm that classifies and stores the data in databases is created. By classifying the data, it is possible to convert unstructured data into structured. In this case, the European platform has semi-structured data and the Spanish one structured data. Thus, they classify the data in a database with the purpose of performing searches and analysis [22].

"The most common techniques for data scraping are manual copy-and-paste, text grabbing and regular expression matching, HTTP programming, HTML parsing, DOM parsing, Web scraping software, Vertical aggregation platforms, Semantic annotation recognizing and Computer vision webpage analysers" [25].

[52] Link to the EUCS PLATFORM API (2021, March 11) https://eu-citizen.science/swagger/ .

Traditional copy and paste is the most common technique for basic-skilled internet users. However, this method is tedious when a lot of data is needed. In comparison, web scraping software is a much easier scraping technique. Nowadays there are several readily available software, mostly designed by using Java, Python and Rubi (See section 3 for more details about technology used). These tools can automatically recognize the data structure of a page and provide interfaces and scripting functions so that the user does not need to manually write code when extracting and transforming content [25].

Therefore, web scraping tools can automate the manual work by visiting each page and extracting data from pages or parsing the HTML pages. There exist several web scraping software such as Import.io[53], a free online web scraper that allows anyone to extract information from a website without writing code; Mozenda[54], a Business Intelligence Software which allows users to extract data from documents as well as web pages or different sources; Scrapy[55], a free open-source python library designed to scrape web content from sites that are composed of many pages of similar semantic structure; Visual Web Ripper[56], one of the most advanced web scraping software that provides functionalities to scrape data from any website; and Web Content Extractor[57] (WCE), a simple user-oriented tool in which users can scrape data from a website with a few clicks and put the data into different formats [25].

2.4.1) Robots.txt

Before starting the extraction, the rights to use APIs or web scraping must be discussed. Some web pages have a robots.txt [58] file in which there is information about what content can be accessed. In more technical words, robots.txt file is a text file that webmasters make to indicate web robots (usually search engine robots) how to crawl the website. It uses the Robots Exclusion Protocol (ERP), a protocol that indicates how robots access and index content. By allowing and disallowing behaviours, the robots.txt file determines if user agents (which are web-crawling software) are allowed to crawl parts of a website [28].

The term known as “spidering” refers to the crawling method in which search engines follow links to go from one site to another. When the web crawler goes to a website it looks for the robots.txt file before spidering it [28]. If found, it will follow the policy and proceed crawling the information that is permitted. Robots.txt files are not obligatory for web pages, but recommended if some content wants to be blocked from robots [28]. To know if a web page has this robots.txt file, it just needs to be added “/robots.txt” to the corresponding URL. So, the robots.txt file will be checked in order not to violate any

[53] Import.io website (2021, March 13) https://www.import.io/ .
[54] Mozenda website (2021, March 13) https://www.mozenda.com/ .
[55] Scrapy website (2021, March 13) https://scrapy.org/ .
[56] Visual Web Ripper website (2021, March 13) http://visualwebripper.com/ .
[57] Web Content Extractor website (2021, March 13) https://www.newprosoft.com/ .
[58] Robots.txt website (2021, March 15) http://www.robotstxt.org/robotstxt.html .

policy of the website that will be scrapped. For the Spanish citizen science platform, there is no restriction:

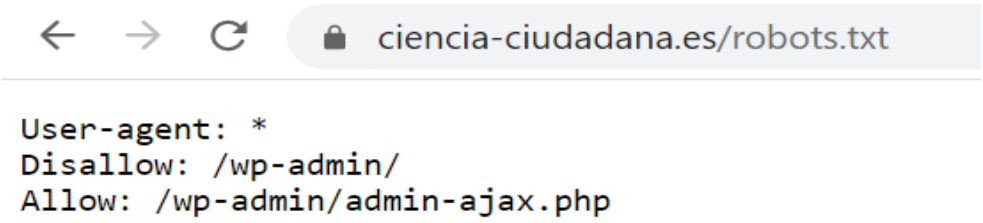


Figure 4. Robots.txt file of the Spanish platform.

For the Barcelonian website, the robots file is the following. As seen, it does not have any restrictions for the content that will be scraped either.

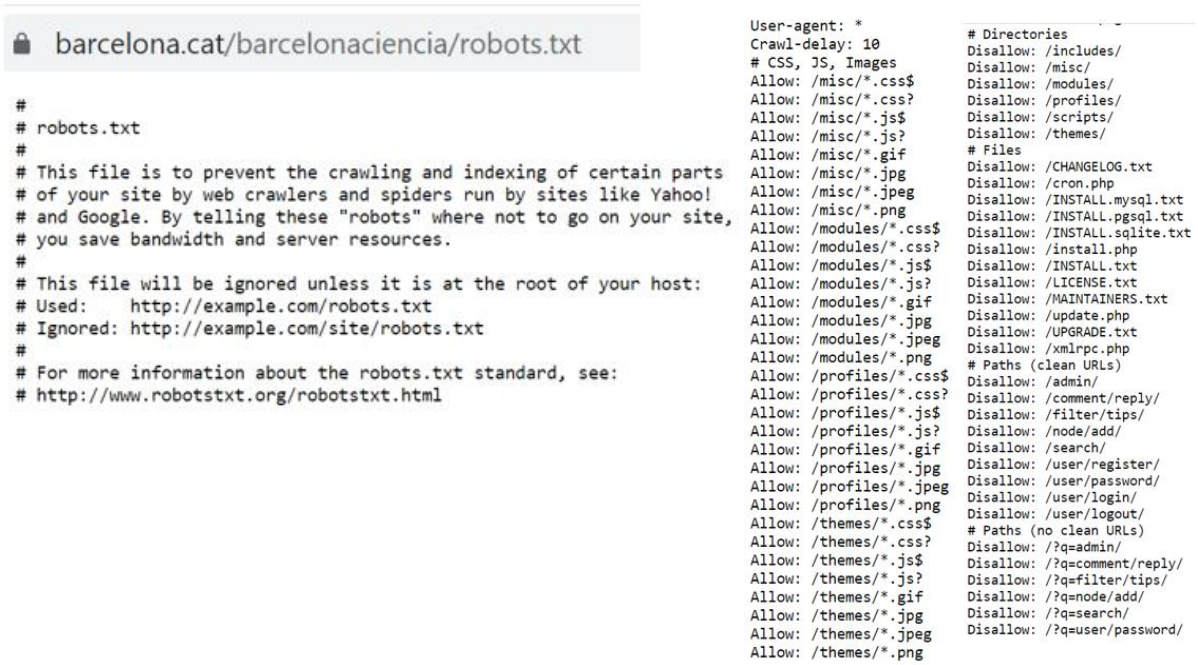


Figure 5. Robots.txt file of the Barcelonian platform.

3. DESIGN AND IMPLEMENTATION

3.1 - Environment set-up, study and selection of the tools

3.1.1) Tools used

In order to perform the analysis of the three platforms, different tools will be used during the whole thesis process. They are described below to understand why each one is necessary and how they have been applied to the project.

3.1.1.1) APIs

An Application Programming Interface (API) is a set of definitions and protocols aimed to develop and integrate application software, allowing communication between two or more software components through a set of rules [27]. These rules determine the method that is used to access the service, plus which and how data is returned. APIs are useful to help non-expert (non-developer) users to interact with the applications and get information, since they are a layer between the application and an external service. The main advantage of APIs is that it is not needed to know how the internal structure of the application works, since just with one command you can get the desired data in a specific format [27].

The most common APIs for websites are Representational State Transfer (REST), which uses HTTP requests and responses to communicate with web services, sending, receiving, deleting or modifying data. They can be seen as an Internet address to access data from a website through a set of protocols and tools, which in Python correspond to libraries [27].

In order to collect data from an API, a request must be done. There are four types of request methods, which indicate what action must be done: GET (to get data from the API), POST (to add new data to the server), PUT (to change already existing data), DELETE (to delete data). The request returns a response, which includes the status code (200 means everything went well) and the data [29].

For this project, only the get method will be called since the objective is to get data from the European API. In this case, the requests will be done through the command line, though the interaction can also be done through scripts.

3.1.1.2) Python

Python[59] is a high-level programming language that makes it easy to work with data, has an open-source license and is free to use. It is object-oriented, has built-in data structures and has a very extensive library of modules and packages that facilitates performing data analysis [30]. Therefore, it has been the chosen language to study the platform's information in order to generate insights.

Natural Language Toolkit (NLTK[60]) is a Python package for natural language processing. A lot of the data to analyse is unstructured and contains human-readable text. Before analysing it programmatically, it needs to be preprocessed. “NLTK provides resources

[59] Python website (2021, March 23) https://www.python.org/ .
[60] NLTK website (2021, March 23) https://www.nltk.org/.

for preprocessing tasks as well as text analysis, such as libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning" [31].

Other Python libraries that have been used in the scripts are:

- Numpy[61]: a package for scientific computing in Python. "It is a Python library that provides a multidimensional array object, various derived objects and an assortment of routines for fast operations" [32].
- Pandas[62]: a fast and powerful open-source data analysis and manipulation tool. It offers features for reading and writing data, creating DataFrame objects to manipulate data with integrated indexing, and much more functionalities to work with data sets [33].
- Matplotlib[63]: "A comprehensive library for creating static, animated, and interactive visualizations in Python" [34].
- Seaborn[64]: "A Python data visualization library based on matplotlib, which provides a high-level interface for drawing attractive and informative statistical graphics" [35].
- Datetime[65]: "The datetime module supplies classes for manipulating dates and times" [36].
- Locale[66]: to change the local time for the Spanish projects.
- re[67]: "This module provides regular expression matching operations" [36].
- Operator[68]: "The operator module exports a set of efficient functions corresponding to the intrinsic operators of Python" [36].
- Math[69]: "This module provides access to mathematical functions" [36].
- String[70]: library for common string operations.
- Unidecode[71]: it provides a function that takes Unicode data and tries to represent it in ASCII characters [37].
- Category_encoders[72]: "A set of transformers for encoding categorical variables into numeric by means of different techniques" [37]. Used in "analysing_european.ipynb" to encode variables.

3.1.1.3) Jupyter Notebook

"The Jupyter Notebook[73] is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text" [38]. It supports several programming languages (including Python) and can be used to clean and transform data, visualize data, perform machine learning… Since it has been the most used tool for programming during the whole degree, it is the chosen application to build and run the necessary Python scripts.

[61] Numpy website (2021, March 23) https://numpy.org/.
[62] Pandas website (2021, March 23) https://pandas.pydata.org/.
[63] Matplotlib website (2021, March 23) https://matplotlib.org/.
[64] Seaborn website (2021, March 23) https://seaborn.pydata.org/.
[65] Datetime documentation (2021, March 23) https://docs.python.org/3/library/datetime.html.
[66] Locale documentation (2021, March 23) https://docs.python.org/es/3.9/library/locale.html.
[67] Re documentation (2021, March 23) https://docs.python.org/3/library/re.html.
[68] Operator documentation (2021, March 23) https://docs.python.org/3/library/operator.html.
[69] Math documentation (2021, March 23) https://docs.python.org/3/library/math.html.
[70] String documentation (2021, March 23) https://docs.python.org/3/library/string.html.
[71] Unidecode documentation (2021, March 23) https://pypi.org/project/Unidecode/.
[72] Category encoders documentation (2021, March 23) https://pypi.org/project/category-encoders/.
[73] Jupyter Notebook website (2021, March 23) https://jupyter.org/.

3.1.1.4) JSON

The language of APIs is JavaScript Object Notation, most known as JSON[74]. It is one format to encode the data structured in a way that makes it understandable for processing. Often the REST API returns the response in this JSON format, which usually contains combinations of dictionaries, lists, strings and integers. To manage this type of data Python has a package called json that is included in the standard library and makes it easy to convert data. Moreover, the requests library also has a .json() method that decodes and transforms JSON to a Python object [39].

3.1.1.5) Selenium

As explained before, web scraping is a technique to obtain data automatically from web pages. Selenium[75] is an open-source web testing automation tool that supports the most used browsers and can be programmed in several languages such as Python, Java or C# [40]. Python has a Selenium library that allows you to instantiate a web browser and perform the web scraping. It will be used to extract data from the Spanish and Barcelonian platforms, by creating a crawler with python and the necessary libraries.

3.1.1.6) Tableau

Tableau[76] is a powerful visual analytics platform that helps to produce visual insights from data. It is a very useful data visualization tool to make data more understandable by creating interactive dashboards and visualizations [41]. It has been presented to UPF data science students this fourth year in a first-trimester subject, which makes it the perfect choice for this project since using it will strengthen the knowledge of it while generating really attractive outputs from the CS platforms data.

3.1.1.7) Wix

Wix[77] is a powerful platform to develop web pages that allows anyone to create a professional website [42]. It has several powerful features such as embedding HTML code which allows displaying Tableau dashboards on the website by copying and pasting the visualization sharing code. Therefore, it has been used to generate a website where the results will be presented so that anyone interested can get insights from the project conclusions.

3.1.1.8) GitHub

"GitHub[78] is a code hosting platform for version control and collaboration" which permits sharing files uploaded to repositories [43]. For this project, a repository[79] will be created in order to post all the generated scripts with the corresponding code and outputs so that anyone interested can consult them.

3.1.2) Generated code and dashboards

In order to get and analyse the content of the platforms, six scripts have been generated. First of all, one called "scraping_data.ipynb" scraps the desired content from the Spanish

[74] JSON website (2021, March 23) https://www.json.org/json-es.html.
[75] Selenium website (2021, March 25) https://www.selenium.dev/.
[76] Tableau website (2021, March 25) https://www.tableau.com/es-es.
[77] Wix website (2021, March 25) https://www.wix.com/about/us.
[78] Github website (2021, March 25) https://guides.github.com/.
[79] Github repository link: https://github.com/nvarasp/TFG_CS_Analysis

and Barcelonian platforms using Selenium and saves the different files. The way to do so is explained later on.

The three scripts called "European_anlaysis.ipynb", "Spanish_analysis.ipynb" and "Barcelonian_analysis.ipynb" analyse individually the content of the corresponding platform in a more extensive way, generating interesting plots to get insights. The "functions_t.ipynb" script includes several created functions that will be used in the analysing scripts.

Finally, the one called "Comparing_platforms.ipynb" imports the project's and resources' data frames from the three platforms (generated as .xlsx files after performing some analysis in the corresponding scripts). On the one hand, it checks if there are projects that appear on more than one website and displays the corresponding cases to compare the different ways in which platforms present the same initiative. On the other hand, it joins all projects from the three pages (taking the repeated ones just one time) to generate a global dashboard with all projects, and the same is done with the resources. This will assist to create a general overview of the citizen science current situation.

These Jupyter notebook scripts will be uploaded into Github[80] so that anyone interested in citizen science can consult them (see Appendix B to consult the locations of the scripts). And the resulting data frames will be imported to Tableau as a means to generate dashboards visualizations that display the data in a more visual and interactive way. All of them will appear in the Tableau Public profile[81] (see Appendix B to consult the locations of the dashboards), and a created website will include the most interesting ones together with some obtained conclusions so that anyone can get all the insights with one click.

3.1.3) Performed analysis

In order to analyse the content of the three platforms, some different analyses have been performed, most of them related to text data mining. By installing and importing the Python Natural Language Toolkit (NLTK[82]) library, several text processing libraries for parsing, stemming, tokenization, tagging and semantic reasoning can be used. This platform is used to create Python scripts that cope with human language data and makes it easier to apply statistical natural language processing. The most important used analysing methods are explained below:

- Tokenizing: First of all, any text can be cleaned by applying some re substitution functions, where re[83] is a module that provides regular expressions matching. Then, the nltk.tokenize[84] package allows dividing a string into its substrings to get the separated terms that compose a text (also called tokenize). The next step in this pre-processing is to remove stop words[85], which are the most common words of a language, like articles and prepositions. It is interesting to remove them since they do not provide any useful information, and the way to do so is by importing the stop words

[80] Github repository link: https://github.com/nvarasp/TFG_CS_Analysis.
[81] Tableau profile: https://public.tableau.com/profile/n.ria4479#!/.
[82] NLTK website (2021, April 10) https://www.nltk.org/.
[83] Re website (2021, April 10) https://docs.python.org/3/library/re.html.
[84] NLTK tokenize documentation (2021, April 10) https://www.nltk.org/api/nltk.tokenize.html.
[85] Stop words explanation (2021, April 10) https://www.geeksforgeeks.org/removing-stop-words-nltk-python/.

from the nltk.corpus[86] module (specifying the corresponding language) and keeping only the terms that do not appear in that stop words list.

- Stemming: Then, the resulting words are reduced to their common base form, stem or lemma, which can be done with stemming or lemmatization. Both of them return the main basic form of a word, but stemming usually cuts the last part removing derivational affixes while lemmatization takes into account the context to return the term base form [44]. In this case, stemming[87] will be the chosen method since both of them have been tried and this has shown to work best. Still, there exist more than one stemming algorithm, being PorterStemmer[88] and Snowball Stemmer[89] the most typical ones. Since Snowball Stemmer is a better version of Porter Stemmer that tries to fix some issues of that one, it is the algorithm that will be used [44]. Once all this pre-processing is done (the resulting text after Snowball Stemmer is applied will be considered stemmed), the most common words of a text can be seen by counting the number of times each one appears.

- TF-IDF: Another text analysis technique is TF-IDF, which is a numerical measure used to get the importance of a word for a document in a collection. The Term Frequency - Inverse Document Frequency technique is based on two elements. First, the term frequency (tf) counts the number of occurrences for a given word in a document, like a projects' description [45]. Then, the inverse document frequency (idf) measures if a term is common or not in a collection of documents. It is obtained by dividing the total number of documents (for example projects) by the number of documents (or projects) that contain that term, and the logarithm of this division is taken as the result. Therefore, the tf-idf value increases with the number of times a word appears in a document but at the same time, it decreases with the repetitions of the word in the whole collection [45]. With this method, the most relevant words for a document can be computed. (It must be taken into account that to perform this analysis the text used will be already stemmed, since otherwise the counting of words would not be fair, interpreting science and sciences as two different terms, for example).

- Cleaning names: Some projects, resources or organisations often have very large names, including a sentence with a short description. Therefore, when searching if another element contains a specific name, most times it would say no since the whole name would not be included, despite it may be mentioned with its concise name. That is the reason behind this technique, that cleans the names by first checking if it can be split with a punctuation mark, and in case this split generates more than one term then getting only the first part. There can also be cases in which short names are inside brackets and contain a dash. An example to cope with those cases is:

[86] NLTK corpus documentation (2021, April 10) https://www.nltk.org/api/nltk.corpus.html
[87] More information about stemming: https://www.geeksforgeeks.org/introduction-to-stemming/
[88] Porter Stemmer website (2021, April 12) https://tartarus.org/martin/PorterStemmer/.
[89] Snowball Stemmer website (2021, April 12) https://snowballstem.org/ .

```
if (len(row["name"].split('('))>1):
        if(len(row["name"].split('(')[0].split('-'))>1):
            SP_initiatives.loc[index,
"clean_name"]=remove_text_marks(row["name"].split('(')[0].split('-')[0])
        else:
            SP_initiatives.loc[index,
"clean_name"]=remove_text_marks(row["name"].split('(')[0])
```

Figure 6. Code example to clean names.

- Variables to data frames: The way to proceed to analyse the categorical or list variables is first of all, working with the data format, since in most cases they were dictionaries that included more information than needed, and generating data frames with one row for each possible value. Then, a data frame is generated with the necessary information for that field, making it easier to count the number of times an element is mentioned, which will show the relevance of a topic, equipment type, audience class... Depending on the project field the pandas data frame[90] included more or fewer columns, being the basic ones the name of the entry, an identifier for it, the number of projects in which it appears (computed with some lines of code) and a list with the ids of the projects.

- Classifying: The way to classify projects or resources using text fields is by generating a list of words that refer to a topic, for example, and looking if that project description contains any of those words. If it does, that value is set to true and otherwise to false. This is also one method used to classify initiatives into educational or not, looking for specific words in the text.

- Correlations: In order to extract correlations among variables such as projects topics and status, those variables need to be converted into numerical values. Therefore, label encoders must be used to determine which number corresponds to a specific type. Once all variables are numerical or Boolean, correlations can be computed.

- More techniques: Finally, if the amount of data was bigger, unsupervised learning algorithms could be applied to describe how projects or resources are grouped in different styles. Therefore, with a clustering technique data would be segmented into groups of similar characteristics, leading to different projects and resources styles. However, in this case, data is not big enough so it does not make sense to cluster it since the results would not be fair.

3.1.4) Functions script

The “functions_t.ipynb” script contains several created functions that are used in the analysing scripts, so that they do not need to be included in each of them and can be reused. To import the functions from this file, the others just need to install and import “nbimporter”[91], which is a python module to import functions from other Python

[90] Pandas DataFrame documeantation (2021, April 14) https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html

[91] Nbimporter documentation (2021, April 14) https://pypi.org/project/nbimporter/.

notebooks. Therefore, by including the following code any script can already called the defined functions.

The created functions are slightly described now:

1. jprint: Function to print json objects in a more visual way.
2. clean_text: this function applies the following functions to clean the text (each one does what its name describes, receiving as parameter a text and returning it with the corresponding process done):
    a. text = text_to_lower_case(text)
    b. text = remove_text_marks(text)
    c. text = remove_punctuation_marks(text)
    d. text = remove_accents(text)
    e. text = remove_emojis(text)
    f. text = split_text_and_numbers(text)
    g. text = remove_alone_numbers(text)
    h. text = remove_multiple_whitespaces(text)
3. convert_to_nltk_text: return the tokens list of the text coming from the corresponding variable.
4. standardize_text: this function receives as inputs the text and its language, and returns the text standardised (by removing stop words and stemming the terms).
    a. remove_stopwords: gets the stop words (explained in 3.1.3) for the given language and removes them from the text.
    b. stem_text: creates a stemmer and stems (explained in 3.1.3) each token in the text, returning the stemmed text.
5. plot_text_length_distribution: computes the vocabulary size for the given text field and plots the distribution of the text length
6. get_bagofwords: creates a bag-of-words with the frequency of each word in the corpus, it creates a stemmed one and another one with the original words.
7. plot_wordcloud and plot_barplot: generate plots of the bag of words with the corresponding words and frequencies.
8. class TF-IDF (Term Frequency–Inverse Document Frequency): to show how relevant a specific word is in a set of documents. Term Frequency indicates the number of times the given word appears in the collection of documents, and Inverse Document Frequency reflects how much information the word provides.
    a. tokenize: tokenize raw text into words and then remove stop words
    b. load_df: turn list of words into pandas dataframe
    c. tf: calculates term frequency
    d. idf: calculate inverse document frequency
    e. count_idf: helper for idf
    f. tf_idf: calculates tf-idf together
9. plot_countplot, plot_countplot3, plot_by_variable, pie_chart (including func) and plot_pie: different functions to create plots.
10. Get_people: a function that first of all, recognises the entity type of the tokens (using the nltk “word_tokenize” and “pos_tag” functions) and then looks for the sequences that have two consecutive proper names, meaning that the expression refers to a person. That is done creating the corresponding chunk pattern ({<NNP>+<NNP>+}) and looking for it with the “RegexpParser” and “parse” nltk functions [46]. This function is used to identify the people's names mentioned in a text.

3.2 - European platform: The “EU-CITIZEN SCIENCE” [92]

3.2.1) Platform data structure

The European platform search section is segmented into 4 parts: projects, resources, training resources and organisations. Also, there is an outstanding section with the elements that they mostly recommend. This project will focus on analysing these four types of content, since the agenda and blogs are not that interesting for the work purpose. The **projects** can be in different states, belong to several topics and countries, and be possible to do at home or not.

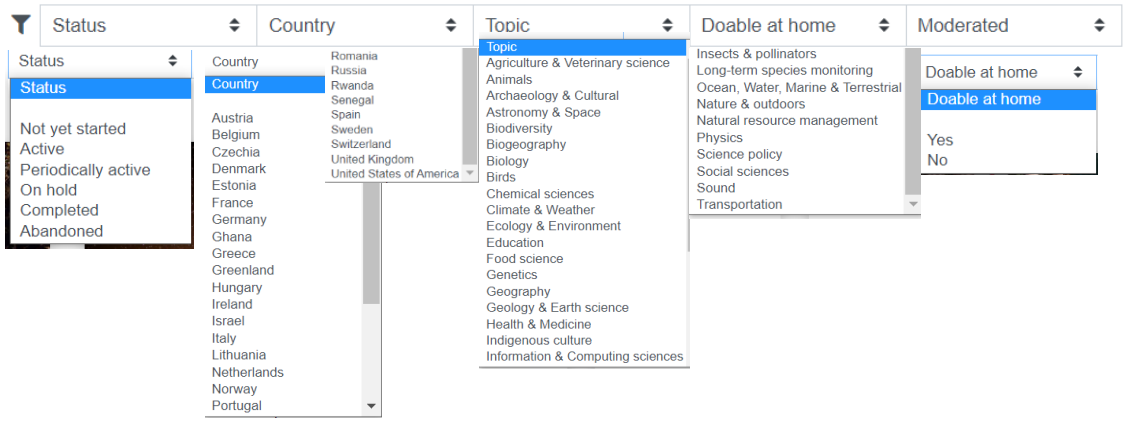


Figure 7.European projects' filters.

When clicking in a project, there appear several fields:

-PAGE 1: DESCRIPTION: Name, aim, description, reviews and URL.

-PAGE 2: PARTICIPATION: How to participate, needed equipment and participation task.

-ON BOTH PAGES:

- Starting and ending date, plus status (active/completed…).
- Science topic (education, physics, social sciences,...).
- Keywords (online games…).
- Location: tags like global, Dinamarca, online globally.
- Contact (host and main organisation).

Both the resources and training resources can be in 11 different languages, belong to distinct 19 themes and belong to one of the 20 resource types.

[92] European platform website (2021, April 21) https://eu-citizen.science/.

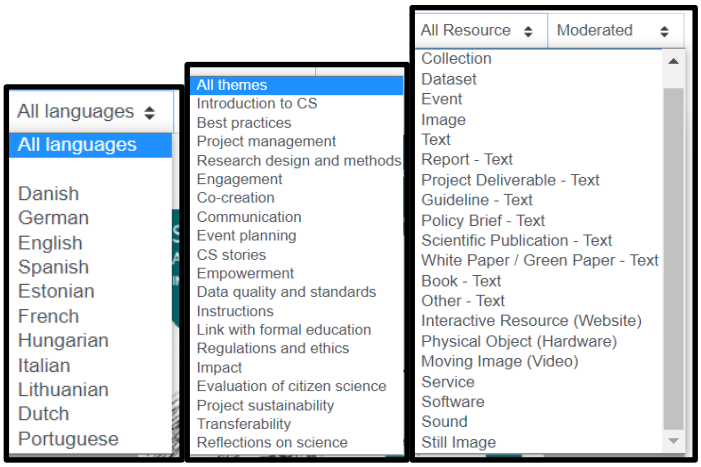


Figure 8. European resources filters.

When clicking on **resource**, there appear several fields:
-Name, reviews, abstract, license, audience, publisher, organisation(s), year of publication, resource's DOI, language, category, theme(s), keyword(s), URL.

When clicking on **training resource**, there appear the same fields plus educational level.

Organisations belong to a specific type, and have these fields:
-Logo, name, description, resources, members, latitude and longitude, country, contact point, contact email, URL.

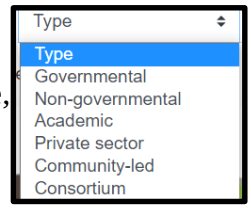


Figure 9. European organisations' filters.

(See Appendix C for more information regarding the EU-Citizen.Science structure).

3.2.2) Extracting the data from the European platform

The information from the European platform will be extracted in the first place since it has an API Therefore, it will be useful to analyse how they structure the data and construct the web scraper.

The method to get this data is quite simple thanks to the EUCS PLATFORM API[93]:

1. The first step is to sign in on the EU-citizen.science website[94] .
2. Then in a command line such as the Anaconda one, we follow the indications mentioned in the API web[93] to obtain the user authentication token. If the correct user and password are introduced, this command returns the requested token that will be used later on to get the desired information.
3. Once the authentication token is received, it just needs to call the API with the corresponding URL from where the data will be extracted[95].

[93] Link to the EUCS PLATFORM API (2021, March 18) https://eu-citizen.science/swagger/ .
[94] EU-Citizen.Science website (2021, March 18) https://eu-citizen.science/ .
[95] The last extraction of data was performed on the 20th of May to get as much content as possible (new entries are being added currently). The resulting datasets contain 176 projects, 117 organisations, 106 resources and 25 training resources.

If everything is ok, the response code will be 200, obtaining the corresponding body. Consider that the command line already downloads the data in json format and extracts it into a file (called EuropeanProjects.json). The same is done with the resources, training resources and organisations (but changing the URL and output file name), getting in this way all the desired information posted in the EU-Citizen.Science platform.

4. Once we have the four files with the needed information, we can open them in a Jupyter Notebook Python script using the address of the file and json.load(file) method.
5. In order to visualize it more clearly, there is a created function called jprint which shows the json object with an easy structure.
6. Also, to work with the information in an easier way, it will be transformed to a pandas dataframe with the function pd.Dataframe.from_dict(), which is very useful since the dataframe shows the data in a more visual way and makes it easy to make analysis.

3.2.3) Analysing the data

3.2.3.1) Analysing projects

When extracted from the request get function through the API, the projects file contains 34 fields for each entry (see Appendix C). However, the fields originDatabase (only filled in 3 projects, being CORDIS[96] in the three cases), originURL (only filled in 8 projects), originUID (only filled in 4 projects) and origin (filled at any project) will not be selected since they do not provide any useful information.

Before starting, some variable types must be corrected (longitude and latitude converted to float and dates to datetime). Afterwards, the text variables will be analysed using text mining. To facilitate the analysis scripts workflows, a file named "functions_t.ipynb" containing all the necessary functions was created.

ANALYSING TEXT FIELDS: AIM, DESCRIPTION, HOWTOPARTIPATE AND EQUIPMENT

First of all, a field called 'description_clean' is added to the project's data frame containing the **description** text with the "clean_text" function applied. Then the text is converted to tokens and standardised by removing the stop words in the corresponding language (in this case English) and stemming the text with the Snowball stemmer. In this way, the 'description_clean' field contains the standardised description and the description_std_tokens variable contains all the tokens for the projects' descriptions.
A plot of the 50 most repeated tokens for the description field is created, where 'citizen', 'project', 'data', 'scienc' and 'research' are respectively the most common terms (stemmed, that is why it is scienc instead of science). The descriptions vocabulary is composed of 176 words, and the average project description length is around 150 words (a graphic of the text length distribution generated with the "plot_text_length_distribution" function is included in Appendix D).

[96] CORDIS is the main source of results for the projects financed by the European funds since 1990. In fact, there is a "CORDIS - EU research projects under Horizon 2020 (2014-2020)" dataset which contains projects and organisations funded by the European Union under the Horizon 2020 framework programme for research and innovation from 2014 to 2020. https://cordis.europa.eu/es

When plotting the word cloud with the most common words in the projects' descriptions, there are two ways to do so:

- Generating a bag of words that counts the number of occurrences of each word as they are and plots the result, counting citizen and citizens as two different words since for this case it takes into account the words as they are.
- Generating a bag of words that first of all, stems the words and then counts the number of occurrences for each stemmed word, giving the following result:

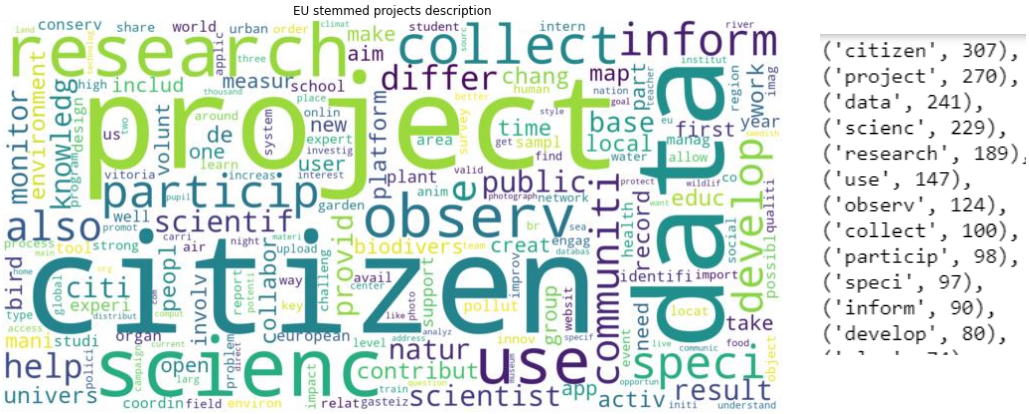


Figure 10. Word cloud (with stemmed terms) of the European projects description.

In this case, the result is more accurate since it counts citizen and citizens as the same word, and we can see the real distribution of the words.

The following analysis of the projects' descriptions aims to find which words are the most important for each project description. The way to do that is with a TF-IDF analysis: Term Frequency - Inverse Document Frequency. The TF*IDF is a numerical measure used to get the importance of a word for a document in a collection [45]. The term frequency (tf) counts the number of occurrences for a given word in a document, in this case in a projects' description. Then, the inverse document frequency (idf) measures if a term is common or not in a collection of documents. It is obtained by dividing the total number of documents (in this case projects) by the number of documents (projects) that contain that term, and the logarithm of this division is taken as result. Therefore, the tf-idf value increases with the number of times a word appears in a project description but at the same time it decreases with the repetitions of the word in the whole collection. With this method, we can get the words that are more relevant for each specific project. (It must be taken into account that to perform this analysis the text used will be stemmed, since otherwise the counting of words would not be fair).

A new column called "description_resultwords" has been added to the projects dataframe with the five more relevant words for each one (words with the highest tf-idf value for each project description). An example of this "description_resultwords" is the following:
[(hornet, 0.246), (insect, 0.112), (invas, 0.089), (wasp, 0.082), (manag, 0.069)]
for the "Vespawatch" project, or
[(cowslip, 0.205), (morph, 0.128), (flower, 0.108), (l, 0.093), (grassland, 0.077)]
for the "Looking for Cowslips" project, or
[(ladybird, 0.111), (differ, 0.055), (experi, 0.047), (swedish, 0.046), (expert, 0.046)]
for the "The Ladybird Experiment", which give a clear idea of what the projects are about without having to read the whole description text.

Regarding the project's **aim**, the same methods can be applied to analyse this field, obtaining the text length distribution plot included in Appendix D and observing that the length of the aim field is much smaller than the ones for the descriptions, being the average around 75 words. The corresponding word cloud is created, where the five most repeated words are exactly the same ones as for descriptions. Also, the five most relevant words (computed with tf-idf) for each project aim are stored in the corresponding variable.

Regarding the **howToParticipate** and **equipment** fields some more analysis will be done. First of all, the same text mining as with the description and aim is performed, but for these variables a more interesting analysis will be done.

Then, a new variable name needed (material) is created from joining the equipment and howToParticipate fields, since both of them mention the necessary material to take the project into practice. After studying several projects, the ways to participate in a project have been differentiated in seven classes:

```
app = ['app']
camera= ["camera", "image", "video", "photo"]
phone= ["phone"]
internet=["internet"]
computer=["computer"]
gps= ["gps"]
nothing=["nothing", "anything", "no equipment"]
```

Figure 11. Code classification to extract the needed material for the projects.

With these elements, a new data frame is created, including for each equipment the corresponding id, name, number of projects in which it appears and list with the ids of those projects. Therefore, the code will analyse all the projects and set to true the corresponding value if that way of participating is mentioned in the howToParticipate or equipment project field. Also, to make it easier each project's row will include a column name "needed_list" which includes the list of the corresponding items if they are mentioned in any of those two sections. In the following bar plot it is shown the number of projects that mention each of the elements in the howToParticipate or equipment text.

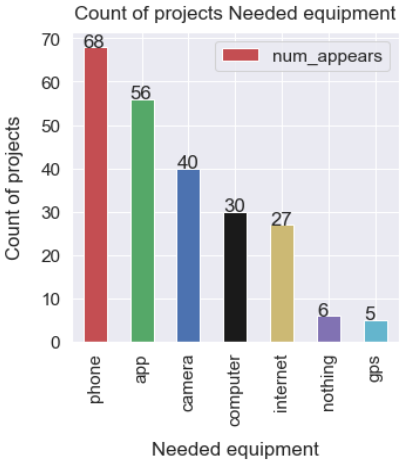


Figure 12. Needed equipment bar plot for European projects.

ANALYSING LIST FIELDS: PROJECTS KEYWORDS AND TOPICS

Each project contains a list of **keywords** and **topics** related to it. When analysing these fields, it is seen that there exist 181 different keywords and 29 distinct topics. The most used ones are "Biodiversity" and "Ecology & Environment" (plots in Appendix D).

In order to obtain a deeper understanding of the projects, the topics have been distinguished in four different groups (depending on their science type) that correspond to the primary disciplines of citizen science projects:

```
humanities_social=['Social sciences',  'Archaeology & Cultural',
'Indigenous culture', 'Science policy','Education','Geography']
physical_sciences=['Astronomy & Space', 'Geology & Earth science',
'Physics', 'Chemical sciences']
life_sciences=['Biodiversity', 'Ecology & Environment', 'Nature &
outdoors', 'Biology', 'Animals', 'Ocean, Water, Marine & Terrestrial',
'Birds', 'Long-term species monitoring', 'Climate & Weather', 'Insects &
pollinators', 'Biogeography', 'Genetics']
applied_sciences=[ 'Health & Medicine', 'Information & Computing sciences',
'Natural resource management', 'Transportation',  'Sound', 'Food science',
'Agriculture & Veterinary science']
```

Figure 13. Lists to classify the projects in the four primary disciplines.

The obtained results are:

- ➢ 73 out of 176 projects belong to humanistic sciences (41.5%).
- ➢ 21 out of 176 projects belong to physical sciences (11.9%).
- ➢ 122 out of 176 projects belong to life sciences (69.3%).
- ➢ 54 out of 176 projects belong to applied sciences (30.7%).

ANALYSING PROJECTS TIME FIELDS: STATUS AND DATES

There are 5 different projects **status**: Active, Completed, Periodically active, Not yet started and On hold. Most projects are active, while some of them are completed and only 6 of them are not started or on hold.

Moreover, we can take the statistics of the **starting**, **ending** and **created date** of the projects, counting the number of projects grouped by years. A lot of projects started in 2018, 2019 and 2020, corresponding to the rise of citizen science. Regarding the ending date, most projects are expected to finish between 2020 and 2030.

From the projects that have a determined starting and ending date, it can be calculated the duration. With these added duration columns (in months or years), more insights are generated: there are 80 projects with a determined starting and ending date. From those, we can extract that the mean duration is 3.25 years while the median is 2, maximum 17 years and minimum less than 1 year.

The projects also include the creation date, which means the day on which the project was uploaded to the "EU-CITIZEN.SCIENCE" platform. When plotting these dates, it is seen that most projects were uploaded between March 2020 and May 2020. Those months correspond to the first release of the platform.

(More information and graphics for this section are available in Appendix D).

ANALYSING PROJECTS FOUNDERS:

There are 47 projects without a determined **host**. Therefore, the other 129 do specify a project host. When showing the five most common hosts it can be seen that the "VA (Public & Science)" host, which is the most common one appearing 9 times, belongs to Sweden (SE). All projects related to it are completed and not done at home, and most of them belong to the fundingProgram H2020 MSCA- NIGHT, which corresponds to European Researcher's night, from Horizon 2020.

The 8 projects related to "Environmental Research Center of Vitoria-Gasteiz city council" belong to Spain and have as funding program the "Summons of assistances for the development of the Scientific, technologic and Innovation Culture (Convocatoria de ayudas para el Fomento de la Cultura Científica, Tecnológica y de la Innovación)"

There are 64 projects that specify a **funding body**. From those, 32 of them are distinct, and the most repeated ones are: "European Commission" (appears as a funding body for 19 projects), "Arts and Humanities Research Council" (finances 5 projects), and "Fecyt-Ministerio de Ciencia y Tecnología" "European Research Council (ERC)" that finance 4 projects each.

The **Funding Program** that appears in more projects (7) is the "Summons of assistances for the development of the Scientific, technologic and Innovation Culture (Convocatoria de ayudas para el Fomento de la Cultura Científica, Tecnológica y de la Innovación)", all of them belong to Spain and have as host the "Environmental Research Center of Vitoria-Gasteiz city council", as seen before.

There are 26 different project **countries**. The most common ones are: Spain (39 projects), United Kingdom (18), Italy (15), Sweden (15) and Netherlands (9). Also, from the totality of projects, almost 48% of them are **doable at home**. And only 6 out of all projects are **highlighted (featured)**: Natura sulle Mura (from Italy), STOC (Suivi Temporel des Oiseaux Communs) (from France), OpenTEK (from Spain), BioBlitz a tutti i costi (from Italy), BeBirds: Belgian Ringing Scheme (from Belgium) and FLAMENCO (from Belgium too).

(Lists with the most common hosts, organisations and main organisations are available in Appendix D).

PROJECTS RELATED TO EDUCATION

After analysing the project's information and fields, some different ways to know if a project is related to education have been extracted.

1) The first and more clear one is if the project has "Education" in the list of topics.
2) Then, if the project contains any of the educational keywords it also means that it is related to education.

education_keywords=['Schools', 'Education', 'Classrooms', 'Special Education', 'Teachers', 'Aire Marine Educative']

3) Finally, if the project mentions any of the following educational words in the description, aim or howToParticipate text fields:

education=["school", "teacher", "student", "educat", "pupil"]

So, two different columns have been added to the projects' data frame, one determining if the project mentions an educational word in one of the text fields and another one indicating if the project has "Education" as a topic or contains an educational keyword.
Out of 176 projects, 54 of them have Education as a topic or an educational keyword, while 69 mention school, teacher, student, educat or pupil in description, aim or howToParticipate field. In total, 86 of the 176 projects are related to education in one of these three ways, which represents a 49% of the projects' totality.

CORRELATION AMONG PROJECTS VARIABLES

A new dataframe to explore correlations is done, containing some columns of the EU_projects plus the needed material (one column for each so that the variables are True/False). Since for correlations we need numerical values and not categories, several variables are generated from the categorical ones, determining if that value is included or not (True/False). For fields such as the status and countries, they will be converted to numerical types by using the OrdinalEncoder, which maps categories to numeric values. Since there are 26 countries, 6 groups have been made depending on their geography. For the status, they have been classified into Active, Completed or not being done at the moment.

The pairs of variables with more correlation are:

- Duration months and duration years, end_ date with duration (months and years), which are obvious.
- Education_in_keywords_topic and topic_social, which means that most projects have educational keywords or belong to social humanities topics, which makes sense.
- Then, need an app and need a phone, which of course are related since if an app is needed to participate in a project, it will need a phone too.
- Also, status and ending dates are correlated since the end date determines the status of the project.
- This correlations' list also shows a fact that may seem weird at the first glance. The thing is that there exists a correlation between the ending date or duration of a project and the need for a phone, app or camera. In the beginning, I thought that it may be by chance, but then using the Tableau dashboard I realised that it is true, depending on if a project needs a phone or not the ending date is modified (most projects that need an app finish at 2030). This makes sense since old projects could not ask for an application since smartphones were not that common. However, nowadays technology is much more evolved and projects take advantage of it and ask citizens to use applications, smartphones and cameras to collect information for the CS projects. That is the reason why the end date and the need for technology are related.

However, this correlation analysis should not be very strictly considered, only as a guide to see what filters to apply while playing with the Tableau visualization. The reason is that several variables are categorical and when encoding them to numeric values we are losing precision, especially if from 26 values as for countries they are reduced into 6. Therefore, these obtained correlations are useful to try and verify them with the Tableau dashboard, where these and more analysis among project variables will be done to extract meaningful insights.

3.2.3.2) Analysing resources

To analyse the European resources, analysis methods very similar to the ones performed with the projects will be applied. First of all, text mining will be used to obtain the abstract word cloud, length distribution (average around 100 words) and the most relevant terms for each resource (with tf- idf). (Word cloud in Appendix D).

Then, depending on if the resources' abstracts mention toolkit, activity, course or webinar they have been classified into these four types (with the same method as for projects):

- Activity: 18 resources.

- Toolkit: 6 resources.
- Webinar: 3 resources.
- Course: 2 resources.

A column named type has been added to the resources to indicate if it belongs to any of these 4 types.

Regarding the resource's **authors**, a list with the most common ones (out of 2014 distinct) has been extracted. However, it is not included here to save their anonymity and respect the GDPR[97], even though they have given consent to appear in the citizen science resources descriptions.

The resources can have more than one **audience,** out of the 7 distinct types. By counting the number of resources each audience appears in, it can be extracted that the most common ones are:

1. Researchers & Academics: 52
2. CS Project Leaders & Initiators: 49
3. Community Members & Citizens: 48
4. ALL Audiences: 39
5. Cos & NGOs: 34
6. Educators: 33
7. Policy & Decision Makers: 31

Resources can also contain more than one **keyword,** out of the 144 distinct types. By counting the number of resources each keyword appears in, it can be extracted that the most common ones are: Environment (8 times), Evaluation (7 times) and Overview (7 times). Also, each resource can belong to more than one **theme**, out of the 19 distinct. Again, by counting the number of resources each theme appears in, we can see that the most common ones are: Best practices (48 times), Introduction to CS (35 times) and Research design and methods (31 times).

Regarding the resources' **categories**, each resource belongs only to one category type, out of the 14 that exist. The most common ones are: Guideline (31 times), Event (14 times) and Interactive Resource (Website) (13 times).

The resources **publication and uploaded dates** can generate insights when grouping resources by the years of those dates. In Appendix D, histograms for the dates in which resources were published and uploaded can be found. It can be seen that most resources are published between 2018 and 2021, coinciding with the years that more projects started (citizen science starting to become more and more relevant to society). The uploaded dates are segmented in months, since the platform was created in March 2020, the month in which most resources were uploaded. After that time, more resources have been added but bit by bit.

76 out of the 106 resources are written in **English**, and only 6 of the totality are featured.

[97] GDPR website (2021, June 3) https://gdpr-info.eu/ .

RESOURCES RELATED TO EDUCATION

After analysing the resources' information and fields, some different ways to know if a resource is related to education have been extracted.

1. The first and more clear one is if the project has "Link with formal education" in the list of themes.
2. Also, if the list of audiences contains "Educators".
3. Then, if the project contains any of the educational keywords it also means that it is related to education.

education_keywords=['Educators', 'Teachers', 'Education', 'high school', 'school', 'secondary', 'primary']

If any of these three conditions is fulfilled in a resource, the created variable "inEducation" is set to True.

Out of 106 resources, 33 of them have Educators as an audience, 12 have "Link with formal education" in themes, and 9 of them contain at least an educational keyword. In total, 35 of the 106 resources are related to education in one of these three ways, which represents a 33% of the resources' totality.

3.2.3.3) Analysing training resources

Exactly the same analysis as with resources is done with training resources, but it is done separately to obtain the different results and contrast them. In Tableau they will be uploaded together since there will be a filter to select what type of resources are desired (normal, training or both). So, the **abstract** text distribution graph and word cloud are generated. Also, the five most relevant terms for each training resource are included in the abstract_resultwords column, to get an overall idea of what the resource is about.

In this case, training resources have also been classified into four **types** (webinar, activity, course and toolkit) depending on if those words are mentioned in the abstract or not, getting that out of the 25 posted training resources, 10 of them are courses (including 5 activities), 3 are webinars and there is 1 toolkit.

The only **author** that appears in more than one training resource is the "UCL", which corresponds to the London's Global University and appears in 5 resources, and the most common **audience** is "Researchers & Academics", appearing in 15 training resources, followed by "CS Project Leaders & Initiators" (in 10) and "Educators", in 8.

Regarding the **categories**, most resources (13 of 25) are "Interactive Resource (Website)", followed by 4 videos and 3 events. Also, the most common **theme** is "Introduction to CS" (appearing in 13 training resources), followed by "Best practices" (in 9) and "Engagement" (in 6).

The resources **publication and uploaded dates** can generate insights when grouping resources by the years of those dates. In Appendix D, histograms for the dates in which training resources were published and uploaded can be found. It can be seen that most are published between 2018 and 2021, coinciding with resources and projects (citizen science starting to become more and more relevant to society). The uploaded dates are segmented in months, since the platform was created in March 2020, the month in which most resources were uploaded.

The 25 training resources are in **English** and only 4 of them are featured, all belonging to "UCL".

TRAINING RESOURCES RELATED TO EDUCATION

The same condition as with normal resources are used to determine if a training resource is related to education or not, getting that 8 resources have Educators as an audience, 2 are linked with formal education and 1 contains an educational keyword. In total, 8 out of 25 resources have an educational keyword, audience or theme, which represents a 32%, a very similar percentage to the one for normal resources (33%).

3.2.3.4) Analysing organisations

In the projects' information there are 43 main organisations and 33 organisations. In the organisations' section there are 117 organisations. The 43 main organisations and the 33 organisations are all included in the organisations' section. Therefore, all organisations related to the projects appear in the organisations' section of the platform, as expected.
First of all, a text analysis with the organisations' descriptions is done, obtaining the typical plots and getting the five most relevant words for each organisation.

In this case, we wanted to know if organisations are related to education, and for this purpose it was checked if the word university was mentioned from one side, and if the description contained any educational word such as education, school, class, student or teacher. With this analysis, it was extracted that the words educat(e/ion), school, class or student appear in 52 organisations while the word university appears in 43 organisations, having a global of 68 organisations out of 117 related to education, which represents a 58%.

Each organisation belongs to a country, where Spain, Denmark and Italy are the most popular ones (plot at Appendix D). Also, each organisation is of one specific type out of the 6 determined, and in this case 52 of the total are Academic organisations, followed by non-governmental (30) (whole list in Appendix D).

3.2.3.5) Relating all of them

An interesting analysis is to investigate if projects are related to organisations, or resources are linked to projects. In this way, we will see the percentage of resources that exist for themselves or if most of them have been created in relation to a specific initiative. Regarding organisations, all organisations related to the projects appear in the organisations' section of the platform, as expected. So the platform contains 117 organisations, 43 of which appear in projects as main organisations and 33 as organisations. In global, 67 organisations do appear in projects (9 of them appear both as organisations and main organisations) while 50 do not. Thus, a 57% of the posted organisations are linked to a posted initiative.

In order to find how many resources are related to projects, a previous step must be done. Names of the projects are often very large, including a whole sentence with a short project description. Therefore, when searching if a resource contains a specific project name, most times it would say no since the whole name would not be included in the resource, despite the project may be mentioned, with its concise name. That is the reason why project names have been split and cleaned to generate a new column named clean_name, which gets only the important part of the name (splitting if the old name contained a punctuation mark, which meant that it included a brief description).

Once the projects contain a clean_name field, it just needs to be checked if a resource mentions a specific project name, which means that they are linked. In that case, the project name is added to the "related_initiative" created column. The obtained results are that 16 resources and 2 training resources are related to projects, which represent a quite small percentage of the total. Also, 37 resources and 4 training resources are linked to organisations.

3.2.4) Generating visualizations

For a more visual and interactive interpretation of the obtained information, several visualizations have been done to play with the "EU.Citizen.Science" platform data and get an overview of the citizen science situation in Europe through its content.

With the script European_analysis.ipynb, the generated data frames EU_projects, EU_resources and EU_organisations were exported to Excel files (.xlsx) containing columns with the text's resulting words list, educational variables, needed material list, topics classification…

Regarding the uploaded **projects**, four different dashboards have been created:

1. A **general** one showing the total number of projects distributed in a pie plot by countries and classified into a tree map by their status, and also divided into doable at home or not and educational or not. Moreover, several bars show the number of projects that need specified equipment to participate in, and there is a diagram for each science discipline indicating the percentage of projects that belong to each. To see the relations among variables, projects can be filtered by any field: doable at home or not, featured or not, educational or not, by the four topics, by country or by starting, ending date or current status.

When playing with it, some conclusions can be extracted. First of all, it is clear that most projects belong to Spain (a 22%) and that the most predominant status is active (only 22% of projects are not active now). Regarding the projects' topics, the most typical one is life sciences, followed by applied sciences and social. The material that more projects require are phones, which is related to apps (that is the second element) and cameras (third).

Filtering by projects doable at home, it can be seen that the percentage of them that require internet increments, and also the number of social projects. Of the 22 completed projects, very few required any technological equipment, since probably they were done before smartphones were so common. The 11 Swedish projects are linked to education and any of them is doable at home, also 10 of those are completed. Obviously, most educational projects belong to social science and most social projects are educational. Finally, from the 7 projects starting this year 2021, 5 of them mention the need for a phone.

2. A dashboard with the projects distributed in a map, in which countries are painted in a darker green depending on the number of projects they have. Also, each individual initiative is pointed in the map with a circle if it is doable and a square if it is not. The colours of the marks represent the projects' status, orange means it is active (active or periodically active), blue means completed, and red is not yet started (including on hold). The size of the shapes is determined by the project duration in years. Some filters can also be applied to see the geographical distribution of projects.

There is another map dashboard that has been made to see the projects' topics geographical distribution. So they can be filtered by their scientific discipline and also the keywords they contain (by writing them in the topics_list_clean filter, which looks for that keyword in the projects' topics list). The names of each topic have been included in order to show which keywords can be searched. Finally, another map dashboard has been created but this one is very similar to the first one with the difference that projects are shown by their starting year, so that the evolution of citizen science projects in Europe through time can be seen.

3. A dashboard that shows the more relevant words from the description and aim of each project, so that you can get an idea of what they are about. These projects can be filtered by topic and if they are educational or not. Also, you can choose to see only the featured projects, or directly search for a specific one with the name filter.
4. A dashboard with the projects' dates, that shows the timeline of starting and ending dates, plus the dates in which projects were uploaded into the platform. The orange bars show the number of projects while the blue line represents the average duration in years of the projects corresponding to that moment. That is why projects starting in the past or ending in the future have a very high duration. Moreover, there is a special graph for this duration field that shows the project's duration (in years) distribution, where the maximum is 17 years. This dashboard can be filtered by the projects' status, to analyse the relation with the projects' dates.

When talking about **resources** (both normal and training resources, they are mixed but can be filtered), three different dashboards have been made:

5. One showing the most relevant **words** of each resource abstract, so that you can get an idea of what they are about. These resources can be filtered by type (normal or training resources), category, educational or not and by name.
6. Another one showing the **timeline** of when resources were published and uploaded. Most of them were uploaded into the platform in March 2020, when it was created. And most resources have been published between 2018 and 2021, when citizen science started to be more popular. They can also be filtered by type of resource, educational, featured, category and language.
7. A **general** dashboard that shows the resources classified by **audiences**, educational, type, languages and categories. It is seen that most of them are guidelines or websites and only 33% are educational. Of the 131 existing resources, 101 are in English (the 25 training resources are all in English) and 67 have as an audience Researchers and academics. By writing "CS" in the themes filter it can be seen that 68 resources contain "Introduction to CS" or "CS stories" as keywords. And writing "Educators" in the audiences' filter 41 resources are got, from which 8 are training resources.
8. Referring to **organisations**, one dashboard showing them distributed in a world map has been created. It also indicates the number of organisations for each type (academic, community-led, consortium, governmental, non-governmental and private sector). From the 117 organisations, 52 of them are academic, other 52 mention educational words and 43 are related to universities.

(Images of the generated dashboards can be seen in Appendix E, and the links to the dashboards are available in Appendix B).

3.3 - Spanish platform: The “Observatorio de la Ciencia Ciudadana en España”

3.3.1) Platform data structure

The platform differentiates between initiatives, which can be citizen science projects, persons or institutions, and resources, which include educational resources, guides and methodologies, infographics, presentations and posters, metrics and indicators, and reports.

The initiatives can be searched by field of knowledge, initiative type (institution / project/person of citizen science) or text / keyword. And the structure of these initiatives is the same for the three types. A table with all the fields is included in Appendix F.

Regarding the resources, the platform shows Educational Resources, Guides and Methodologies, Infographics, Presentations and Posters, Metrics and Indicators, and Reports[98]. All of them have the same basic information, which is what will be extracted: name, URL, abstract and keywords (the ones in bold style).

3.3.2) Extracting the data from the Spanish platform

As mentioned before, the European platform has an API that is very useful to get the information but the Spanish one does not, therefore an important part of this project was to design a crawler in order to extract the information from the Spanish website[99]. In the state of the art, it has been mentioned that there exist several web scraping software that would work for this case, from which Selenium has been the chosen one. This tool will make it possible to get the desired data by examining the website elements and access all the required fields using python, the programming language data analysts are more familiar with.

Now, the steps followed to design the web crawler in a Jupyter Notebook environment are explained:

```
pip install selenium
```

Figure 14. Code.

1. Download the web driver: https://chromedriver.chromium.org/downloads and save it in a known path.
2. Import the webdriver:

```
from selenium import webdriver
driver= webdriver.Chrome('./chromedriver.exe')
```

Figure 15. Code to import the webdriver.

3. Go to the main page that contains the projects, specifying the URL:

```
driver.get('https://ciencia-ciudadana.es/proyecto-cc/')
```

Figure 16. Code.

[98] It also has videos as resources, but these ones will not be analysed.

[99] The last extraction of data was performed on the 16th of May to get as much content as possible (new entries are being added currently). The resulting datasets contain 295 initiatives, 117 organisations, 106 resources and 25 training resources.

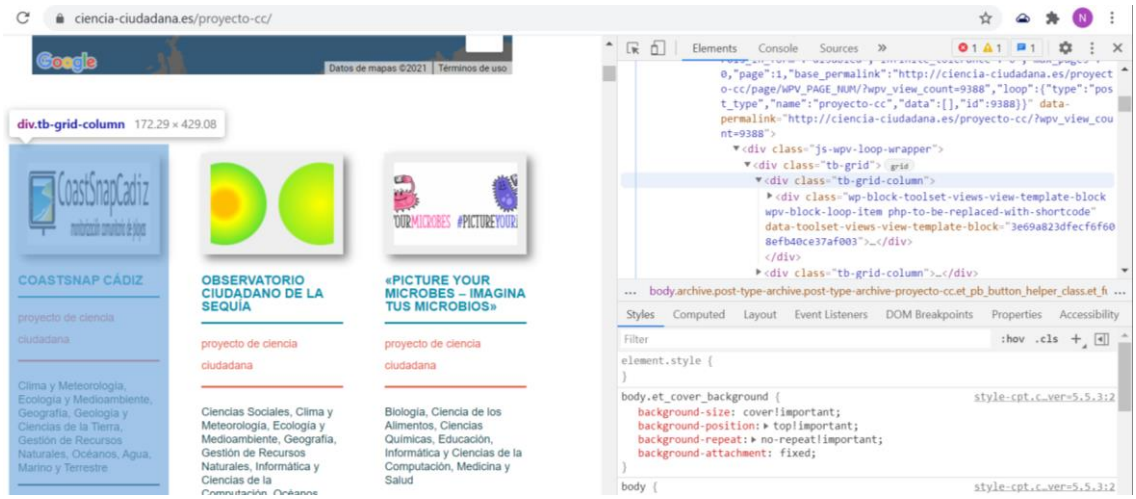


Figure 17. Elements of the Spanish platform website.

4. Analyse the webpage: “Ajustes” -> “Más herramientas” -> “Herramientas para desarrolladores” and find the object that contains the project titles and URLs:
5. Go to this element by using the method “find_elements_by_xpath”

```
projects = driver.find_elements_by_xpath('//div[@class="tb-grid-column"]').
```

Figure 18. Code.

Now in projects there are all the objects that contain the projects, one for each.

6. Iterate through all the projects and store the project title (text) and URL (href attribute) into a projects_list, so that later on each project can be analysed.

```
projects_list=[]
for project in projects:
    title=project.find_element_by_xpath('.//h4[@class="tb-heading has-text-color"]/a')#.text
    link = title.get_attribute('href')
    this_project={"Title": title.text, "Link": link}
    projects_list.append(this_project)
driver.close()
```

Figure 19. Code to store projects' titles and URLs.

7. Once the list of projects with all the titles and URLs is completed, it can be iterated to get all the necessary information of each project. In order to do that, all the desired fields will be stored in a dictionary that will be inserted into the dictionaries_list. To access the desired attributes, it must be checked to which element they belong and then get it with the “find_element_by_xpath” method.

```
driver= webdriver.Chrome('./chromedriver.exe')
    #Create a list of dictionaries
    lista_diccionarios=[]
    #Iterate, for every project in the list: store the desired fields in a dictionary and append the dictionary to the lista_diccionarios
    for i in range(len(projects_list)):
        project_dict={}

        driver.get(projects_list[i]["Link"])
        left_area=driver.find_element_by_xpath("//*[@id='left-area']")
```

```
        idd=left_area.find_element_by_xpath(".//*").get_attribute("id")

        project_area=left_area.find_element_by_xpath("//*/div[1]/h1")
        project_dict["id"]=idd
        project_dict["name"]=project_area.text
        project_dict["type"]=left_area.find_element_by_xpath("//*/div[2]
/div/div/div[2]/div[3]").text
       #get also all the other fields (code included in the script)
       lista_diccionarios.append(project_dict)
driver.close()
```

Figure 20. Iterating the projects to get the desired information.

8. The same is done for all the resources and the Barcelonian initiatives, by adapting the code to get what is necessary. For the cases in which the list occupies more than one page, the element that stores the different pages is found and saved in a variable called num_pages, so that later on it is iterated to store the corresponding link of each page (by getting the "href" attribute).

```
num_pages = driver.find_elements_by_xpath('//a[@class="wpv-archive-
pagination-link js-wpv-archive-pagination-link page-link"]')
links_pages=[]
    for page in num_pages:
        links_pages.append(page.get_attribute("href"))
```

Figure 21. Code to ensure it iterates through all the pages.

Then, the link_pages variable is iterated, navigating to all the pages to get all elements appearing in the platform.

9. Finally, the list with the corresponding content is transformed into a dataframe and exported as a json file, with the "utf-8" encoding, since otherwise text would not be exported as expected.

```
df_SP_projects=pd.DataFrame.from_dict(lista_diccionarios)
if (EXPORT_JSON):
    with open('df_initiatives_SP.json', 'w', encoding='utf-8') as fp:
        df_SP_projects.to_json(fp, force_ascii=False)
```

Figure 22. Exporting the generated file.

3.3.3) Analysing the data

In the Spanish platform there are 204 citizen science projects, 69 citizen science institutions, 22 citizen science persons, 24 educational resources, 20 guides, 9 infographics, 22 reports and 6 metrics.

3.3.3.1) Analysing initiatives

Since Spanish initiatives can be of three types: projects, institutions or people, first of all, a graphic with this distribution is shown. What is done next is cleaning the initiative names as it was done with European projects and organisations. That is because usually the names contain information that is not vital, and by splitting the string the clean names are obtained and stored in a new column.

First of all, some text mining is performed with the **aim** and **description** fields, cleaning them, generating the corresponding word clouds and obtaining the most relevant terms

for each initiative. Then, the **starting** and **ending dates** fields are analysed (199 initiatives determine a starting date and 140 an ending date. By grouping them, it can be seen that most initiatives started from 2010 on, having a peak in 2015 and 2020. Regarding the ending dates, a big amount of them are in 2020, since probably oldest projects set as ending dates that year. With those values, statistics from the initiative's duration can be extracted. From the 138 computed durations (initiatives that have a starting and ending date), the median of initiatives duration in years is 6.

Of the 295 specified **main Organisations**, 294 are distinct, only the "Identificación De Asteroides Cercanos A La Tierra" appears twice. Also, from the 123 different mentioned **responsible entities** (139 initiatives specify this field), the most common ones are "Bifi de la Universidad de Zaragoza" (7 times) and "Fundación Ibercivis" (5 times).

Creating a "get_people" function that uses nltk to identify people names from text [45], from the workTeam field a dictionary containing all people appearing in workTeams of initiatives has been created. Once this list of people is computed, the number of initiatives each person appears in can be computed by iterating for all the initiatives and checking if they mention that name. In the Spanish_analysis script this list is shown, however it has been preferred not to show the people's names in this document to preserve their anonymity. However, by printing the lines in which they appear we can deduce what is their job position, obtaining that the professionals more related to CS initiatives in Spain are:

1. Executive director of Ibercivis (appears 13 times)
2. Ibercivis main investigator of a project (11 times)
3. Ibercivis Foundation (7 times)
4. Head of communication at the Ibercivis Foundation (7 times)
5. Coordination and events manager at the Ibercivis Foundation (7 times)
6. General coordinator of Ibercivis foundation (7 times)
7. Ibercivis IT developer (5 times)

Therefore, it can be seen that all of them belong to the Ibercivis foundation.

There are 32 different keywords, from those the ones appearing in more than 100 initiatives are "Ecology & Environment" and "Biodiversity", as for the European ones. These 32 topics have been distinguished in humanities_social, physical_scieces, life_sciences and applied_sciences, following the same classification as for the European ones. A Boolean column for each science discipline has been added, being true if the corresponding initiative belongs to that science or not (being possible to belong to more than one). The obtained results are:

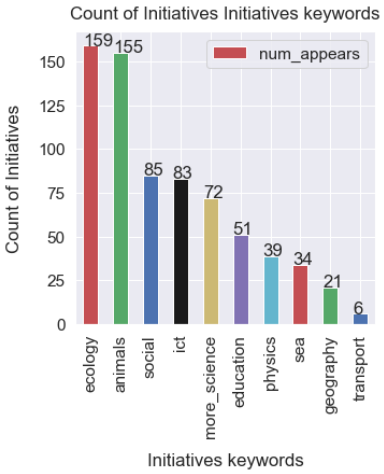


Figure 23. Spanish initiatives' keywords

- ➢ 124 out of 295 initiatives belong to humanistic sciences (42%).
- ➢ 47 out of 295 initiatives belong to physical sciences (15.9%).
- ➢ 181 out of 295 initiatives belong to life sciences (61.4%).
- ➢ 151 out of 295 initiatives belong to applied sciences (51.2%).

As for the European case, a column name **needed_list** has been added, determining if the initiative needs an app, camera, device, GPS, kit or nothing to participate in it. Also, in this case the values social_network, survey and event are included. A plot with the required material, **public** and **participants** distribution is included in Appendix G.

There is a field that determines the **province** of the initiatives, but some of them are not correct, so a column named new_province is created including the standardized province names[100]. However, for a big number of initiatives this province field is empty, even though they refer to a specific province in the initiative description. Therefore, we are going to induce the provinces for some initiatives that do not have it filled but do mention a province, meaning that they belong there. After that, 56 initiatives that did not have an assigned province have been deduced with this text analysis (plot of the provinces' distribution at Appendix G).

For the **results summary** field, five distinctions have been made, determining if the initiative has a result report, map or tool, and if they are in the starting or developing phase (results in Appendix G). And regarding the **results link**, out of 295 initiatives only 82 have a direct link to a web page and 2 have available a DOI resource code.

In the Spanish platform, initiatives have a field called **impact** in which they determine the impact for that specific project, being educational, scientific, social, environmental, technological, political or economic. The **impact examples** field contains text describing specific impacts that the initiative has generated. For this case, a classification into 6 types has been made: "noticia, comunicar, participación, educación, institución y visibilidad", depending on what they have achieved. Moreover, there is an initiative motivation field that explains the reason behind the initiative. It is filled only for 63 of them, from which 29 have as purpose persuading citizens to collaborate into science.

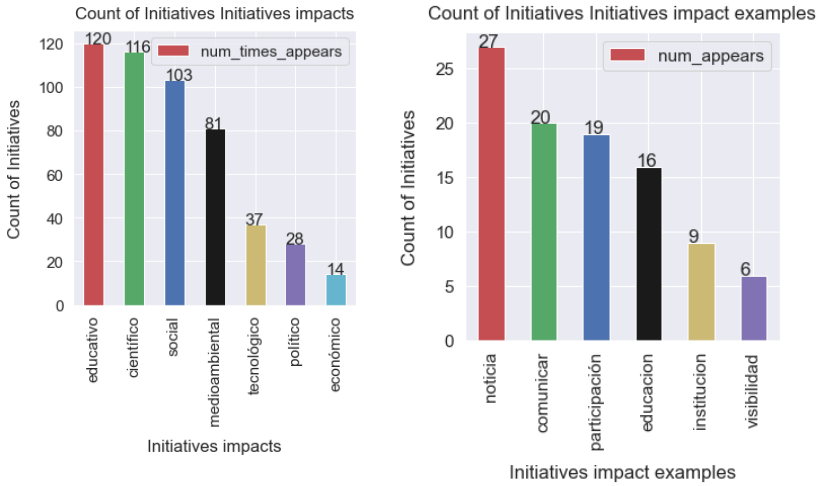


Figure 24. Spanish initiatives' impacts.

INITIATIVES RELATED TO EDUCATION:

After analysing the initiatives' information and fields, some different ways to know if an initiative is related to education have been extracted.

[100] *The initiatives having more than one related province have been assigned to the first-mentioned one.

1) The first and more clear one is if the initiative has “Educación” in the list of keywords.
2) Then, if the project contains “educativo” in the impacts it also means that it is related to education.
3) Finally, if the project mentions any of the following educational words in the name, aim, subtitle, description, howToParticipate, equipment, results summary or motivation text fields:

education=["educativ", "educación", "alumn", "profesor", "docente", "escuela", "instituto", "colegio", "enseñanza"]

Out of 295 initiatives, 132 of them have Education as keyword or in impact, while 96 mention an educational element in any of the text fields (34 of them mention education in text but it is not in keywords or impact). In total, 166 initiatives are related to education in one of these three ways, which represents a 56% of the initiatives’ totality.

3.3.3.2) Analysing Spanish resources

In order to analyse the resources published at “Observatorio de la Ciencia Ciudadana en España”, a data frame containing all the resources from the different types has been generated, with 24 educational resources, 22 reports, 20 guidelines, 9 infographics and 6 metrics. The only information that could be obtained from these resources are the names, URLs, the text that was in bold and the abstract. Therefore, the only analysis that can be done is performing text mining with the abstract field, generating a word cloud, text length distribution plot and resulting more relevant words for each resource.

3.2.3.3) Relating Spanish initiatives and resources

By checking if resources contain any project name, it can be determined that 66 resources are related to initiatives, which represents a 81,5%. From those, 33 are connected to a CS institution, 31 to a CS project and 2 to a CS person.

3.2.4) Generating visualizations

Regarding the uploaded **initiatives**, four different dashboards have been created:

1. A **general** one showing the total number of initiatives distributed by projects, institutions and persons. There is also the list of initiatives that have each type of impact, that belong to each of the four disciplinary sciences and that are connected to education in one or another way. Also, you can see the number of initiatives that require each material to participate in, and two pie plots showing the project's participants distribution and the public they are aimed at. To see the relations among variables, initiatives can be filtered by any field: type, province, topic, educational, participants and public.

When playing with it, some conclusions can be extracted. First of all, it is clear that most initiatives are projects, followed by institutions and then 22 CS persons. 120 out of the 295 initiatives have an educational impact, while 116 also have scientific and 103 social. This is connected with the fact that 181 initiatives belong to life sciences, 151 to applied and 124 to social sciences. The most required equipment are devices (phones or computers), and lots of initiatives are linked to events. Regarding the participants, only a 36% of the initiatives have this field filled. From those, 32 have between 100 and 1.000 participants and 28 between 1.000 and 10.000. Something similar happens with the public, where only a 58% of initiatives indicate the public they are directed for. From

those that do, 142 are aimed at any kind of public, 13 to adults and 15 to young people. The 16 initiatives that have more than 10.000 participants are directed to any kind of public, and when the participants decrease sometimes they have a specific public (meaning that they are aimed at a specific society range and that is why they have fewer participants). From the 12 CS projects aimed at youngsters between 12 and 18 years, 10 of them have an educational impact, and also 10 are connected to physical and social sciences.

2. Then there is a dashboard with the initiatives (filtered by projects, institutions and persons) distributed in a map, in which Spanish provinces are painted in a dark colour depending on the number of initiatives they have. It can be seen that Madrid, Barcelona and Zaragoza are the ones with more citizen science movement, coinciding with the most important cities of Spain. There is a region in France called without where the projects that do not have a related province are indicated. That has been done in order to not ignore those initiatives and take them into account, but knowing that they are not linked to a specific region in Spain. The timeline button can be actioned and the top map will show the initiatives started by year, being able to see then the evolution of citizen science in Spain (it can be seen that each year that goes by more and more CS initiatives are created).

3. One showing the most relevant **words** of each initiative aim and description, so that you can get an idea of what they are about. These can be filtered by type (project, institution or person), by topic or directly by name.

4. A dashboard with the initiatives' **dates**, that shows the timeline of starting and ending dates. The orange bars show the number of initiatives while the blue line represents the average duration in years of the activities corresponding to that moment. That is why initiatives starting in the past or ending in the future have a very high duration. Moreover, there is a special graph for this duration field that shows their duration (in years) distribution. This dashboard can be filtered by type (projects, institutions or people). When selecting projects (the only type that makes sense to analyse its duration), it can be seen that most projects start between 2014 and the actuality, and a big amount ended in 2020 (probably all projects uploaded that were already finished were set as ending date 2020). For those, the median project duration is 4 years, coinciding with the duration of the "COMPASS" Spanish citizen science project.

5. Regarding the Spanish **resources**, a dashboard that shows the amount of each resource type has been done. It also shows the most relevant words for each resource abstract, to get an idea of what it is about. Of the 81 Spanish resources, 24 are educational resources, 22 are reports and 20 are guidelines.

(Images of the generated dashboards can be seen in Appendix H, and the links to the dashboards are available in Appendix B).

3.4 - Barcelonian platform: The “Oficina de la Ciència Ciutadana”

3.4.1) Platform data structure

The Barcelonian platform includes projects, activities, news and a calendar. But these last three fields have very little information due to the fact that the program is relatively new and the pandemic has stopped the activity of the city in all its aspects, including CS. Therefore, the only section that will be analysed is the one including citizen science projects. When entering a project, the appearing fields are: the project name and webpage, a short description, status, topic and different activities related to it.

3.4.2) Extracting the data from the Barcelonian platform

Since the Barcelonian platform does not have an API either, the same process as for the Spanish one has been followed. With the Selenium library, a web scraping script has been generated to get all the desired information posted on the website (always checking that the robots.txt allows doing so, done in section 2.3.1).

So the first three steps are reused for this case, changing the URL to 'https://www.barcelona.cat/barcelonaciencia/es/proyectos-ciencia-ciudadana',

4. Analyse the webpage: “Ajustes” -> “Más herramientas” -> “Herramientas para desarrolladores” and find the object that contains the project titles and URLs:

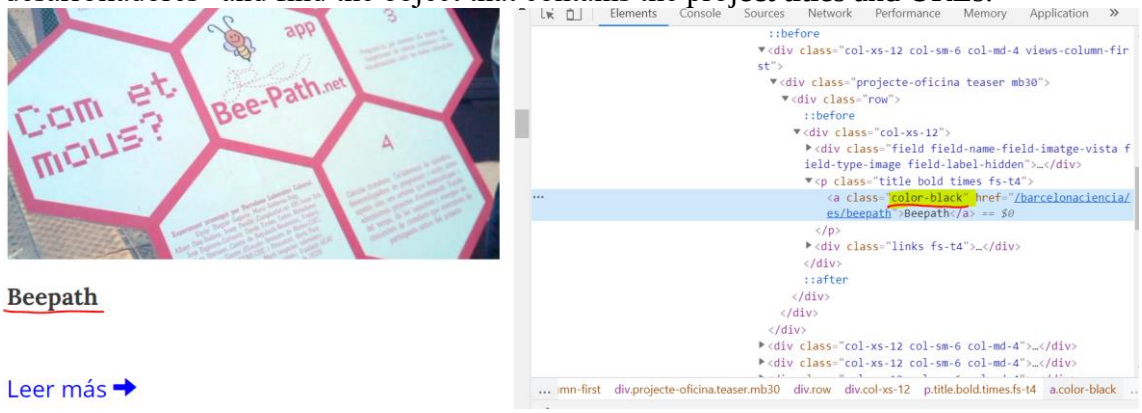


Figure 25. Inspecting Barcelonian website elements.

5. Go to this element by using the method “find_elements_by_xpath”.

```
projects = driver.find_elements_by_xpath('//div[@class="color-black"]').
```

Figure 26. Code.

Now in projects there are all the objects that contain the projects, one for each.

6. Iterate through all the projects and store the project title (text) and URL (href attribute) into a projects_list, so that later on each project can be analysed.
7. Once the list of projects with all the titles and URLs is completed, it can be iterated to get all the necessary information for each project. In order to do that, all the desired fields will be stored in a dictionary that will be inserted into the dictionaries_list. To access the desired attributes, it must be checked to which element they belong and then get it with the “find_element_by_xpath” method.
8. Since not all projects contain all the fields (or they have different strong expressions), the programming expressions try and except are used to check if it is possible to get the desired information. Here an example is shown:

```
try:
   try:#if the strong expression does not include the white space
       text = driver.find_element_by_xpath("//p[strong =
'Estado:']").text.strip()
   except:#in this case the strong expression includes the space
       text = driver.find_element_by_xpath("//p[strong = 'Estado:
']").text.strip()
       project_dict["status"] = text.split(":")[-1].strip()
except:#in case that field does not exist
    project_dict["status"]=""
```

Figure 27. Code example.

9. Finally, the list with the corresponding content is transformed into a dataframe and exported as a json file, with the "utf-8" encoding, since otherwise text would not be exported as expected.

3.4.3) Analysing the data

The Barcelonian initiatives only contain the project names, descriptions, URLs, status, topics and related activities. Therefore, few analyses can be done. A word cloud with the most common terms in **descriptions** can be generated, as well as a text length distribution plot and the obtaining of the most relevant words for each project. Also, by checking if the description text contains the word "phone", it can be deduced if an initiative is or not related to phones (8 out of 20 projects mention phones). Plots with the **topics** and **status** (active, temporally active or finished) are also included in Appendix I.

Finally, a data frame is created to see which Barcelonian **activities** have more projects participating, being it "Comunidad de práctica". In this case, the method to know if a project is related to education is checking if it belongs to the "Ciencia ciudadana en las escuelas" program, which connects different schools of Barcelona with scientists to collaborate on real research projects [15]. It has been tried to search for educative words such as ["educ", "escuela", "instituto", "profesor", "alumn"] in the projects' abstracts as with the European and Spanish platforms but for this case it is not a good option since almost any initiatives mention these words even though they are focused on schools.

3.4.4) Generating visualizations

For the Barcelonian initiatives, a Tableau dashboard has been done, showing the most relevant words for each project so that an idea of what they are about can be extracted. It also shows the projects distributed by status, by topics (most of them are related to environment or health), by if they are connected to the "Programa en las escuelas" or not, and finally by if they mention the need for a phone. From the 2 projects that are finished, any of them belonged to the Schools program and all of them are related to health.

(Images of the generated dashboard can be seen in Appendix I, and the link is available in Appendix B).

3.5 - Comparing the format of the platforms

Before comparing the current state of citizen science through what the platforms spread, it is interesting to compare the information they present and the way it is structured. As seen before, the European platform "EU-Citizen.Science" contains data about projects, resources, training resources and organisations. Moreover, it includes a blog, a section with the upcoming events and a forum to share knowledge. The Spanish "Observatorio de la Ciencia Ciudadana en España" also presents initiatives (which can be projects, institutions or people), resources (including training resources), the blog and the upcoming events section. Therefore, it offers the same content as the European plus CS interviews, articles and news. And despite being a local platform, the Barcelonian "Oficina de Ciencia Ciudadana" also offers projects, activities, news and upcoming events.

Regarding the projects, different fields are presented in each of the cases. The PPSR Core metadata standard mentioned before requires the following fields for a project description:

ID	Field	Data Type	Description
1	ProjectGUID	GUID	A Globally Unique Identifier (GUID) for each PPSR project
2	ProjectName	text	Name of project
3	ProjectDataProvider	text	Name of data provider / source / initial first registry
4	ProjectDescription	text	Description of project related to aspects such as goals, objectives, purpose, etc.
5	ProjectDateLastUpdated	ISO 8601 DateTime (UTC)	Date the project information was last updated; (required in DBs but not in XML)
6	DatabaseContactName	Person Object / Construct	Database contact first and last name
7	ProjectStatus	text / categorical	Current status of the project activity: Starting/ active/ complete/ hiatus

Figure 28. PPSR project fields.

The European platform follows this standard guide and describes the 34 fields shown in Appendix C, even though for some projects a few fields are empty. The Spanish platform displays the 24 fields shown in Appendix F, some of them differ from the European ones. Basically, the European platform includes the status, created date, more information about the project location, if it is doable at home and the origin database while the Spanish does not. And the Spanish includes data about the public, participants, results and impacts that the European platform does not show. In the Spanish case, some of the fields required by the PPSR Core metadata standard are not included (status, ProjectDataProvider and DatabaseContactName). Regarding the keywords, it has been verified that they are the same for the European and Spanish platforms, which means that they have been standardised.

Regarding the Barcelonian website, it contains much less information, since it just shows the name, description, status, topic, URL and related activities for each project.

An interesting analysis is to see if any project is repeated in some of the platforms, which would be logical since Barcelona is in Spain and Spain is in Europe, therefore the European platform should contain the projects of the other two, as well as the Spanish one should contain the Barcelona projects. However, these platforms may not be perfect, meaning that it is possible that they do not show all the projects that are being carried out in their respective territories. This is what aims to be answered with the Python script "Comparing_platforms".

With the necessary code, the names of the projects repeated in the European and Spanish platforms are shown. The "Mosquito Alert" project is printed two times since in the Spanish web there are two different entries for this project, and same with "Coact" and "FLOODUP". Therefore, it is seen that not all Spanish projects are included in "EU-Citizen.Science", but only some of them.

```
['MammalNet', 'RiuNet', 'WeCount', 'CoActuamos para la Salud Mental',
 'CoAct', 'Ocean Initiatives', 'Cos4Cloud', 'Identificación de asteroides
cercanos a la Tierra', 'Mosquito Alert', 'Mosquito Alert',
 'MelanogasterCTF', 'Melanogaster', 'FLOODUP', 'FLOODUP', 'Paddle Surfing
for Science on microplastic pollution', 'Cities at night']
```

The same happens with the Spanish and BCN platforms, where the following projects are repeated. Again, all projects in the Barcelonian platform should be included in the Spanish but are not, which shows that there is still work to be done.

```
['WeCount', 'Cities-Health', 'Observadores del Mar', 'Mosquito Alert',
'Saca La Lengua', 'RiuNet', 'Mosquito Alert']
```

Finally, when comparing the projects posted in the European and Barcelonian platforms the next projects are obtained:

```
['RiuNet', 'Genigma', 'WeCount', 'Mosquito Alert']
```

Therefore, the projects that appear in the three platforms are:

```
['RiuNet', 'WeCount', 'Mosquito Alert', 'Mosquito Alert']
```

3.6 - Comparing and analysing the content of the platforms

First of all, to get a general overview of citizen science initiatives' current state, all projects from the three platforms will be joined in the "Comparing_platforms" script in order to generate a global dashboard with the main projects' data. To do that, only some fields will be selected since some of them, as seen before, are not shown in the three platforms. However, most of them will be empty for the Barcelonian projects since those initiatives contain fewer fields.

For the projects appearing in more than one platform, they will be included only once (otherwise it would be replicated data), and the information used will be the one from the "EU-Citizen.Science" in case the project appears in that platform or from the

"Observatorio de la Ciencia Ciudadana en España" if it only appears in the Spanish and Barcelonian.

Once this data frame is computed and extracted, it is imported to Tableau in order to generate a general CS initiatives visualization. This dashboard shows different things. The first aspect indicates the number of initiatives that belong to each platform, being the Spanish the one with more content followed by the European one. Then, the needed equipment is shown, seeing that devices, apps and smartphones are the most typical required material. Regarding the topics, 290 out of the 468 initiatives belong to life sciences and 196 to applied sciences, and also 51% of the projects are connected to education.

Filtering by origin platform, this dashboard can be used to contrast the citizen science situation in the three regions through what their web pages offer. From the 176 projects of the European website, 70% of them belong to life sciences, 42% to social sciences, 30% of them to applied sciences and just 12% to physical sciences. Also, the most required material to participate are devices (phones or computers), followed by apps and cameras. Half of the European projects are related to education, and their durations mainly vary between 3 and 14 years. From the 279 Spanish initiatives (16 of them are already included in the European), also 60% of them belong to life sciences, but the next discipline is applied sciences (50%), followed by social (42%). Therefore, in both cases most projects belong to life sciences, and also have the same proportion of social initiatives (42%) and physical (12% vs 14%). However, in Spain there are more projects related to applied sciences than in Europe. The most required material is also devices, followed by apps and cameras, so this does not change. Also, 55% of the initiatives are related to education, a very similar percentage to the European one, and durations mainly vary between 7 and 34 years, so Spanish initiatives take more time than the European ones. From the 13 projects that are only uploaded to the Barcelonian platform, not much information can be extracted. However, it is seen that devices are again the most common needed material and life sciences is the most popular CS projects discipline, followed by applied and social sciences. So physical sciences is the scientific area with fewer citizen science initiatives in the three levels, European, Spanish and Barcelonian.

Regarding the initiatives' dates, it can be seen that from 2014 on the number of initiatives is increasing, since citizen science is more popular every year (2021 does not have that many starting projects since it is the current year and there has been no time to upload the projects recently created, plus the fact that with covid-19 it has been very difficult to generate new CS projects). When looking at the end dates, 48 initiatives ended in 2020 (probably because when uploading them they were already finished and the date was set to 2020), 15 in 2025, 14 in 2030 and 17 in 2050. This distribution through time is very similar in the European and Spanish cases, since for both platforms projects mainly started between 2014 and 2021 and lots of them end in 2020 or 2030. The main difference regarding dates is the fact that in Spain there are projects that end much later, which relates to the fact that Spanish initiatives have larger durations.

The same has been done with the 212 resources, joining all of them in the same dashboard that shows how many of them belong to each platform (62% to "EU-Citizen.Science" and 38% to the "Observatorio de la Ciencia Ciudadana en España"), their names and categories (most of them are guidelines, reports, websites or educational resources) and

if they are linked to a citizen science project or not (40% of the total are, while the resting 60% are not related to any initiative).

Again, filtering by origin platform, the styles of citizen science resources that both European and Spanish websites upload can be compared. From the 131 European uploaded resources, most of them are guidelines, websites, events, collections or videos. And only 14% are related to an existing citizen science project. In contrast, from the 81 Spanish resources, 82% of them are connected to a Spanish initiative, showing that in Spain resources are much more related to initiatives than in Europe. The Spanish resources are mainly educational, guidelines or reports. So the most common category for both cases is guidelines.

(Screenshots of the dashboards are attached in Appendix J, and the links to the dashboards are available in Appendix B).

4. WEB PAGE CREATION

The reason behind generating a web page[101] as a final product is to have all the results together in one space so that anyone interested can consult the webpage and get all the desired data, metrics and insights. So, this website will contain useful information about the current state of the citizen science European, Spanish and Barcelonian platforms.

As mentioned above, the Wix Website Editor has been the chosen tool to design this citizen science analysis website. The site is structured in six main pages:

1. HOME: this is the main page that presents this project, describing its aims and objectives. It also justifies the existence of the website and explains what content shows and how to navigate through it. Moreover, it contains two dashboards, one with the citizen science projects of all platforms joined together, so that a global analysis of CS initiatives current state can be performed, and another one that does the same with CS resources.

2. EU: this section presents the "EU-Citizen.Science" European platform, explains a little bit how it is structured and the data that it contains and then it shows several visualizations of the generated dashboards, for CS projects, resources and organisations. Each of these parts goes with a little interpretation of what can be observed and extracted, getting meaningful insights.

3. SP: it follows the same structure as the previous section but with the Spanish "Observatorio de la Ciencia Ciudadana" platform, having a little introduction followed by displays and analysis of the uploaded initiatives and resources.

4. BCN: this section presents and analyses the Barcelonian platform "Oficina de la Ciència Ciutadana", following a similar pattern to the European and Spanish websites, with the difference that this one contains less information.

5. CONCLUSIONS: the main conclusions of this work will be explained there.

6. ABOUT ME: in this last page I introduce myself so that anyone consulting the webpage can get to know me a little better.

Finally, in order to make sure that the website accomplishes the expected objectives, it will be presented to some experts. They will have access to the website and then complete a small form[102] containing questions about their satisfaction with the page and how interesting the project has been for them. By receiving this feedback from few experts of the CS Track project group, it will be possible to verify that the performed analyses and the generated metrics and dashboards are useful. Then, the idea is that just by consulting a webpage anybody can get interesting insights into how citizen science is spread through the European, Spanish and Barcelonian platforms. Therefore, thanks to this little form it will be seen if the webpage is understandable and if the overall result of the project accomplishes its expectations, doing novel analyses not seen before.

(Images of the different sections of the webpage are included in Appendix K).

[101] https://nvarasp.wixsite.com/cs-analysis
[102] Link to the form: https://forms.gle/KCftoRgwdhCyceYo6 .

4.1 - Feedback about the created website

As mentioned before, the created website has been tested with seven experts related to citizen science. One question asks for their relation to CS, five of them are researchers and/or leaders of CS projects, one works for a CS platform and another one in the area of communications of CS Track. The form (included in Appendix L) contains some quantitative (evaluating from 1 to 5) and some qualitative questions about the utility of the website and its content.

Regarding the webpage structure and facility to navigate through it, the mean answer is 3.25 points out of 5, which means that it could be improved so that it is even easier to navigate. In relation to its content, all the respondents found it understandable, being the general punctuation 4 out of 5. Also, all the experts have answered that the content was interesting for them, being all the answers between 4 and 5. This relates to the fact that all the seven experts have learnt some new knowledge. All except one of the respondents think that the website is useful and fulfils its purpose[103], and most of them have punctuated that they like it 4 out of 5. (Appendix L).

Concerning the suggestions for improvement proposed by the experts, they mostly suggest rethinking the highlighting and the format, unifying the text and the way the information is emphasized. They also coincide in giving more importance to the generated dashboards since it is what they consider the added value of the website, its interactive features. Also, two of them mentioned that few words of the dashboards are not in English.

All in all, the received feedback has been very meaningful, since from one side it is seen that the performed analysis is really interesting for the experts and provides new knowledge. The fact of showing data in an interactive way has been very appreciated, which means that the dashboards are useful. Therefore, one of the main goals of the work and the website is accomplished since people get interesting insights about citizen science. However, thanks to the experts' opinions it has been seen that the webpage can be reformatted in a way that makes it easier to read and process its content, since a lot of findings are shown. Thus, this will be done.

There were questions about the website structure and suggestions about its design, but the main idea behind this test was to confirm the utility of the generated insights and created dashboards, more than the website itself. Receiving feedback from experts in the citizen science field is a perfect way to evaluate this work. Therefore, the result is satisfactory since they have reported that the performed analyses and computed data are really interesting. "In my opinion, it is a very useful website that should be taken into consideration given its analytical nature, containing comparisons between citizen science platforms. Very good work."

[103] That just by consulting it anybody can get interesting insights into how citizen science is spread through the European, Spanish and Barcelonian platforms.

5. RESULTS AND CONCLUSIONS

This project intended to examine how citizen science is spread through its platforms, analysing which role CS currently has in society through what is exposed in the platforms and giving a special emphasis to education. Therefore, it studied the European, Spanish and Barcelonian citizen science platforms, analysing from one side their format and structures and from the other side their content. This section is separated into five parts:

PLATFORMS' STRUCTURE

It has been seen that the European and Spanish platforms have a similar structure, both presenting projects, resources, blogs and events calendars, while the Barcelonian includes projects and activities but in a lighter way. Regarding the websites' format, the most completed one is the European "EU-Citizen.Science", since it follows the standard metadata protocol and offers more fields, for projects and also resources. The Spanish initiatives also show several variables, but some important ones such as the projects' status are not included. Moreover, resources uploaded to the "Observatorio de la Ciencia Ciudadana en España" do not include specific fields, just the titles and their descriptions, while for the "EU-Citizen.Science" their information is more complete. This is also related to the fact that the European platform has an API which makes it really easy to extract the desired data. In contrast, for the Spanish and Barcelonian cases a web crawling script has been created since there was no other way to get the information (of course following the robots.txt restrictions). All in all, it has been proven that the bigger range a platform has, the more complete it is. This makes sense since the European Commission has more funds to invest in the citizen science European platform than the Spanish government has and clearly more than the Barcelonian city council.

PLATFORMS' CONTENT

Regarding the content, it has been verified that most citizen science initiatives are related to life sciences, mainly belonging to "Biodiversity" and "Ecology & Environment" topics, as the "The European citizen science landscape - a snapshot" [7] report already shown in 2018. Also, it has been seen that most CS projects and resources have been created from 2014 on, coinciding with the boom of citizen science in society. This is related to the fact that devices are becoming more and more required to participate in citizen science projects, as well as cameras and apps. Since the utility technology can have in this field is really big, making it really easy to collect and share data.

By checking if the uploaded resources are connected to citizen science projects or organisations, it has been confirmed that lots of resources come from CS institutions or have born thanks to specific projects. This means that almost half of the resources are addressed to specific citizen science purposes, however, the other half have been created to promote citizen science in a more global way.

Also, to give specific attention to educational aspects of citizen science, the number of projects related to education has been studied. The obtained result is that around 50% of citizen science initiatives have an educational motivation, impact or are directly connected with formal education. This is a very good metric since as mentioned before, by making students participate in scientific activities, they get a deeper understanding than just reading about a specific topic.

APPLIED TECHNIQUES

During this project, several analyses and techniques have been applied. First of all, using an API and creating two web crawling scripts for the platforms with no API was an essential step to get the content. Then, a lot of data pre-processing was necessary in order to clean the data and have it in a format that makes it easy to work with it. For example, converting dates that were written in Spanish into numerical data time format, unifying all the provinces names and looking for them, cleaning the names so that it can be seen if other initiatives mention them…

Most applied techniques were about text mining, since almost all the information posted on the platforms is text. That is why techniques such as tokenizing, stemming, cleaning text, getting the text length distribution or word clouds with the most repeated words have been very used to study some fields. Also, tf-idf has been used to extract the more relevant words for each initiative, so that it can be deduced what they are about. Entity name recognition has also been used to get the names of people belonging to the project team. An important part of the scripts consisted of classifying text fields into categorical variables, by studying which were the important elements of a field, creating lists with the corresponding words and looking for them in the text. Leading in this way to a classification of the most required material to participate in projects, for example. Moreover, in one case variables have been encoded with the purpose of computing correlations, which were not very accurate due to the nature of the data but gave interesting insights.

Finally, a big part of this project has been focused on creating interactive visualizations, by generating dashboards with Tableau. An infinity of graphics have been used, studying which were the most correct for each case and applying several features of Tableau in order to get the most useful visualizations.

CREATED WEBSITE

Finally, one objective of this work was to generate an output with the generated insights so that anyone interested in citizen science can get an overview of what this project has done and what is the citizen science current state through what the platforms show. The created website[104] achieves this goal since it presents all the visualizations so that anyone consulting the page can play with them and get the desired data about the platforms' content. It has been tested with some citizen science experts in order to verify that it fulfils the determined aims, which are approaching citizen science to all kinds of public, analysing the current situation of this field in a European, Spanish and Barcelonian level.

PERSONAL CONCLUSIONS

All in all, this project has accomplished its objectives, since it has extracted the data from the three platforms and analysed it, generating several visualizations and a useful webpage to consult any desired information. While carrying it out, I have consolidated knowledge learnt in the degree and also new scopes. First of all, I have learnt about the different ways to extract information from websites and moreover how to design my own web crawling script, something completely new for me. Moreover, I have studied and investigated several ways to analyse data, since some techniques applied were already seen but others were totally novel. So, looking at the data you have and thinking about how to process it in order to get its maximum potential is one of the key aspects of being

[104] https://nvarasp.wixsite.com/cs-analysis

a data scientist, and I think this project has helped me to improve on that. Moreover, generating the different dashboards has provided me with a lot of confidence and experience using Tableau, which is one of the most important tools for data analysts.

Of course, during the whole project I have learnt a lot about citizen science, which was a completely new field for me, but I think it has been very interesting to use data science to take profit and extract meaningful insights from this topic, since at the end it shows in which aspects citizens are more willing to help or participate. And finally, the fact of doing a quite long and important project as this is one provides different learnings by itself, such as managing situations in which not everything goes as expected and impediments arise.

6. FURTHER WORK AND IMPROVEMENTS

As further work, different lines have been considered:

- Including more citizen science platforms would allow extending the analysis to a wider range, and therefore compare the current state of citizen science in more territories. One example could be the "Federal Crowdsourcing and Citizen Science Catalog"[105] of the United States government. However, for the purpose of this thesis it was not relevant since it aimed at analysing the citizen science situation in our territories: Europe, Spain and Barcelona.

- One idea that came out while doing this project was to study repositories of educational resources to get educational resources related to citizen science. However, this was not possible since all the consulted platforms (OER Commons, Teaching Commons, BC Campus, MIT OpenCourseware, Merlot.org) had very few resources related to citizen science (26 at most). It is true that when searching for "Citizen science" in the search engines, hundreds of results were shown. Nevertheless, when analysing those resources, it was seen that those resources contained the words "citizen" or "science" separately, and they did not have any relation with citizen science. Therefore, as there were not enough resources really connected to citizen science, it was decided to not continue with this line of work and focus more on the three platforms' content.

- As mentioned before, if the amount of data were bigger, it would be interesting to apply unsupervised learning algorithms such as clustering to group the CS projects and resources into different styles and see which characteristics each group has.

- Proposing a standardization of citizen science platforms so that all of them contain the same information would be a very interesting approach that would make it easier to compare and contrast projects or resources, since all of them would contain the same attributes. This is what the PPSR_CORE Program Data Model Metadata Standard aims for.

- Finally, presenting this project to local institutions and schools would promote citizen science and approach it to a wider audience, since nowadays still most citizens are not aware of the existence of this science. As mentioned during the whole project, citizen science can play a key role in education because the closer students are to a scientific topic, the more implicated they become and the more knowledge they get.

[105] https://www.citizenscience.gov/catalog/#

7. BIBLIOGRAPHY

1. Chari, R., Blumenthal, M., & Matthews, L. (2019). Community Citizen Science: From Promise to Action. RAND Corporation. https://doi.org/10.7249/RR2763
2. National Geographic Society. (2009). National Geographic Society. Retrieved January 25, 2021, from https://www.nationalgeographic.org/encyclopedia/citizen-science/
3. CS Track. (2020). CS Track. Retrieved January 27, 2021 from https://cstrack.eu/
4. Serrano Sanz, F., Holocher-Ertl, T., Kieslinger, B., Sanz García, F., & G. Silva, C. (2014). WHITE PAPER on Citizen Science for Europe. Socientize: Citizen Science Projects.
5. Zooniverse, T. (2009). The Zooniverse is Go. Zooniverse. Retrieved February 7, 2021, from https://blog.zooniverse.org/2009/12/12/the-zooniverse-is-go/
6. Robinson, L. D., Cawthray, J. L., West, S. E., Bonn, A., & Ansine, J. (2019). Ten principles of citizen science. In Citizen Science (pp. 27–40). UCL Press. https://doi.org/10.2307/j.ctv550cf2.9
7. Hecker, S., Garbe, L., & Bonn, A. (2018). The European citizen science landscape – a snapshot. In Hecker S., Bonn A., Haklay M., Bowser A., Makuch Z., & Vogel J. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy (pp. 190-200). London: UCL Press. Retrieved February 10, 2021, from http://www.jstor.org/stable/j.ctv550cf2.20
8. Ferran-Ferrer, N. (2015). Volunteer participation in citizen science projects. Profesional De La Información, 24(6), 827-837. https://doi.org/10.3145/epi.2015.nov.15
9. European Citizen Science Association. (2020). ECSA's characteristics of citizen science.
10. Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (n.d.). The Science of Citizen Science.
11. Sauermanna, H., & Franzonib, C. (2015). Crowd science user contribution patterns and their implications. Proceedings of the National Academy of Sciences of the United States of America, 112(3), 679–684. https://doi.org/10.1073/pnas.1408907112
12. Observatorio de la Ciencia Ciudadana en España. (2020). Informe del Observatorio de la Ciencia Ciudadana 2019/2020.
13. Storksdieck, M., Shirk, J. L., Cappadonna, J. L., Domroese, M., Göbel, C., Haklay, M., Miller-Rushing, A. J., Roetman, P., Sbrocchi, C., & Vohland, K. (2016). Associations for Citizen Science: Regional Knowledge, Global Collaboration. Citizen Science: Theory and Practice, 1(2), 10. https://doi.org/10.5334/cstp.55
14. Wagenknecht, K., Woods, T., Sanz, F. G., Gold, M., Bowser, A., Rüfenacht, S., Ceccaroni, L., & Piera, J. (2021). EU-Citizen.Science: A Platform for Mainstreaming Citizen Science and Open Science in Europe. Data Intelligence, 3(1). https://doi.org/10.1162/dint_a_00085
15. Unit, S. C. (2013). Science for Environment Policy In-Depth Report: Environmental Citizen Science. Report Produced for the European Commission DG Environment, (9), 32. Retrieved February 17, 2021, from http://ec.europa.eu/science-environment-policy
16. European Citizen Science Association (ECSA) – Engage with us. (s. f.). European Citizen Science Association. Retrieved February 17, 2021, from https://ecsa.citizen-science.net/
17. Law, E., Williams, A. C., Shirk, J., Wiggins, A., Brier, J., Preece, J., & Newman, G. (2017). The science of citizen science: Theories, methodologies and platforms. CSCW 2017 - Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 395–400. https://doi.org/10.1145/3022198.3022652
18. Mitchell, N., Triska, M., Liberatore, A., Ashcroft, L., Weatherill, R., & Longnecker, N. (2017). Benefits and challenges of incorporating citizen science into university education. PLoS ONE, 12(11). https://doi.org/10.1371/journal.pone.0186285

19. Jenkins, L. L. (2011). Using citizen science beyond teaching science content: A strategy for making science relevant to students' lives. Cultural Studies of Science Education, 6(2), 501–508. https://doi.org/10.1007/s11422-010-9304-4
20. Shah, H. R., & Martinez, L. R. (2016). Current Approaches in Implementing Citizen Science in the Classroom. Journal of Microbiology & Biology Education, 17(1), 17–22. https://doi.org/10.1128/jmbe.v17i1.1032
21. Ciència Ciutadana | Barcelona Ciència. (n.d.). Retrieved March 9, 2021, from https://www.barcelona.cat/barcelonaciencia/ca/ciencia-ciutadana
22. Bowser, A., Brenton, P., Stevenson, R., Newman, G., Schade, S., Bastin, L., Parker, A. and Oliver, J. (2017). Citizen Science Association Data and Metadata Working Group: Report from CSA 2017 and Future Outlook. Woodrow Wilson International Center for Scholars. https://www.wilsoncenter.org/sites/default/files/media/documents/article/wilson_171204_meta_data_f2.pdf
23. PPSR_CORE Metadata Standard. (n.d.). Retrieved April 3, 2021, from https://citizenscience.org/2015/10/09/ppsr_core-metadata-standard/
24. S.C.M. de S Sirisuriya. (2015). A Comparative Study on Web Scraping. 8th International Research Conference KDU, (November), 135–140. Retrieved April 3, 2021, from http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf?sequence=1&isAllowed=y
25. Saurkar, A. v, & Gode, S. A. (2018). An Overview On Web Scraping Techniques And Tools. In International Journal on Future Revolution in Computer Science & Communication Engineering. Retrieved April 3, 2021, from http://www.ijfrcsce.org.
26. Zhao, B. (2017). Web Scraping. In Encyclopedia of Big Data. https://doi.org/10.1007/978-3-319-32001-4_483-1
27. Application Programming Interface (API). (n.d.). Retrieved March 23, 2021, from https://www.ibm.com/cloud/learn/api
28. Carl Drott, M. (2002). Indexing aids at corporate websites: The use of robots.txt and META tags. Information Processing and Management, 38(2), 209–219. https://doi.org/10.1016/S0306-4573(01)00039-5
29. Python Requests - Make a request. (n.d.). Retrieved March 23, 2021, from https://docs.python-requests.org/en/master/user/quickstart/#make-a-request
30. Welcome to Python.org. Python.org. (n.d.). Retrieved March 23, 2021, from https://www.python.org/ .
31. Natural Language Toolkit — NLTK 3.6.2 documentation. (n.d.). Natural Language Toolkit. Retrieved June 14, 2021, from https://www.nltk.org/
32. What is NumPy?. What is NumPy? - NumPy v1.20 Manual. (n.d.). Retrieved March 23, 2021, from https://numpy.org/doc/stable/user/whatisnumpy.html .
33. Pandas (n.d.) Retrieved March 23, 2021, from https://pandas.pydata.org/ .
34. Matplotlib (n.d.) Retrieved March 23, 2021, from https://matplotlib.org/ .
35. Seaborn (n.d.) Retrieved March 23, 2021, from https://seaborn.pydata.org/ .
36. Documentation (2021). Retrieved March 23, 2021, from https://docs.python.org/ .
37. Python Package Index (2021) Retrieved March 23, 2021, from https://pypi.org/
38. Project Jupyter. (2021). Retrieved March 23, 2021, from https://jupyter.org/
39. Introducción a JSON. (n.d.) Retrieved March 25, 2021, from https://www.json.org/json-es.html
40. Selenium (n.d.) Retrieved March 25, 2021, from https://www.selenium.dev/
41. Tableau (n.d.) Retrieved March 25, 2021, from https://www.tableau.com/es-es
42. Wix (n.d.) Retrieved March 25, 2021, from https://www.wix.com/about/us
43. Github Guides (n.d.) Retrieved March 25, 2021, from https://guides.github.com/

44. Snowball Stemmer – NLP. (n.d.). Retrieved May 23, 2021, from https://www.geeksforgeeks.org/snowball-stemmer-nlp/
45. Salton G; McGill MJ (1986). Introduction to modern information retrieval. McGraw-Hill. ISBN 0-07-054484-0.
46. Chunking. (n.d.). Retrieved May 27, 2021, from https://www.nltk.org/howto/chunk.html

8. APPENDICES

8.1) Appendix A

The CS Track Project must be mentioned in this document since the directors of this thesis belong to the TIDE-UPF[106], which is a partner of the CS Track Project. That is why they proposed some thesis topics that are framed in the context of the objectives pursued by CS-Track. Mainly, this work will use computational methods to learn more about Citizen Science, analysing the content of three platforms to get insights about the CS current situation at different levels and study the connection between CS and education.

CS Track[107] is a project that aims to expand the knowledge of Citizen Science and maximize the potential benefit of Citizen Science activities on individual citizens, organizations, and society in general. "CS track investigates what activities are called citizen science and by whom? What are their societal, economic, educational and scientific impacts? Who participates in such projects? How and why? What are the (dis)incentives and rewards for all participants? What enables citizen science, what creates barriers in citizen science, and what are its limits?" [23].

There already exists information regarding these questions. However, it is mostly limited to specific initiatives. That is why CS Track analyses a wider range of projects, narrowing the knowledge gap by conducting research on these issues and others across a broad range of projects across the European Union. The main goals of CS Track are:

- Investigate a large number of Citizen Science activities: observing and analysing a large number of different citizen science activities to increase knowledge about what makes them successful in response to a given set of research questions.
- Provide useful tools to measure and analyse Citizen Science activities: providing tools and frameworks to analyse citizen science activities in different settings that can be used by the people interested.
- Share information, analysis and tools: sharing knowledge and resources it creates, together with recommendations to the citizen science community.

In the following *Figure 29* it can be seen the main outputs the project has with the corresponding dates.

[106] TIDE-UPF group website (2021, January 27) https://www.upf.edu/web/tide .
[107] CS Track website. (2021, January 27) https://cstrack.eu/

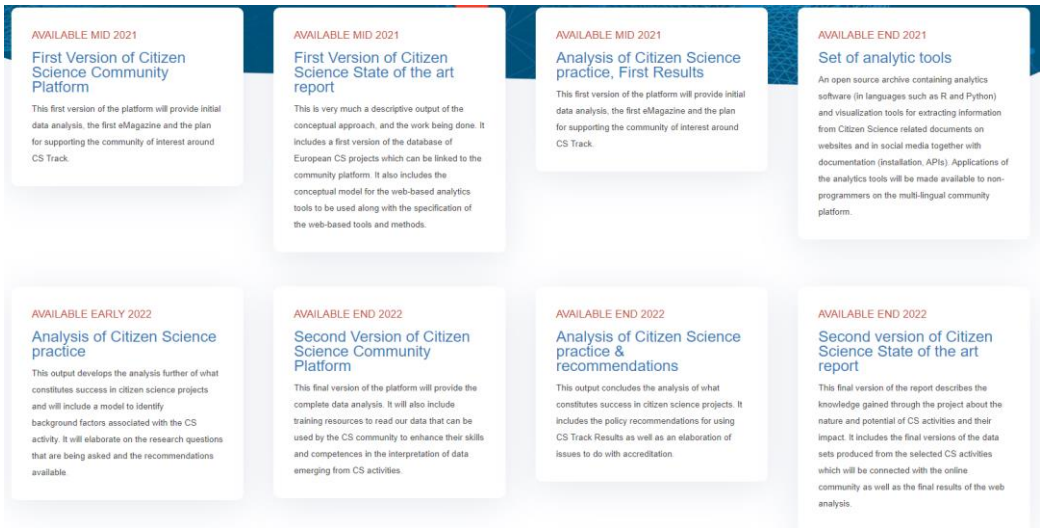


Figure 29. CS Track project main outputs timeline [3].

8.2) Appendix B

Location of the scripts:

Script name	Link
scraping_data.ipynb	https://github.com/nvarasp/TFG_CS_Analysis/blob/main/scripts/scraping/scraping_data.ipynb
functions_t.ipnyb	https://github.com/nvarasp/TFG_CS_Analysis/blob/main/scripts/analysing/functions_t.ipynb
Comparing_platforms.ipynb	https://github.com/nvarasp/TFG_CS_Analysis/blob/main/scripts/analysing/Comparing_platforms.ipynb
European_analysis	https://github.com/nvarasp/TFG_CS_Analysis/blob/main/scripts/analysing/European_analysis.ipynb
Spanish_analysis	https://github.com/nvarasp/TFG_CS_Analysis/blob/main/scripts/analysing/Spanish_analysis.ipynb
Barcelonian_analysis	https://github.com/nvarasp/TFG_CS_Analysis/blob/main/scripts/analysing/Barcelonian_analysis.ipynb

Figure 30. Location of the scripts.

Links to the dashboards:

EUROPE	Projects	General	https://public.tableau.com/app/profile/n.ria4479/viz/Mapfilters_EU_projects/Generalfilters
		Map	https://public.tableau.com/app/profile/n.ria4479/viz/Maptime_EU_projects/map_time links to map topics and map filters
		Words	https://public.tableau.com/app/profile/n.ria4479/viz/Words_EU_projects/GeneralWords
		Dates	https://public.tableau.com/app/profile/n.ria4479/viz/Mapfilters_EU_projects/Dates
	Resources and training resources	Words	https://public.tableau.com/app/profile/n.ria4479/viz/Mixresources/Abstract
		Dates	https://public.tableau.com/app/profile/n.ria4479/viz/Mixresources/dates
		General	https://public.tableau.com/app/profile/n.ria4479/viz/Mixresources/Audiences
	Organisations	https://public.tableau.com/app/profile/n.ria4479/viz/organisationsmap/Organisations	
SPAIN	Initiatives	General	https://public.tableau.com/app/profile/n.ria4479/viz/spanishglobal/global_initiatives
		Map	https://public.tableau.com/app/profile/n.ria4479/viz/spanishglobal/map
		Words	https://public.tableau.com/app/profile/n.ria4479/viz/spanishglobal/words
		Dates	https://public.tableau.com/app/profile/n.ria4479/viz/spanishglobal/Dates
	Resources	https://public.tableau.com/app/profile/n.ria4479/viz/SPresourcestypes/SPresourceswords	
BCN	Projects	https://public.tableau.com/app/profile/n.ria4479/viz/bcnprojects/bcnprojects	
MIXED	Initiatives	https://public.tableau.com/app/profile/n.ria4479/viz/MIX_projects/Dashboard1	
	Resources	https://public.tableau.com/app/profile/n.ria4479/viz/CSresources/Dashboard1	

Figure 31. Location of the dashboards.

8.3) Appendix C

European project fields, divided by shown in the platform or not. From the visible fields, all have been used except the URL. And from the non-visible, the table shows the division of used and non-used.

Visible fields.	Project id	Name	Aim	Description
	Keywords	Status	Topic	URL
	Start_date	End_date	country	host
	howToParticipate	equipment		
Non-visible fields, but used.	mainOrganisation	organisation	latitude	longitude
	dateCreated	doingAtHome	fundingBody	fundingProgram
	featured			
Non-visible and non-used fields.	origin	originDatabase	originURL	originUID
	imageCredit3	image1	imageCredit1	image2
	imageCredit2	image3	customField	

Figure 32. European projects' fields.

8.4) Appendix D

For descriptions:

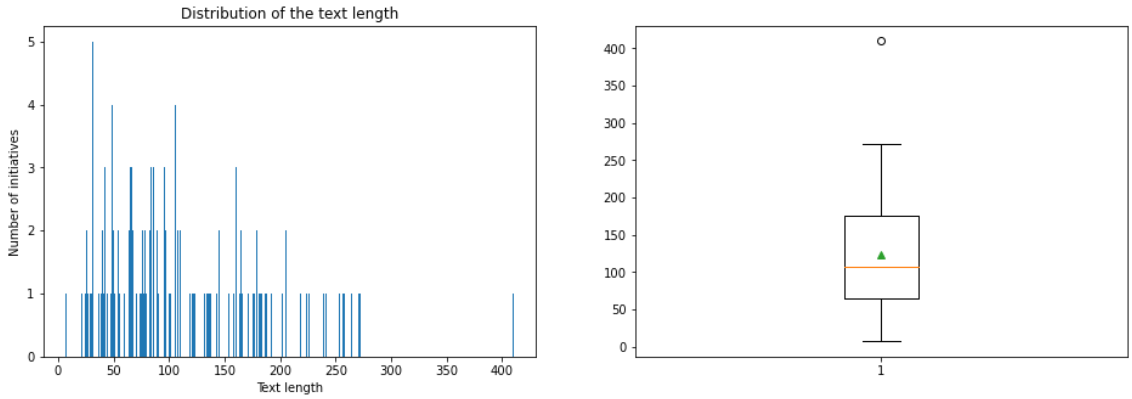


Figure 33. European projects, description field plots.

For aim field:

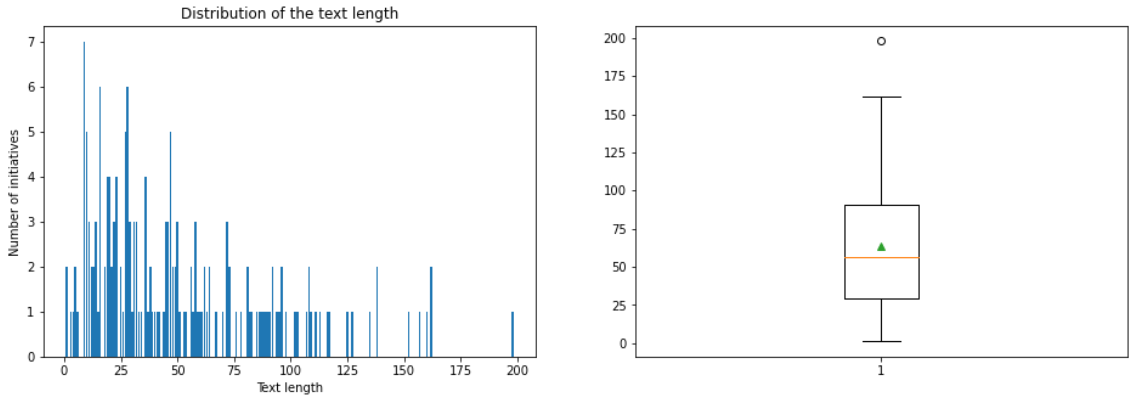


Figure 34.European projects, aim field plots.

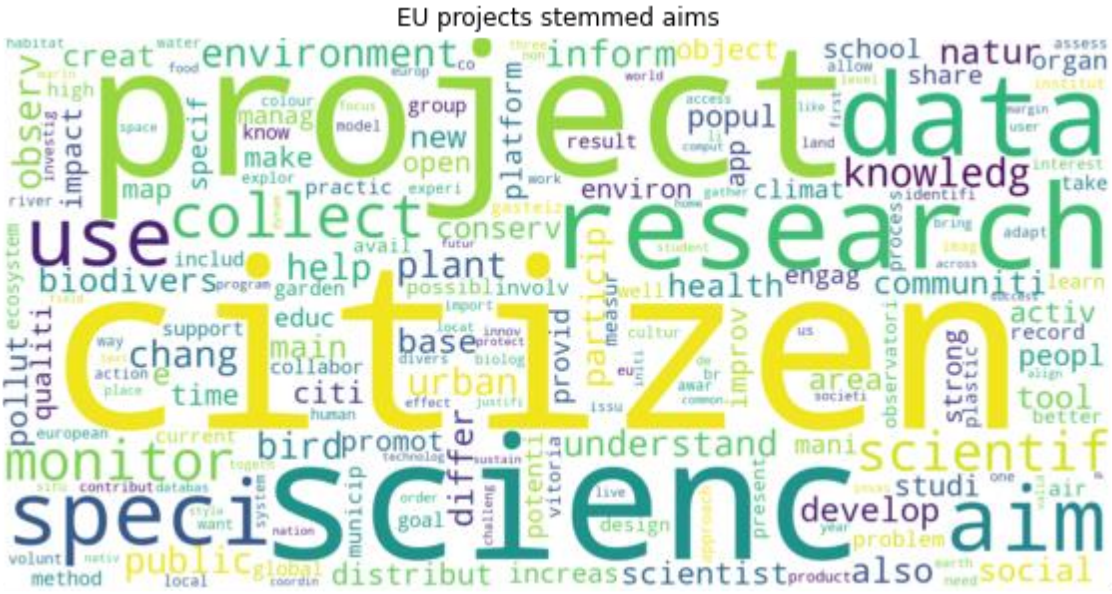


Figure 35.European projects, aim word cloud.

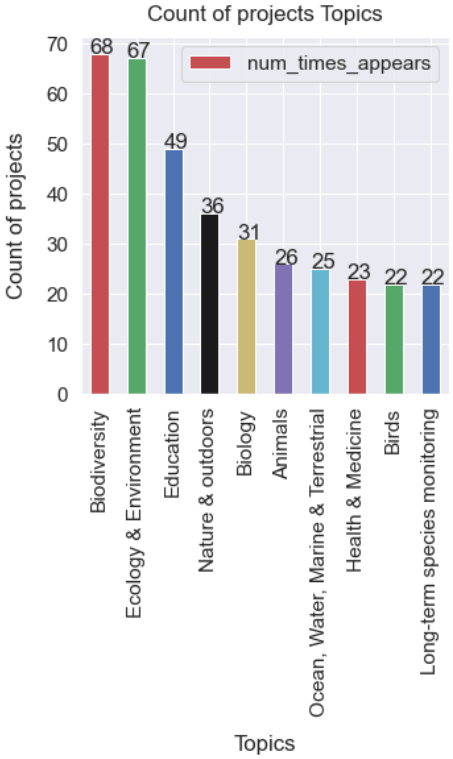


Figure 36. European projects, topics plot.

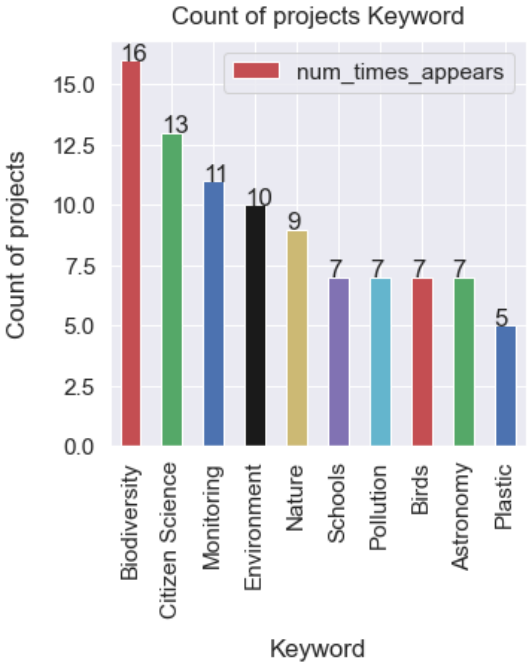
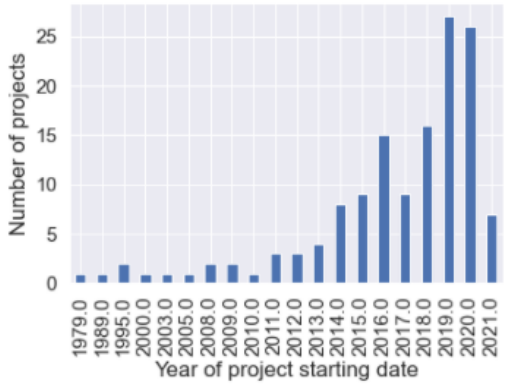


Figure 37.European projects, keywords plot.




```
EU_projects['duration_years'].describe()
```

```
count    80.000000
mean      3.250000
std       4.320982
min       0.000000
25%       0.000000
50%       2.000000
75%       3.250000
max      17.000000
Name: duration_years, dtype: float64
```

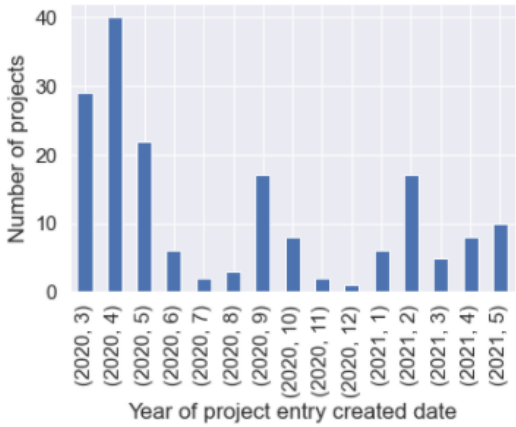


Figure 38.European projects, timelines.

List of the most common hosts:

1. VA (Public & Science): 9 times
2. Environmental Research Center of Vitoria-Gasteiz city council: 8 times
3. Leiden University: 4 times
4. Natural History Museum: 3 times
5. Ibercivis Foundation: 2 times

The **main organisations** that appear more times are:

	id	dateCreated	name	description	latitude	longitude	country	creator	type	num_times_appears	list_projects
25	2	2021-03-12T16:14:44.312362Z	Ibercivis Foundation	\r\nIBERCIVIS is the national private non-for-...	41.684275	-0.887436	ES	1	None	4	[26, 7, 6, 4]
7	11	2021-04-09T11:03:03.325507Z	Maremma Natural History Museum	Among the precursors of citizen science in Ita...	42.760494	11.116094	IT	19	None	4	[247, 251, 250, 258]
18	50	2020-10-07T09:47:24.539706Z	Surfrider Foundation Europe	Surfrider Foundation Europe is a non-profit or...	43.459946	-1.540784	FR	1390	None	3	[172, 178, 242]
31	19	2020-09-08T09:10:48.338867Z	Leiden University	Leiden University is one of Europe's leading i...	52.156635	4.486740	NL	45	None	2	[125, 135]
19	73	2021-02-24T17:23:21.652503Z	Museum für Naturkunde Berlin	The Museum für Naturkunde Berlin – Leibniz In...	52.530490	13.379115	DE	946	None	2	[19, 21]

Figure 39. European projects' main organisations.

The most common **organisations** are:

	id	dateCreated	name	description	latitude	longitude	country	creator	orgType	num_times_appears	list_projects
9	1	2021-05-06T07:47:15.622051Z	European Citizen Science Association (ECSA)	Since 2013, ECSA has evolved from an informal ...	52.530490	13.379115	DE	3	2	5	[180, 237, 21, 18, 74]
0	6	2020-09-08T06:48:03.452860Z	Vetenskap & Allmänhet (VA – Public & Science)	VA (Public & Science) is a Swedish non-profit ...	59.335830	18.075792	SE	24	2	4	[185, 226, 180, 227]
14	2	2021-03-12T16:14:44.312362Z	Ibercivis Foundation	\r\nIBERCIVIS is the national private non-for-...	41.684275	-0.887436	ES	1	2	3	[237, 21, 145]
12	60	2020-11-04T06:35:04.459879Z	Universitat de Barcelona	The University of Barcelona (UB) is a public u...	41.384362	2.119789	ES	2316	3	2	[227, 235]
31	116	2021-02-15T13:26:01.639736Z	Universitat Autonoma de Barcelona	The Universitat Autònoma de Barcelona (UAB), l...	41.757870	2.031182	ES	2545	3	1	[254]

Figure 40. European projects' organisations.

Figure 41. European resources' abstract word cloud.

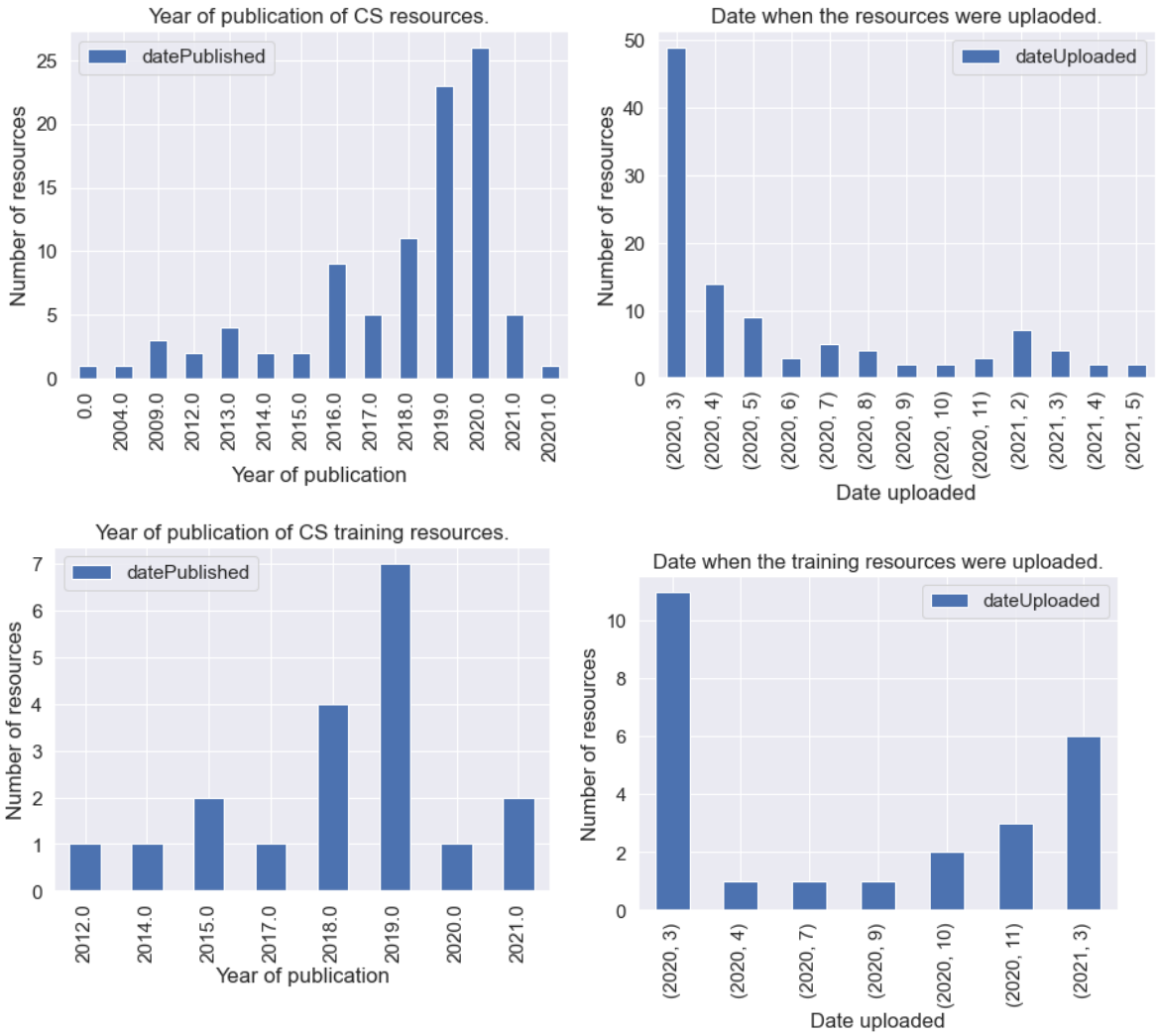
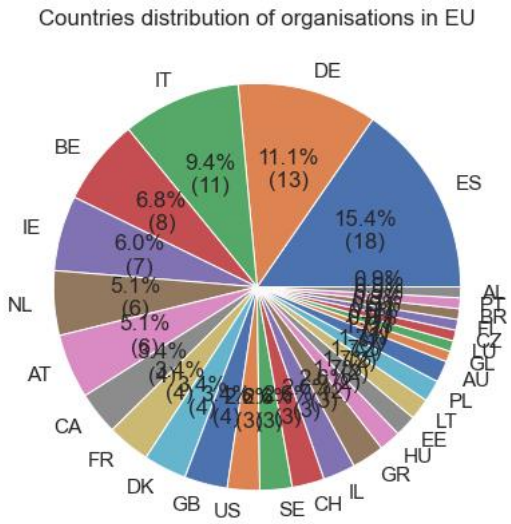


Figure 42.European resources' timelines.



orgType	num_times_appears
Academic	52
Non-governmental	30
Private sector	14
Governmental	13
Community-led	4
Consortium	4

Figure 43. Organisations' types.

Figure 44. Countries distribution of organisations in Europe.

8.5) Appendix E

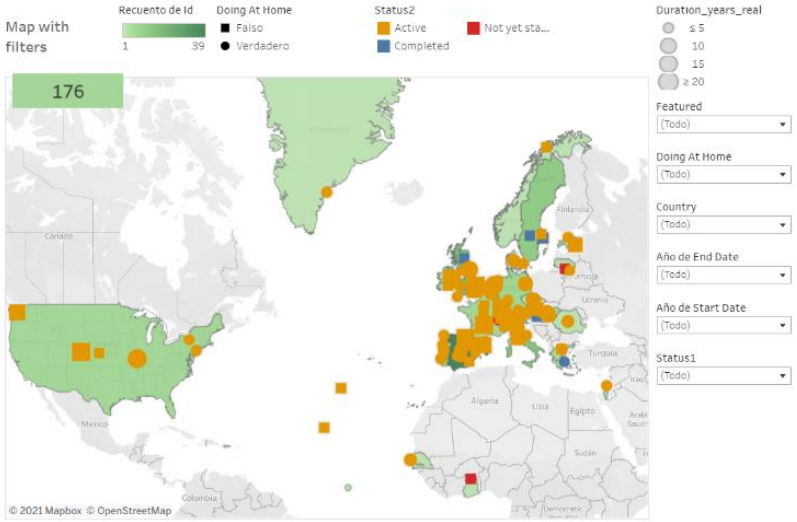


Figure 45. Dashboard of European projects distributed in a map.

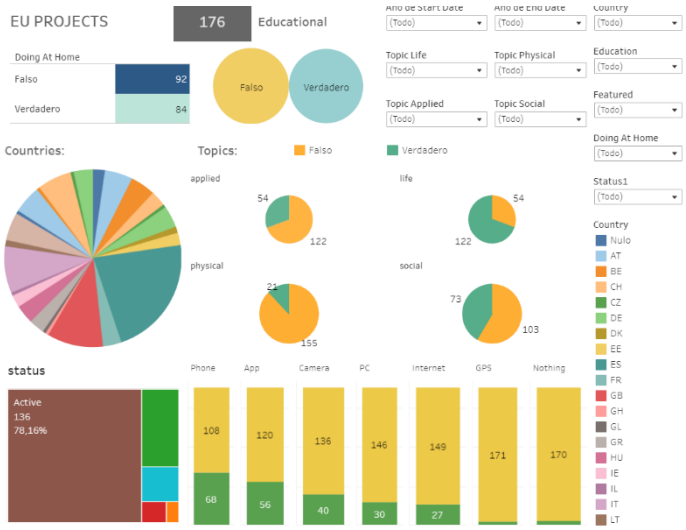


Figure 46.Dashboard showing main features of European projects.

Figure 47.Dashboard showing most relevant words of European projects.

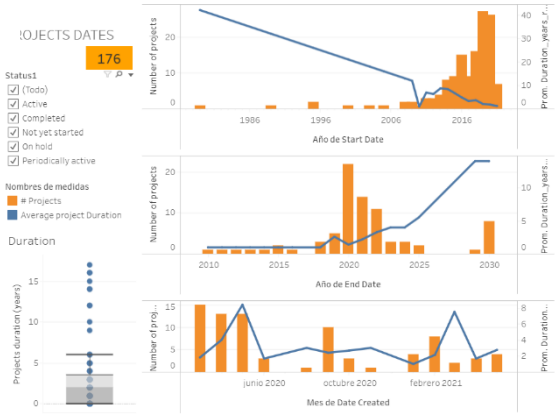


Figure 48.Dashboard with European projects' dates.

EUROPEAN RESOURCES DASHBOARDS

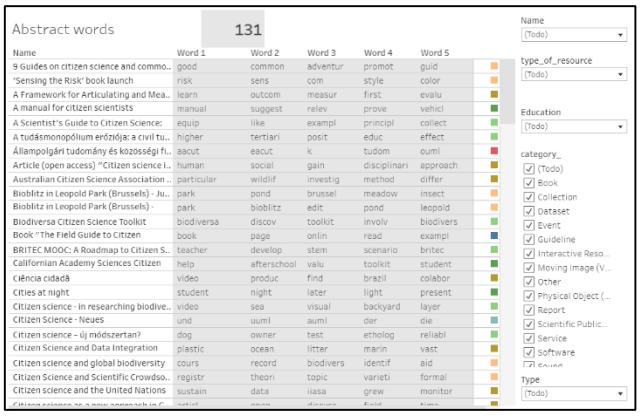


Figure 51.Dashboard with European resources' words.

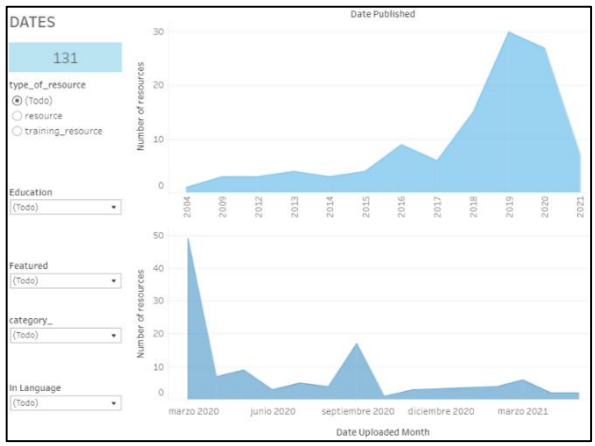


Figure 50.Dashboard with European resources' dates.

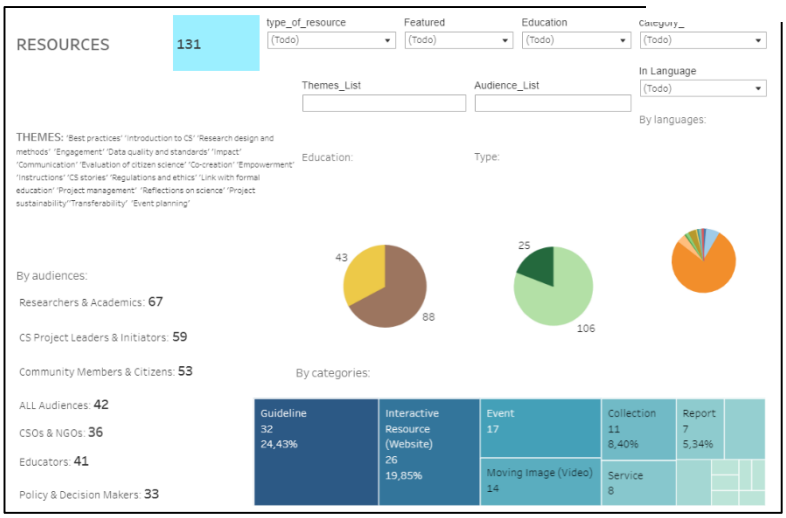


Figure 49.Main dashboard of European resources.

ORGANISATIONS DASHBOARDS

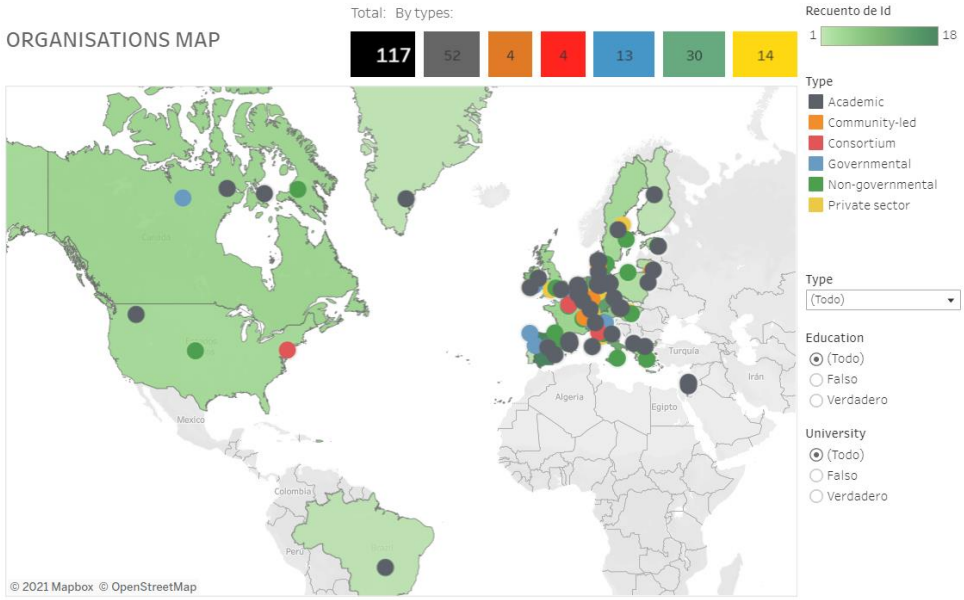


Figure 52.Dashboard of European organisations distribution.

8.6) Appendix F

Spanish initiatives' fields:

project id	name	mainOrganisation	subtitle	type
keywords	start_date	end_date	public	province
participants	URL	aim	description	responsibleEntity
foundingTeam	moreEntities	howToParticipate	equipment	results_summary
results_link	impact	impact_examples	motivation	

Figure 53. Spanish initiatives' fields.

8.7) Appendix G

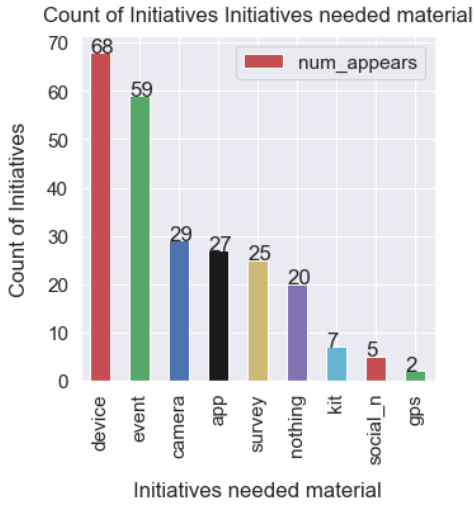


Figure 55. Needed material of Spanish initiatives.

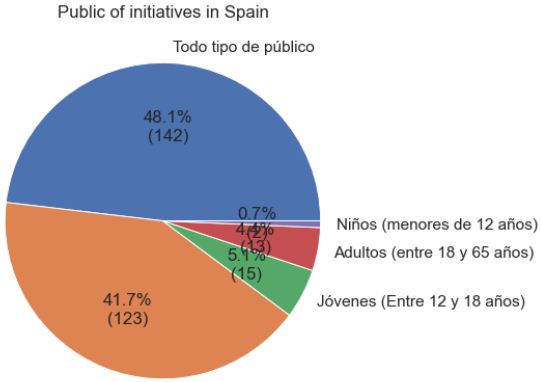


Figure 54. Public of initiatives in Spain.

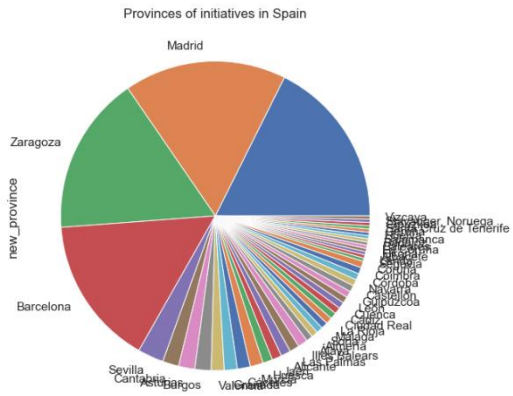


Figure 56.Provinces of Spanish initiatives.

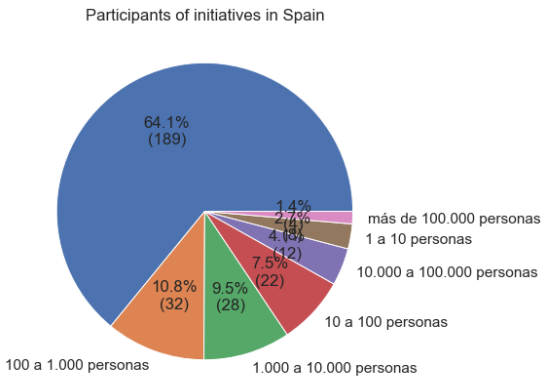


Figure 57. Participants of initiatives in Spain.

name	num_appears
en_proceso	20
informe	7
herramienta	7
inicio	5
mapa	5

Figure 58.Results summary table.

8.8) Appendix H

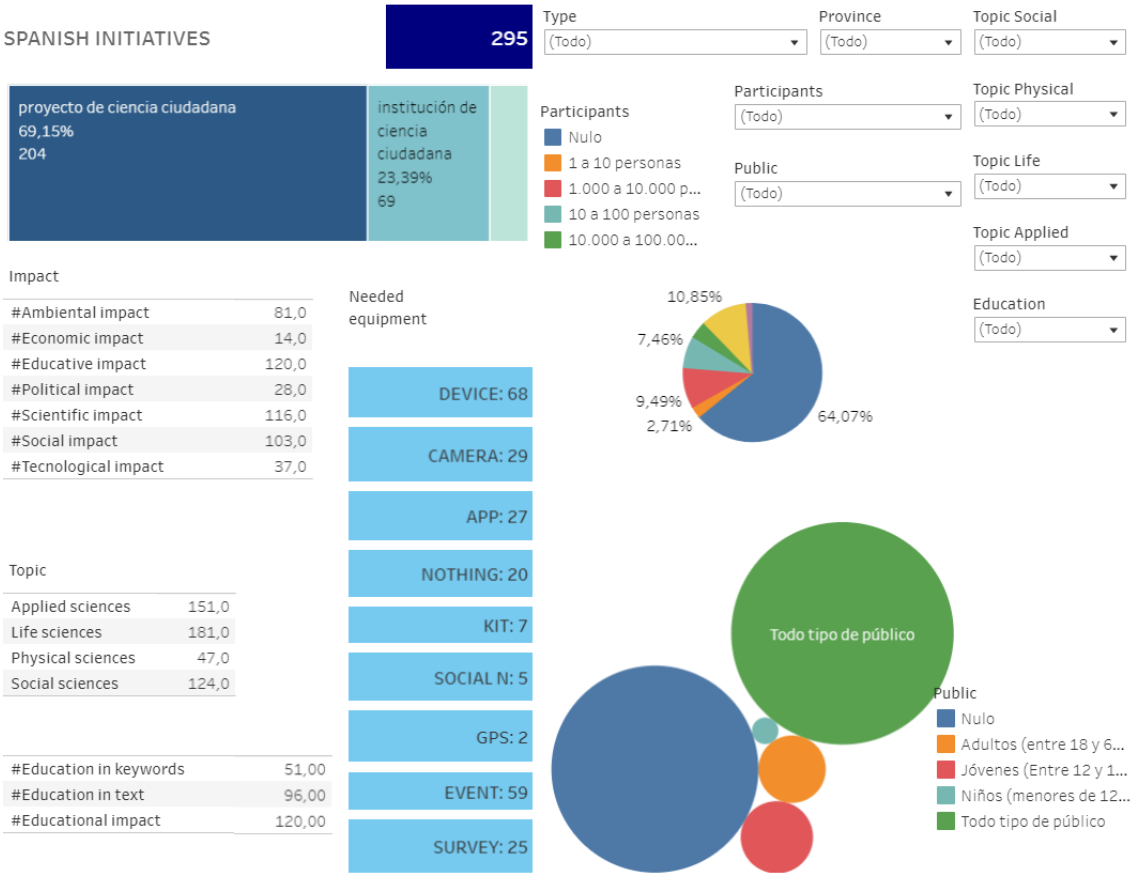


Figure 59.Main dashboard of Spanish initiatives.

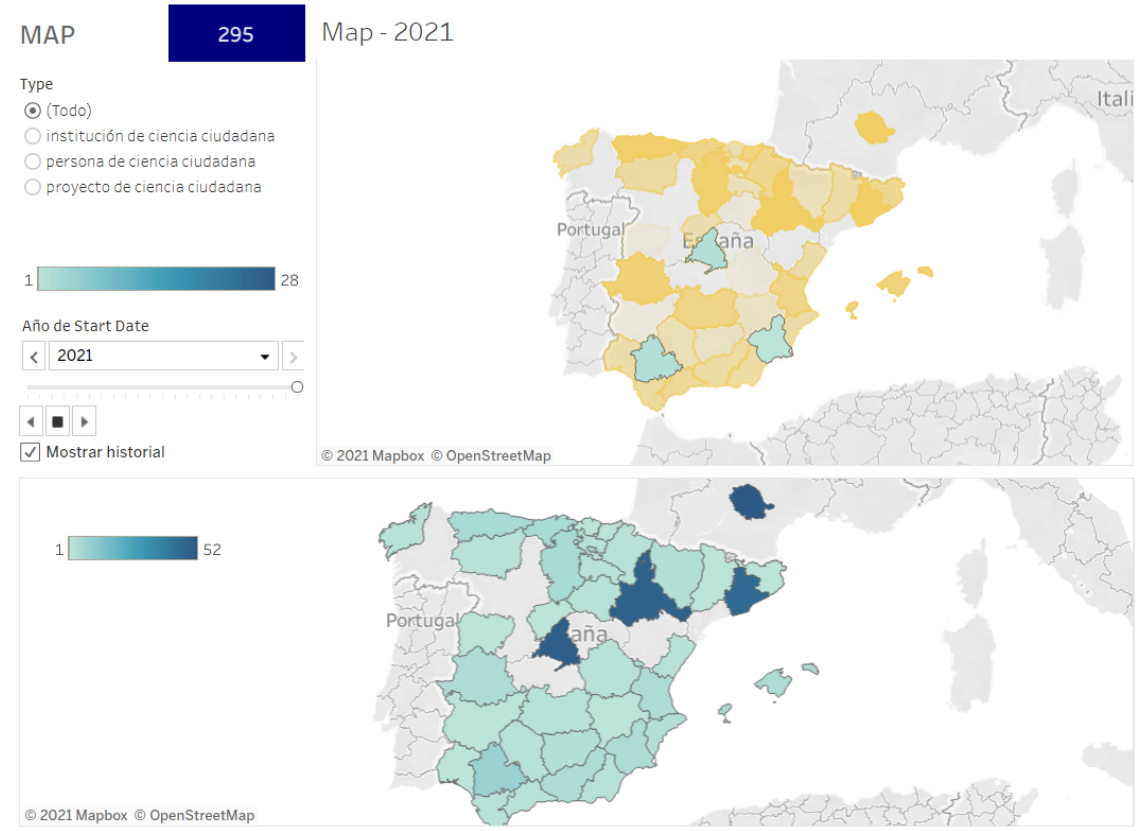


Figure 60.Spanish initiatives distributed in a map.

Name	Word 1	Word 2	Word 3	Word 1 d	Word 2 d	Word 3 d
«FLOODUP. Explorando ..	inund	observ	floodup	episodi	impact	permit
«Picture your microbes	viv	obes	microb	microbiot	particip	discusion
&DIGITALFAB				pol	representacion	sensorial
#FakeNurses	enfermeri	profesional	mujer	enfermeri	fakenurs	enfermer
AcercaCiencia	acercacient	com	category	dam	difusion	cienci
AEV – Centro de Ciência ..	cual	ciudadan	cinc	may	ener	fas
Aficca	aficc	demostr	institu			
AGEO	atlant	age	riesg	riesg	observatori	geolog
Agrupación				aam	astronom	telescopi
Alerta Forestal				salud	societat	bosc
Algoritmo Arturo	ajust	artur	ayudan	vam	nombr	artur
Algoritmo Mercè	merc	utilizacion	algoritm	habit	urban	agend
Alimentus	agronom	alimentacion	alimentus	nutricion	alimentacion	ambit
ALMUDENA DE LA ENCA..	encant	particul	quier			
AluCIENCIAnante				aluciencian	cientif	cultur
AMBER	atlas	barrer	geolocaliza..	barrer	fluvial	atlas
AmBioBlitz UAL	campus	inventari	ual			
Amigos de la Tierra				amig	tierr	local

Figure 61.Dashboard with words of Spanish initiatives.

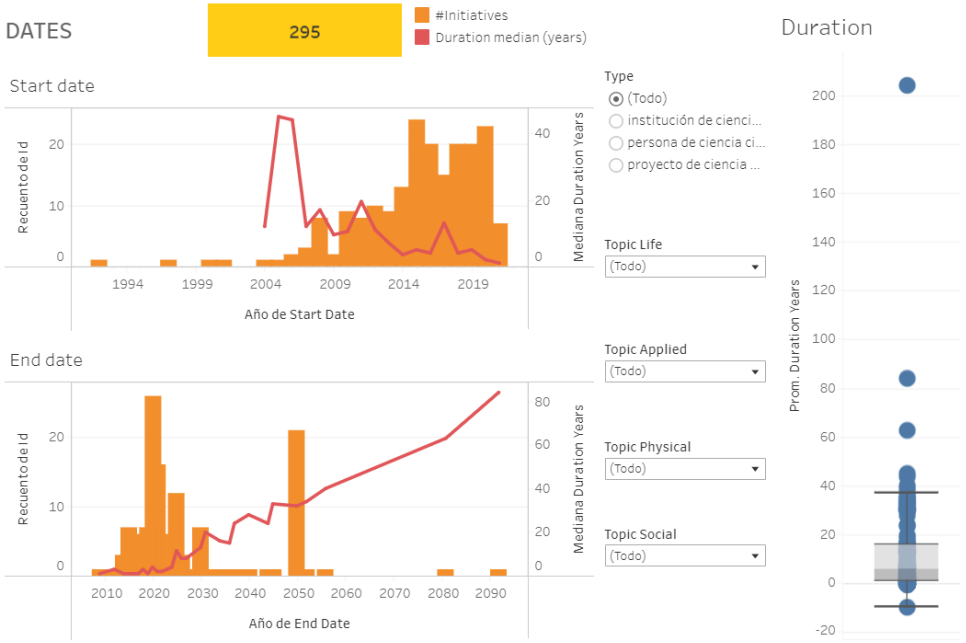
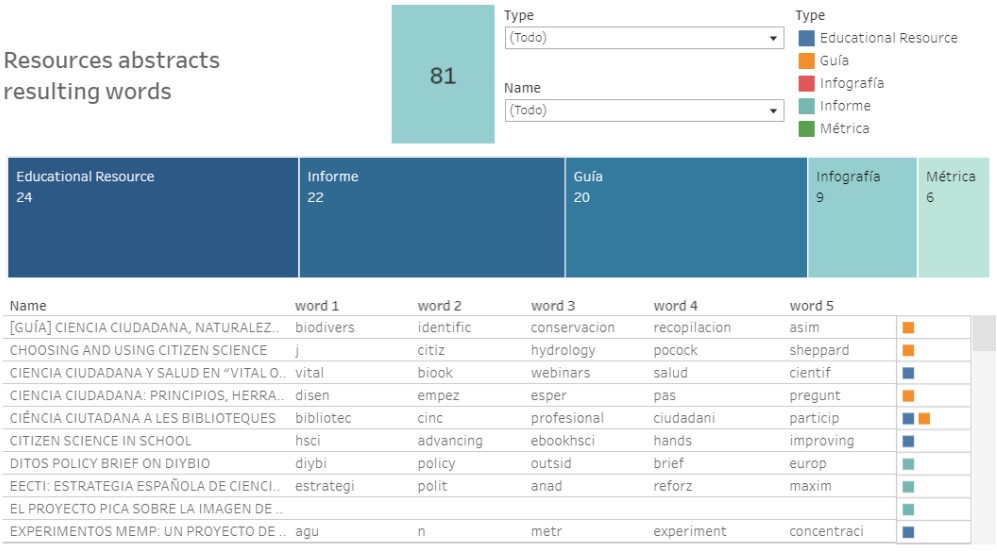


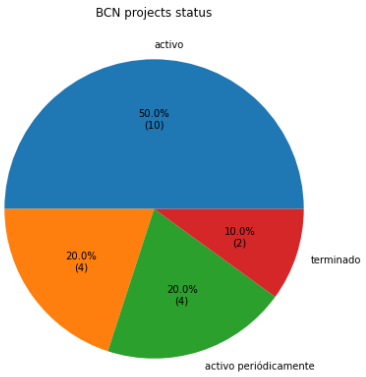
Figure 62.Dashboard with timelines of Spanish initiatives.



Name	word 1	word 2	word 3	word 4	word 5
[GUÍA] CIENCIA CIUDADANA, NATURALEZ..	biodivers	identific	conservacion	recopilacion	asim
CHOOSING AND USING CITIZEN SCIENCE	j	citiz	hydrology	pocock	sheppard
CIENCIA CIUDADANA Y SALUD EN "VITAL O..	vital	biook	webinars	salud	cientif
CIENCIA CIUDADANA: PRINCIPIOS, HERRA..	disen	empez	esper	pas	pregunt
CIÉNCIA CIUTADANA A LES BIBLIOTEQUES	bibliotec	cinc	profesional	ciudadani	particip
CITIZEN SCIENCE IN SCHOOL	hsci	advancing	ebookhsci	hands	improving
DITOS POLICY BRIEF ON DIYBIO	diybi	policy	outsid	brief	europ
EECTI: ESTRATEGIA ESPAÑOLA DE CIENCI..	estrategi	polit	anad	reforz	maxim
EL PROYECTO PICA SOBRE LA IMAGEN DE ..					
EXPERIMENTOS MEMP: UN PROYECTO DE ..	agu	n	metr	experiment	concentraci

Figure 63.Spanish resources' dashboard.

8.9) Appendix I



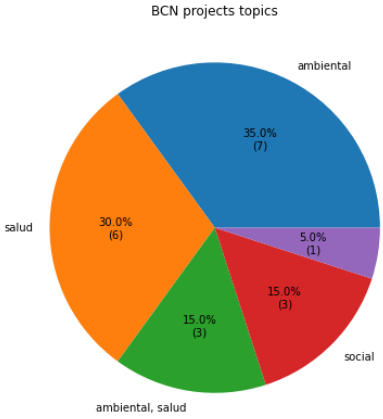


Figure 64. BCN projects' status.

Figure 65.BCN projects' topics.

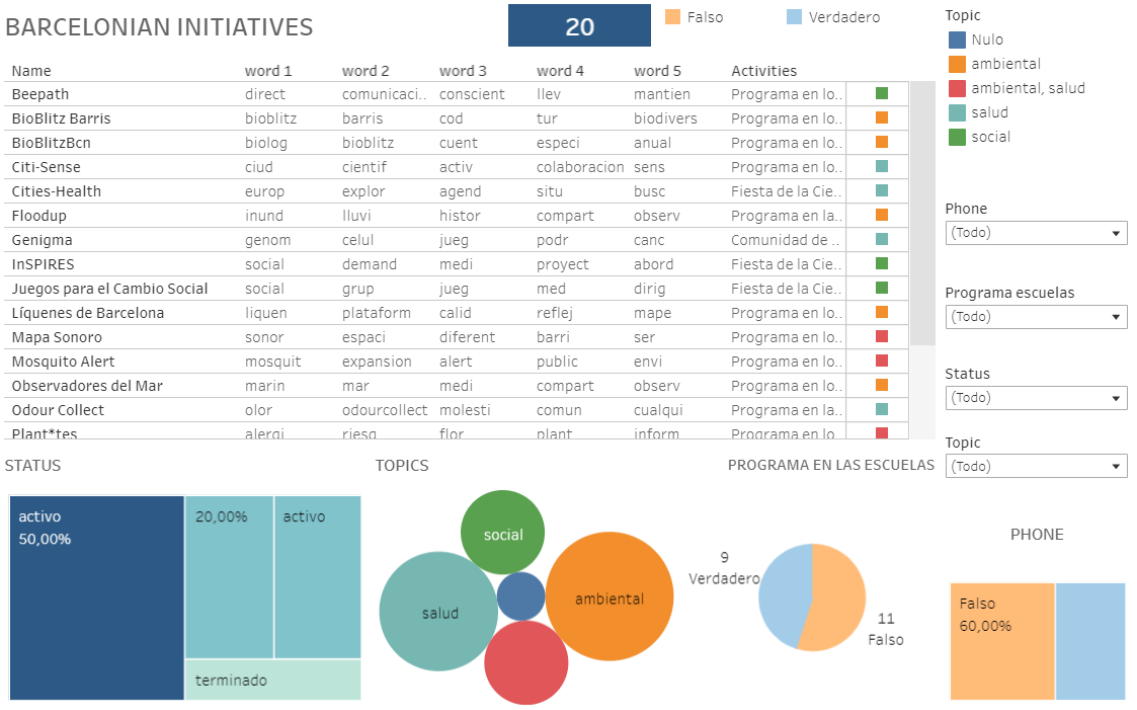


Figure 66. Barcelonian projects dashboard.

8.10) Appendix J

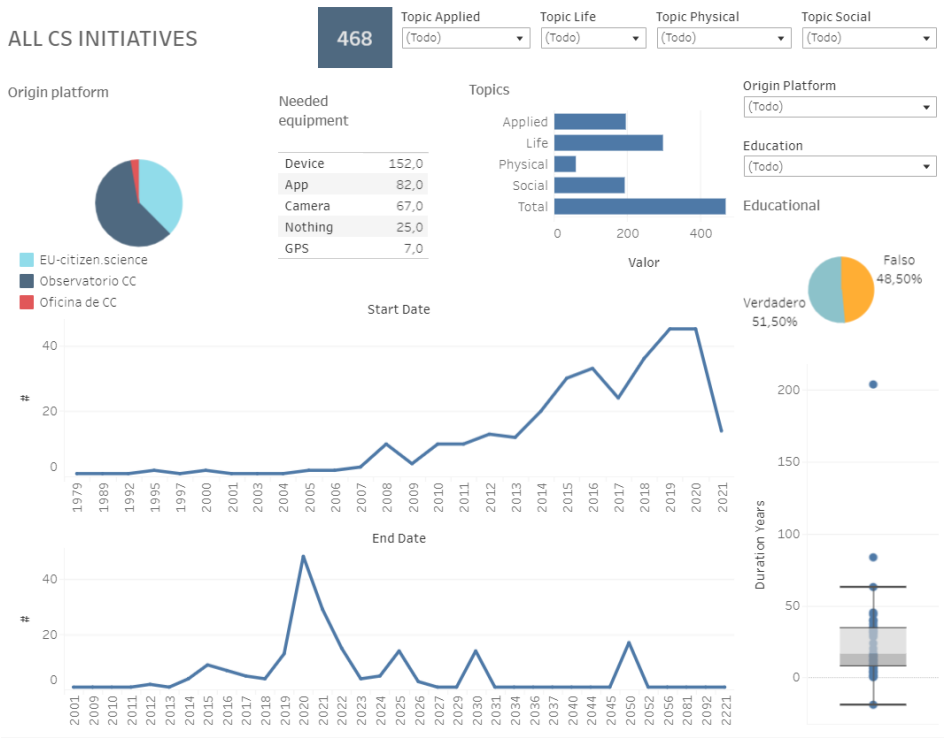


Figure 67.Dashboard with all CS initiatives.

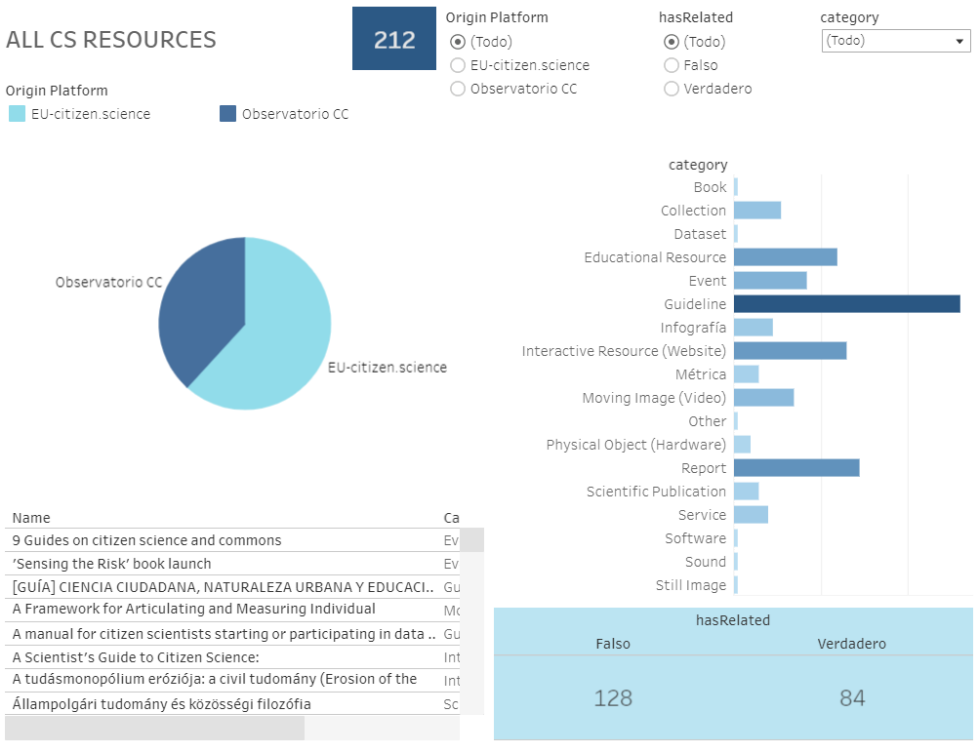


Figure 68.Dashboard with all CS resources.

8.11) Appendix K

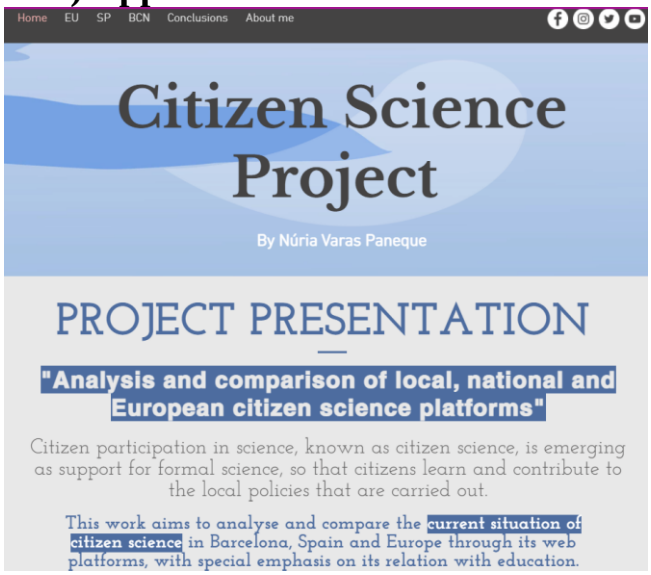


Figure 69.Website main page.

Figure 70.Website European section.

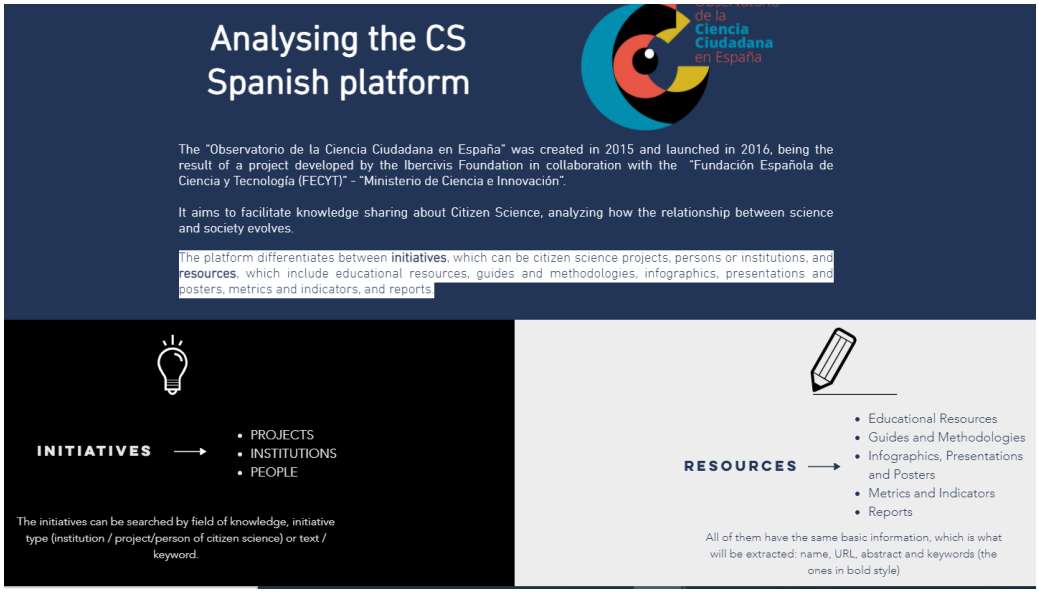


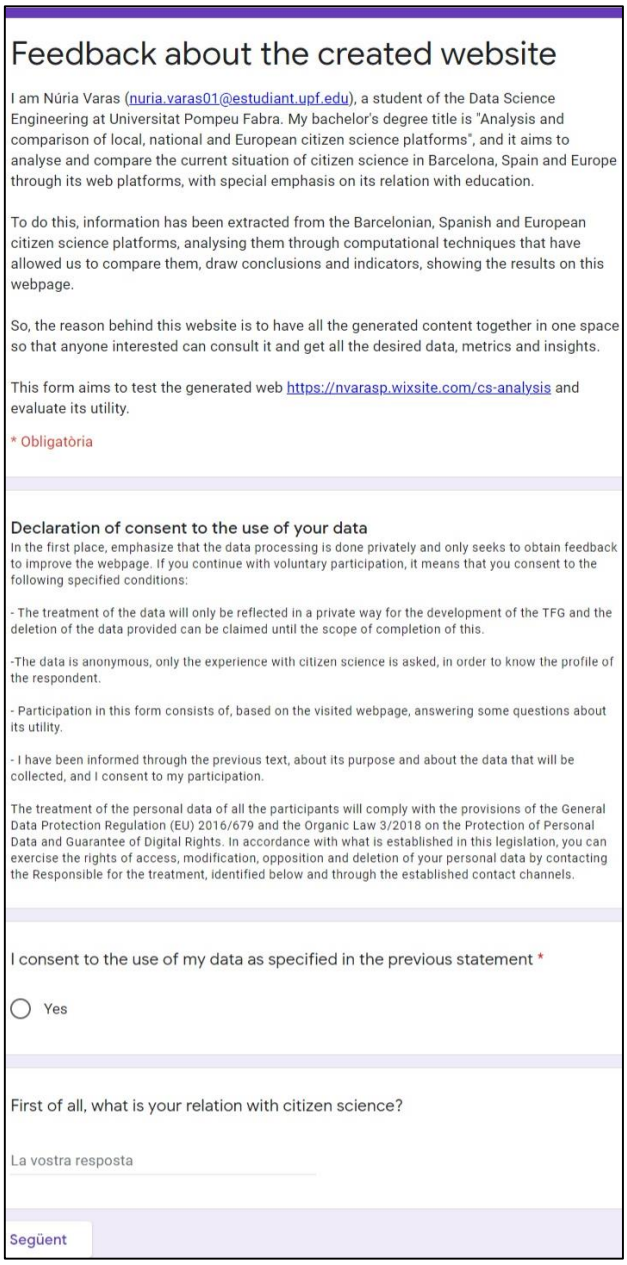
Figure 71.Spanish section of the website.

Figure 72.Barcelonian section of the website.

Figure 73.Conclusions section.

Figure 74.About me.

8.12) Appendix L



Feedback about the created website

I am Núria Varas (nuria.varas01@estudiant.upf.edu), a student of the Data Science Engineering at Universitat Pompeu Fabra. My bachelor's degree title is "Analysis and comparison of local, national and European citizen science platforms", and it aims to analyse and compare the current situation of citizen science in Barcelona, Spain and Europe through its web platforms, with special emphasis on its relation with education.

To do this, information has been extracted from the Barcelonian, Spanish and European citizen science platforms, analysing them through computational techniques that have allowed us to compare them, draw conclusions and indicators, showing the results on this webpage.

So, the reason behind this website is to have all the generated content together in one space so that anyone interested can consult it and get all the desired data, metrics and insights.

This form aims to test the generated web https://nvarasp.wixsite.com/cs-analysis and evaluate its utility.

* Obligatòria

Declaration of consent to the use of your data

In the first place, emphasize that the data processing is done privately and only seeks to obtain feedback to improve the webpage. If you continue with voluntary participation, it means that you consent to the following specified conditions:

- The treatment of the data will only be reflected in a private way for the development of the TFG and the deletion of the data provided can be claimed until the scope of completion of this.

-The data is anonymous, only the experience with citizen science is asked, in order to know the profile of the respondent.

- Participation in this form consists of, based on the visited webpage, answering some questions about its utility.

- I have been informed through the previous text, about its purpose and about the data that will be collected, and I consent to my participation.

The treatment of the personal data of all the participants will comply with the provisions of the General Data Protection Regulation (EU) 2016/679 and the Organic Law 3/2018 on the Protection of Personal Data and Guarantee of Digital Rights. In accordance with what is established in this legislation, you can exercise the rights of access, modification, opposition and deletion of your personal data by contacting the Responsible for the treatment, identified below and through the established contact channels.

I consent to the use of my data as specified in the previous statement *

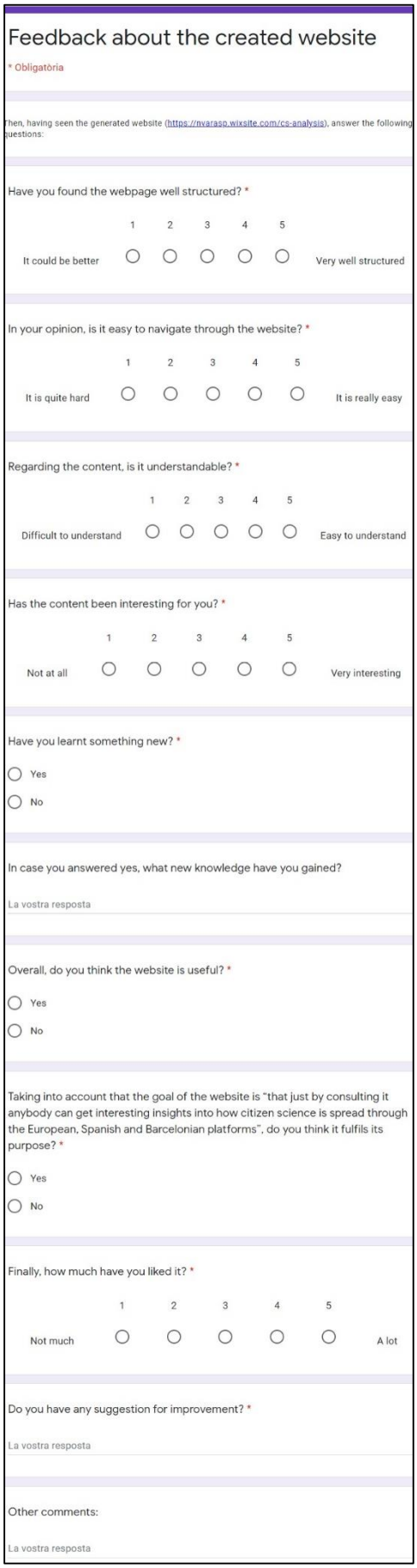
Yes

First of all, what is your relation with citizen science?

La vostra resposta

Següent

Figure 75. Screenshot of the form.



Feedback about the created website

* Obligatòria

Then, having seen the generated website (https://nvarasp.wixsite.com/cs-analysis), answer the following questions:

Have you found the webpage well structured? *

1 2 3 4 5

It could be better — Very well structured

In your opinion, is it easy to navigate through the website? *

1 2 3 4 5

It is quite hard — It is really easy

Regarding the content, is it understandable? *

1 2 3 4 5

Difficult to understand — Easy to understand

Has the content been interesting for you? *

1 2 3 4 5

Not at all — Very interesting

Have you learnt something new? *

Yes

No

In case you answered yes, what new knowledge have you gained?

La vostra resposta

Overall, do you think the website is useful? *

Yes

No

Taking into account that the goal of the website is "that just by consulting it anybody can get interesting insights into how citizen science is spread through the European, Spanish and Barcelonian platforms", do you think it fulfils its purpose? *

Yes

No

Finally, how much have you liked it? *

1 2 3 4 5

Not much — A lot

Do you have any suggestion for improvement? *

La vostra resposta

Other comments:

La vostra resposta

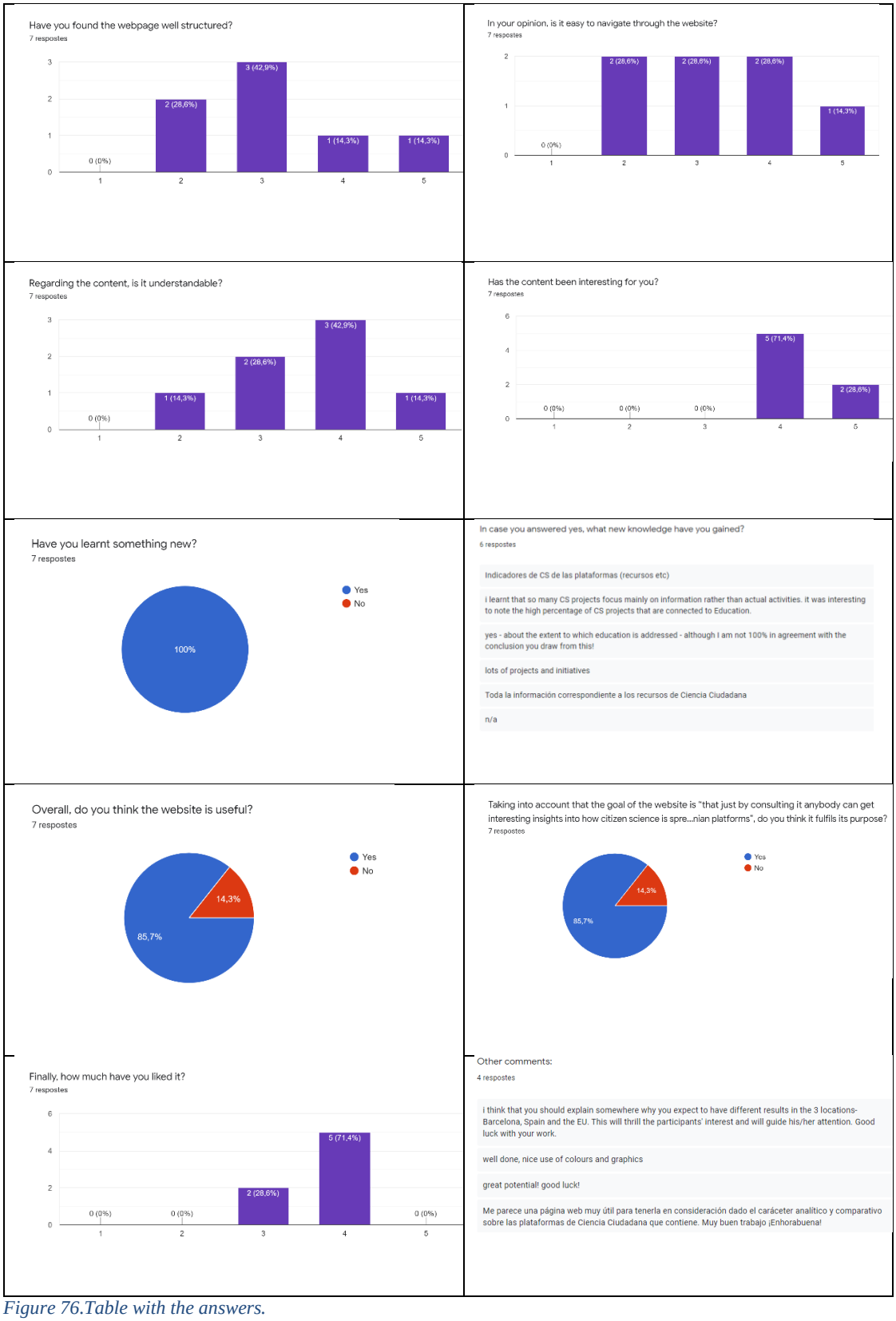


Figure 76.Table with the answers.