Comunicar, n. 74, v. XXXI, 2023 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com

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

Cómo automatizar la extracción y análisis de información sobre ciencia ciudadana con propósitos educativos

Miriam Calvera-Isabal. Research Assistant, Research Group of Interactive and Distributed Technologies for Education (TIDE), University of Pompeu Fabra, Barcelona (Spain) (miriam.calvera@upf.edu) (https://orcid.org/0000-0003-4117-6953)
Dr. Patricia Santos. Researcher, Research Group of Interactive and Distributed Technologies for Education (TIDE), University of Pompeu Fabra, Barcelona (Spain) (patricia.santos@upf.edu) (https://orcid.org/0000-0002-7337-2388)
Dr. H.-Ulrich Hoppe. Full Professor, Rhine-Ruhr Institute for Applied Systems Innovation, Duisburg (Germany) (uh@rias-institute.de) (https://orcid.org/0000-0003-3240-5785)
Cleo Schulten. Research Assistant, Rhine-Ruhr Institute for Applied Systems Innovation, Duisburg (Germany) (cs@rias-institute.de) (https://orcid.org/0000-0003-3082-6084)

ABSTRACT

There is an increasing interest and growing practice in Citizen Science (CS) that goes along with the usage of websites for communication as well as for capturing and processing data and materials. From an educational perspective, it is expected that by integrating information about CS in a formal educational setting, it will inspire teachers to create learning activities. This is an interesting case for using bots to automate the process of data extraction from online CS platforms to better understand its use in educational contexts. Although this information is publicly available, it has to follow GDPR rules. This paper aims to explain (1) how CS communicates and is promoted on websites, (2) how web scraping methods and anonymization techniques have been designed, developed and applied to collect information from online sources and (3) how these data could be used for educational purposes. After the analysis of 72 websites, some of the results obtained show that only 24.8% includes detailed information about the CS project and 48.61% includes information about educational purposes or materials.

RESUMEN

El interés y la práctica de la ciencia ciudadana (CC) ha aumentado en los últimos años. Esto ha derivado en el uso de páginas web como herramienta de comunicación, recolección o análisis datos o repositorio materiales y recursos. Desde una perspectiva educativa, se espera que al integrar información sobre proyectos de CC en un entorno educativo formal, se inspire a los maestros a crear actividades de aprendizaje. Este, es un caso interesante para usar bots que automaticen el proceso de extracción de datos de webs de CC que ayuden a comprender mejor su uso en contextos educativos. Aunque esta información está disponible públicamente, se deben seguir las reglas de la ley de protección de datos o GDPR. Este artículo tiene como objetivo explicar: 1) cómo la CC se comunica y promueve en los sitios web; 2) cómo se diseñan, desarrollan y aplican los métodos de web scraping y las técnicas de anonimización para recopilar información en línea; y 3) cómo se podrían usar estos datos con fines educativos. Tras el análisis de 72 webs algunos de los resultados son que solo el 24,8% incluye información detallada sobre el proyecto, y el 48,61% incluye información sobre propósitos o materiales educativos.

KEYWORDS | PALABRAS CLAVE

Citizen science, informal learning, algorithms, automatization, education, privacy protection.
Ciencia ciudadana, aprendizaje informal, algoritmos, automatización, educación, protección de la privacidad.

Received: 2022-05-30 | Reviewed: 2022-06-27 | Accepted: 2022-07-19 | OnlineFirst: 2022-10-30 | Published: 2023-01-01
DOI https://doi.org/10.3916/C74-2023-02 | Pages: 23-34

1. Introduction and state of the art

Citizen Science (CS) is the active engagement of the general public in scientific research tasks (Vohland et al., 2021). CS activities are typically organized in projects with a strong online presence via web pages and platforms which are used as data dissemination, participation and repository tools (Vohland et al., 2021). There are several international CS associations: The Citizen Science Association (CSA-North America), the European Citizen Science Association (ECSA) and the Australian Citizen Science Association (ACSA). In addition, there are national or regional associations such as *Observatorio de la ciencia ciudadana* (Spain) or Bürger schaffen Wissen (Germany) or individual projects such as Cities-Health. Information on CS activities can also be found on the websites of research institutes, universities, museums, etc. The variety of CS institutions demonstrates that communication about projects can be done through different channels (individual, as part of a 'network' or association, at local, regional or larger scale). Although the communication approach will vary throughout the project and might be different for each type of project, it is important to define it well in order to engage, retain, motivate or inform volunteers (Vohland et al., 2021; Veeckman et al., 2019). As Lin-Hunter et al. (2020) concluded in their analysis about the volunteers' tasks described in the CS project description and its connection to participant's scientific literacy development, how CS project is communicated may affect volunteers' engagement and might imply changes on public science perception and awareness of the problem to be addressed.

The Internet (through websites) or the television has historically contributed to informal science learning and science communication (Stocklmayer et al., 2010). The existence of various formations in CS demonstrates that communication about projects can be done through different channels (individual, as part of a "network" or association, at local, regional or larger scale). The materials provided on these platforms have a great potential to be used for educational purposes, especially in relation to Sustainable Development Goals (SDGs) taking into account that many CS projects address sustainability issues (Fraisl et al., 2020; Storksdieck et al. 2016). However, although multiple projects are collected in the national or global platforms, there is no centralized database that contains global information about all CS projects (Vohland et al., 2021).

Among the potential educational benefits of CS activities, we see the improvement of, scientific knowledge and understanding, the development of technical/scientific skills, STEM career motivation and values such as sustainability or respect for the environment (Hiller & Kitsantas, 2014; Bonney et al., 2016; Kobori et al., 2016; Vohland et al., 2021). Although CS projects do not usually primarily aim at fostering citizen's scientific literacy and knowledge, they often develop educational materials or conduct training activities to prepare participants for participating in scientific activities such as collecting or classifying data (Bonney et al., 2009). More and more frequently, the participation of schools in CS projects is promoted by institutions (e.g. the Oficina de Ciencia Ciudadana in Barcelona has an open call for schools to participate in CS projects: https://bit.ly/3cB1lMH), and this is increasing. However, there is still a lack of knowledge about how CS can be more centrally integrated in schools as a guide or source of inspiration for teachers to create activities aligned with current research and societal problems addressed by the CS projects. All the materials and data generated by CS projects could be used for students' learning about specific topics or support teachers' practice. This is a task for both scientists and educators to work together, so communicating science (through workshops, learning activities or informal conversations) might have an impact on the public understanding of scientific facts and knowledge (Bickford et al., 2012; Stocklmayer et al., 2010).

Given the massive presence and availability of online information on CS projects and activities, it appears promising to use computational analytics techniques to generate specific insights into the functioning and evolution of CS activities. There are many fields in which such tools have been used, especially to massively extract data from online sites and store these in databases (Diouf et al., 2019). There are few examples of use in the CS field (Ponti et al., 2018).

From a European perspective, there is a specific interest in better understanding the role of CS in science and society, e.g., the actual distribution and contribution in geographical regions, distribution over disciplines, as well as the importance of science communication in the CS field and the impact on education. There is still a lack of knowledge as to how CS projects are distributed for further developing and supporting

specific types of CS (Warin & Delaney, 2020). The work reported here is part of the EU project CS Track (https://cstrack.eu/) that operates in this line of research. For this purpose, CS Track relies on a combination of web analytics techniques and classical social studies methods. CS Track has built up a database comprising information about 4,949 CS projects that were gathered from different sites. This is the basis for the on-going extraction and further enrichment of descriptive information related to these projects. All the data centralized will allow us to know more about how CS is communicated online and to broaden our knowledge on the connections to education.

In this paper, we explain how to build a central point of knowledge about CS using as a base the information about CS projects distributed on different websites. It will allow us to see the differences and similarities between the data structures of the websites to report the data. As part of this data extraction and analysis, we have particularly tried to identify the potential for supporting educational purposes. In this work, we have been aware of constraints that are legitimately imposed by privacy and data protection principles, especially the European General Data Protection Regulation (GDPR). The GDPR aims to give citizens control over their personal data and enforces the anonymization of data unless there is no specific individual consent. A dataset can be considered anonymous if a person can't be re-identified (Gruschka et al., 2018). Although the data extracted about CS projects describes project characteristics, sometimes direct or indirect personal data is informed through the texts. The work reported here has been guided by the following research goals:

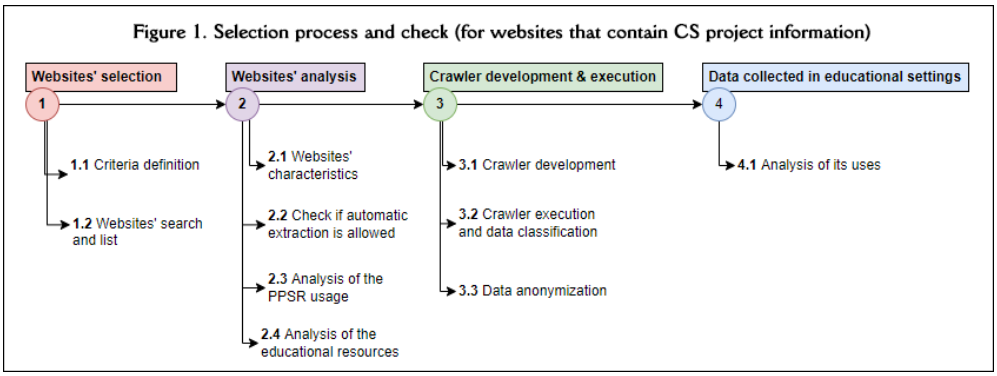
- (RG1) Design and implement an automatic algorithm to extract data from CS platforms in a unique central point (database). The extracted data should be aligned with the PPSR metadata, extended if necessary.
- (RG2) Find technical solutions to comply with GDPR requirements in this context.
- (RG3) Identify the potential educational uses of the data collected.

2. Methodology and data selection

The source of information for this study was websites that contain information about CS projects. The following criteria of inclusion was applied to identify online data from CS projects (unit of analysis):

- The website contains a list of CS projects information or are the websites of a single project.
- From Europe, associated countries or are fully conducted online.
- It is allowed to extract the data either automatically or manually.

Figure 1. Selection process and check (for websites that contain CS project information)



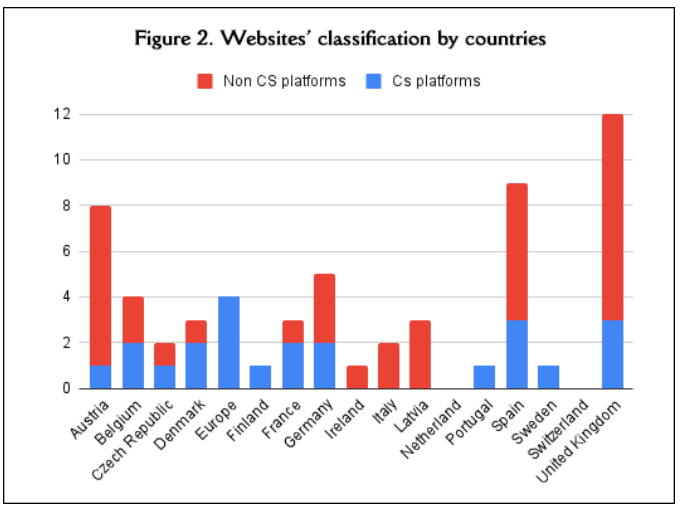
In the first phase, all the consortium members were asked to do manual online research of all the websites that could contain information about CS in European regions. After this, we manually explored each one to identify which ones contain specific information about CS projects and follow the criteria defined below. The websites' identification, selection and analysis were done manually and consist of 72 online sites. This list can be extended in next iterations. It is possible we could not identify all the existing websites that follow the criteria but, the most relevant ones were selected. The manual analysis of the website had two main objectives: (1) to identify how information of CS projects is reported, the main elements of information, the geographical distribution of websites and languages (2) to understand

the technical structure of the data and how it can be aligned with the PPSR metadata standard. Figure 1 shows the process followed during this research.

3. Citizen science presence in online platforms

This section explains the second phase of the process followed (Figure 1) and the results and findings obtained from the analysis. We classified the websites into two categories: CS platforms (29 websites) and non-citizen science platforms (43 websites). CS platforms are those digital platforms that share information about CS projects, activities, events, materials or resources, news about the field, communication tools (i.e., comments or forums) or sometimes, they are also used as a participatory tool (Sanz et al., 2019). Non-citizen science platforms' first objective is not to inform about CS, but has been created as a communication tool, as a repository or even to allow user's interaction. We analyze platform descriptions obtained from the project websites. For those classified as non-citizens' science platforms we can find a diversity of associations (i.e. Helmholtz Association), museums (i.e. Natural history museum UK) or research institutes (ICM Divulga). In the description, these websites use terminology such as "national scientific communication networking center", "Museum", "independent non-profit organization" or "Research Transfer Office" to define the association or organization.

The classification of CS platforms has been carried out following the criteria proposed by Vohland et al., (2021) which differentiates between five types of platforms. Due to the criteria followed for data selection, the category "World-wide citizen science platform" has been added for those platforms that have projects from all over the world. After analyzing the platform descriptions, we categorized them into: Commercial Platforms for CS Initiatives (2 websites), CS Platforms for Specific Projects (8 websites), CS Platforms for Specific Scientific Topics (2 websites), National CS Platforms (15 websites), EU Citizen Science Platforms (1 website) and World-wide citizen science platform (2 websites). In these texts we read the terms such as "citizen science portal" or "online citizen science hub" which are used to identify it as CS platforms and others such as "center of citizen science" or "citizen science network" in reference to the CS associations that coordinate the website. It is common for CS projects to use websites as a participatory tool, for this reason, when we read the CS platforms for specific projects, they use terms such as "simulator" or "webtool".



Europe is a continent in which cultures and languages coexist. To understand the distribution of websites across Europe, the websites have been analyzed from two points of view: the geographical location of the platform and the languages available. A total of 17 out of 44 countries have been identified in the list of websites. Figure 2 shows the countries distribution by the two types of platforms. All the online platforms considered to be "World-wide" such as SciStarter (https://scistarter.org/), iNaturalist network (https://www.inaturalist.org/), Zooniverse (https://www.zooniverse.org/) and Instant wild (https://instantwild.zsl.org/intro) have been excluded because, although we could assign to each one

a single country, they share information about projects or initiatives from all over the world. In order to better understand this geographical distribution and the citizen outreach they could achieve, it is important to also understand the linguistic diversity of Europe. Several online platforms facilitate the use of more than one language. For instance, 29.7% platforms facilitate the use of two languages (i.e. Iteritalia), 8.1% of platforms facilitate the use of three languages (i.e. OpenSystems UB) and 4.1% of platforms facilitate the use of more than three languages (i.e. EU Citizen science). 58.1% of platforms only support the use of one language (i.e. Desqbre). As stated in the Charter of fundamental rights of the European Union (European Union, 2010), "the union shall respect (...) linguistic diversity" as well as "any discrimination based on (...) language (...) shall be prohibited". The EU has 24 official languages and other regional and minority languages. 17 out of 24 languages have been covered by the websites selected. Moreover, three regional languages (Euskera or Catalan) and one extra community language (Arab) have been identified as languages in them. CS platforms cover 84.3% of languages identified.

Although we indicate which country each website can be related to, when we read the descriptions, we realized that 35.6% provides information about the region covered (such as "in Flanders" (Citizen Science Vlaanderen) or "Globally"). The Websites cover regional or national areas (i.e., Barcelona CS platform (Barcelona)), Europe (i.e., EU Citizen science platform) and all areas of the world (i.e., iNaturalist). Geographic region covered by the online platforms is aligned with the language available.

From the same information used to classify websites, key terms have been extracted to assign specific research areas. Terminology such as "protecting our planet", "science used in the investigation of crime science, laboratory analysis and the presentation of scientific evidence within the courts" or "meteorological and geophysical services" has been selected to identify the category. The platforms were classified into the six broad research areas defined in Web of Science Core Collection (Clarivate analytics, 2022): Arts & Humanities (1.37%), Life Sciences & Biomedicine (50.68%), Physical Sciences (1.37%), Social Sciences (2.74%), Technology (0%) and All (43.84%).

3.1. Websites functionalities and applications

For this research, we applied manually the platform's taxonomy defined by Derave et al. (2020). Although it defines seven categories, we have only analyzed the first three due to the websites selected being participation and communication oriented and we focus our attention on this.

Table 1. Platforms classification based on "Market side", "Affiliation" and "Centralization"

	Zero-side		**One-side**		**Multi-side**			
Market side	All websites	CS platforms	All websites	CS platforms	All websites		CS platforms	
	20.83%	17.24%	52.78%	41.38%	26.39%		41.38%	
Affiliation			**Registration**					
			All websites		**CS platforms**			
			43.06%		37.93%			
			Subscription					
			All websites		**CS platforms**			
			51.39%		55.17%			
					No transaction		**Transaction**	
					All websites	**CS platforms**	**All websites**	**CS platforms**
					79.17%	72.41%	20.83%	27.95%
							Investment	
							0%	
Centralization					**Centralized**			
					All websites		**CS platforms**	
					33.33%		51.72%	
					Decentralized			
					All websites		**CS platforms**	
					66.67%		48.28%	

- Market sides: It is the first category that defines the number of user groups. We also included the term Zero-side. From the selected sites, we identified 15 sites Zero-side (no interactions between users, only between them and website manager), 38 One-side (users' interaction is

with the platforms and indirectly with other users, via comments or posts) and 19 Multi-side (interaction within the users and platform). CS platforms are commonly One-side or Multi-side oriented although, as they are sometimes used as a participatory tool, it is common that they allow users to interact with each other (i.e., via forums or comments) (Table 1).

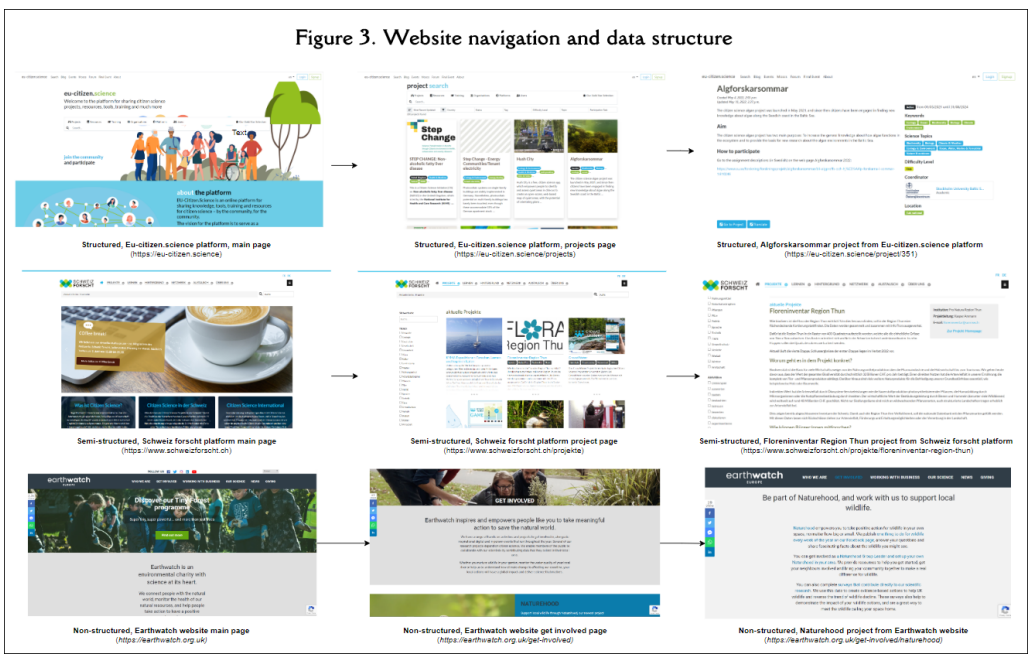
- Affiliation: It refers to how the users interact with others and the website. There are many functionalities designed for giving users the opportunity to be connected (i.e., forums, newsletters, etc.) and can be combined in a single site. We identify 31 sites that allow registration, 37 that allow subscription, 19 that allow main content creation or commenting, 15 that allow transaction and 0 that allow investment. For Multi-side platforms it is common to involve users in commenting or content creation, ask for registration and give the option of being connected and informed via subscription. Nevertheless, for the other types, registration is not a main requirement but subscription is highly recommended to be connected. The most common tool used for user interaction is the forum, but only 8 websites have one (5 of them are CS platforms). The second option for user interaction is adding comments (no direct messages). Only 6 of the platforms allow this type of functionality. As sites for scientific research, dissemination of results and participant" engagement is important. There are special pages for news or a blog created in these sites: 47 of them have one, 20 of them being for CS platforms.
- Centralization: the third category is aligned with the second one because it indicates the way the users connect among each other. There are 48 Decentralized websites while there are 13 Centralized websites.

3.2. Citizen science information online

In this section, it is explained how CS information is shared in websites from two points of views: How data are structured and what kind of data are shared. The analysis is necessary for the algorithm and database design.

Although each web page follows its own design and data structure, it is common for all to have a main page, then a page with the list of projects and indexed links to another page with the information of the associated CS project (Figure 3). However, there is an exception for CS platforms for specific projects since it contains information about a single project distributed in pages, not a list of individual projects as the others.

Figure 3. Website navigation and data structure



As a first step, the original data source was classified according to three types of web pages (Figure 3):

- Structured: information is presented in categories (i.e. https://eu-citizen.science/).
- Semi-structured: some information is presented incategories and another is presented in paragraphs (i.e. https://www.zooniverse.org/).
- Non-structured: no information categorized (i.e. https://www.scivil.be/en).

From the 72 websites we can identify 13 websites with structured data, 27 with semi-structured data and 34 with non-structured data. Several working groups from CS associations (Data & Metadata working group (CSA) or Working Group on "Data, Tools and Technology" (ECSA)) are focused on promoting standardization of CS data. This is the case, for example, of the Public Participation in Scientific Research - Core (PPSR-Core) data model which proposes a data standard and works on promoting it to be accepted and used by CS websites (Bowser et al., 2017). Regarding the CS projects information, in order to analyze how this data is shared online and define a common structure for the database to store the data classified, it has been necessary to identify how the information associated with a CS project can be classified according to the PPRS-Core metadata standard attributes and some that are newly created. We analyzed how this standard is followed in the websites, and the Title was well identified (commonly it is the first shown and bigger than other texts) and Description (below the title). For the other categories, we have created a dictionary of terms containing 19 categories (15 included in the metadata standard and 4 added to the standard), based on similar terms/sections contained in the different websites analyzed:

- Social media: the name of the social media platform (i.e., "twitter" or "facebook") or general terms (i.e., "blog.", "REDES SOCIALES:" or "PERFILES EN REDES SOCIALES:").
- Online resources: file extension formats (i.e., ".pdf) or general terms (i.e., "OTROS RECURSOS DEL PROYECTO:", "Desktop:").
- Tools and materials: only one expression selected "Tipo de medios".
- Applications used: applications repositories names although could be integrated into Tools and materials category (i.e., "play.google" or "apple.com") or general terms (i.e., "Mobile:").
- Images: images file extension (i.e., ".jpg", ".png", ".JPG" or ".jpeg").
- Geographical location: general terms used (i.e., "Geographical", "Geographic Scope", "WHERE", "Ubicación", "places", "Project Location" or "Location") or specific terms for regions or areas (i.e., "Country", "PROVINCIA:" or "País ").
- Status: general terms (i.e., "Project Status", "Status" or "ESTADO DEL PROYECTO:").
- Methodology - Participants tasks: general terms (i.e., "Participation Tasks" or "Tasks") and open questions about the participation (i.e., "HOW TO GET STARTED", "RELACIÓN CON LA CIENCIA CIUDADANA:" or "¿Cómo participan los voluntarios/as?").
- Start date: general terms (i.e, "Start Date", "FECHA DE INICIO DEL PROYECTO:" or "Projektstart:").
- Investment or support: general terms (i.e., "Sponsor", "TOTAL EXPENSE" or "Project Funding").
- Field of science: general term (i.e., "Fields of Science", "TOPICS", "ÁREA DE CONOCIMIENTO:").
- Development time: general terms (i.e., "Intended Outcomes", "IDEAL FREQUENCY", "When? " or "Période : ").
- Main objectives: general terms (i.e., "Goal", "Waarom doe je mee?" or "Objet : ").
- Participants age: only the term "IDEAL AGE GROUP".
- Participants profile: general terms (i.e., "Wie kan meedoen?", "Usuarios", "/people/", "PÚBLICO AL QUE SE DIRIGE EL PROYECTO:", "INTEGRANTES DEL PROYECTO:", "INTEGRANTES:", "Who can take part?", "Public: ", "Project Partners" or "Users").
- Development place: general terms to explain the space or area to develop activities research (i.e., "SPEND THE TIME", "Region", "Ubicación", "ÁMBITO DE ACTUACIÓN:" or "Type of activity:").
- Dedication time: general terms to explain how much time participants will invest in participation (i.e., "AVERAGE TIME" or "How long will it take? ").

- Contact information: "@" is used in the email addresses.
- Project update date: only the term "PROJECT UPDATED".
- "Main program or person in charge": in this category it is combined information about the information creator, coordinators or managers and associations that support or collaborate (i.e., "PRESENTED BY", "Wie organiseert het?", "/researcher/", "Creado por:", "Administradores de proyecto:", "Administrador de proyecto:", "MIEMBROS DEL EQUIPO:", "OTROS GRUPOS O INSTITUCIONES COLABORADORES:", "OTRAS PERSONAS O ENTIDADES COLABORADORAS:", "Project Manager", "Project Co-ordinator" or "Kontakt:").

4. Algorithm development and execution

Our goal is to have all the information stored in a database, so it is essential to choose the correct ones based on the type of data extracted from the online platforms selected. In this section, the process followed in the third phase of the process is described (see section 2).

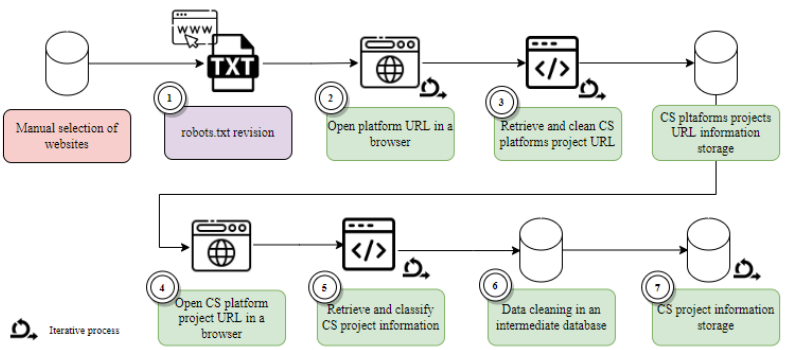
4.1. Database selection

In order to select the database, the comparison of the type of databases was made between relational databases (those accepting Structured Query Language, (SQL)) and non-relational databases (those which do not accept SQL) (Li & Manoharan, 2013). MongoDB database version 4.2 was selected because it can store structured and unstructured data; it can easily grow; the database structure can be changed independently to other data collections and documents (data structure depends on each project description) and it admits queries and data consumption.

4.2. Algorithm development and execution

The algorithm is adapted to the three types of web structures. In accessing CS project information, we applied a two-step process: first to access a main page where projects are listed, and second, select a certain CS project to see its information. Figure 4 shows the process followed by the algorithm. It was developed in Python programming language (using selenium, b4soup, requests and PyMongo libraries). In this process, it was necessary to take into account if the web pages had an Application Programming Interface (API) (the EU. Citizen science and iNaturalist websites) which allows the automatic extraction of data from the database source.

Figure 4. Process followed by the algorithm



The robot exclusion protocol regulates, for bots, access to the source code of the website (Kolay et al., 2008). It is defined by each website and informed in the robots.txt file which is accessible via the website URL. The restrictions combine blocking all or partial content for all or certain bots. This information is checked before data extraction. We identified: 4 websites didn't define robots' exclusion protocol; 20 websites contain restrictions but not for the algorithm created or the specific content to extract and 1 website whose protocol does not allow ".pdf" extraction. For each website, the parser searches the source code manually identified previously and extracts specific web elements (Parvez et al., 2018). Furthermore,

user actions such as pressing buttons have to be replied to because some pages which contain CS project information have more than one tab.

4.3. Data classification, cleaning and storing

The CS project data classification is done automatically searching by keywords, symbols or sentences previously defined (See section 3.3). A cleaning process was applied to remove wrong data or all the elements unclassified and the final information is stored in the database. To avoid duplicates, it is checked if the project title exists in the database. Additional information such as storage date, update date (in case it already exists), origin or ID is added before storing. The information obtained through the API has to be mapped to the defined database structure (cleaning is not needed). This process can be reproduced when needed, so the number of projects and information can increase. Figure 5 shows an example of the information stored.

Figure 5. The Neureka project information stored in the database


```
{
    "_id": {
        "$oid": "5efc3ae868516f3f3e568278"
    },
    "TITLE": "The Neureka Project",
    "WEB": ["https://www.neureka.ie/", "WEBSITE The Neureka Project"],
    "SOCIAL MEDIA": ["https://www.youtube.com/watch?v=TkWtYafAAT4", "https://twitter.com/@neurekaApp"],
    "DESCRIPTION": ["Neureka is a collection of fun brain games and challenges that allows you to solve cogr
    "Plat Id": ["65", "17"],
    "Plat country": ["USA", "EU"],
    "Insert date": "2020-07-01",
    "APPS": ["https://play.google.com/store/apps/details?id=com.gillanlab.neureka.beta", "https://apps.apple
    "GEOGRAPHICAL LOCATION": ["WHERE Online"],
    "METHODOLOGY": ["HOW TO GET STARTED Download neureka from either the Google Play Store or Apple App Stor
    "INVESTMENT": ["TOTAL EXPENSE 0.00"],
    "TOPICS": ["Biology", "Psychology", "Health & Medicine", "brain", "dementia", "depression", "neuroscienc
    "DEVELOPMENT TIME": ["IDEAL FREQUENCY Twice per day"],
    "DEDICATION TIME": ["AVERAGE TIME 1 hour"],
    "DEVELOPMENT SPACE": ["SPEND THE TIME Indoors"],
    "PLATFORM UPDATE DATE": ["PROJECT UPDATED 06/22/2020, 08:25 pm GMT+2", "2021-03-03T10:35:12.479221Z"],
    "Url platform": ["https://scistarter.org/the-neureka-project", "https://eu-citizen.science/projects"],
    "MAIN PROGRAM OR PERSON IN CHARGE": ["PRESENTED BY Gillan Lab at Trinity College Dublin", "Trinity Colle
    "MAIN OBJECTIVES": ["GOAL Improve mental health and dementia research"],
    "TOOLS AND MATERIALS": ["MATERIALS A smartphone (either Android or Apple)"],
    "PARTICIPANTS PROFILE": ["SPECIAL SKILLS None"],
    "Wp2 Id": ["3", "17"],
    "Language": ["English"],
    "STATUS": ["Active"],
    "START DATE": ["2020-07-01"],
    "END DATE": ["2025-12-31"],
    "LATITUDE": ["53.343667"],
    "LONGITUDE": ["-6.254445"],
    "COUNTRY": ["IE"],
    "IMAGE": ["https://eu-citizen.science/media/media/images/2021-02-08_020202020404_546_NEUREKA%20LOGO.png"
    "Date update": ["2021-03-19", "2021-04-08", "2021-05-14"],
    "MAIL": null
}
```


We use Named Entity Recognition (NER) paired with the Entity Ruler to identify phone numbers, email addresses and personal accounts based on given regular expression (RegEx) patterns. The algorithm then checks if names of individuals are occurring in connection with personal data found by the Entity Ruler. Names can in some cases remain not-anonymized, such as if there is a wikipedia article for that name, as this indicates that this is either a person in the public eye and their name carries meaning beyond naming a person (i.e. Albert Einstein), or it is a common name that does not identify one single person. Another reason can be if the person's name has an overlap with the project name as we found that some projects center around analyzing data about a single person or a family. Both these reasons are disregarded however if a phone number, email or personal account information is mentioned in the same sentence.

For the anonymization we opted to go for a pseudonymization by repeatedly hashing the selected text parts (Kasgar et al., 2012) and adding the previously ascribed type of data into the text. This way the individual persons as well as the personal data cannot be identified, but we retain some readability of the text and connections over multiple projects via common persons are still possible.

5. Results

In our analysis, we included 4 new categories in addition to the existing 43 categories included in the PPSR-Core standard: 11 required fields, 21 optional fields and 15 new attributes identified in the websites. This classification is aimed to facilitate the automatic analysis of data in next steps. The main

problem identified is the inconsistency when reporting data about CS projects online, since most of the websites or platforms do not follow metadata standards. The most common attributes (required) are title, description, website link (if exists), social media, contact information or project topic. 91.56% of them are informed. All the information related to data origin, language, storage date and other information necessary for data management is always informed but is not considered for this analysis. Less common attributes (defined as optional or news created by us) such as geolocation or member's age are informed in 24.8%. Data mining techniques have been used to extract data from texts in other fields. An important source of information is the project description, these techniques can be applied to automatically extract information from this category and fill other attributes in the database. For instance, as explained in previous sections, NER has been used to identify persons' names but it can also be used to get information from the text about the organizations, cardinal numbers, countries, cities or states, etc.

Other computational methods oriented for data classification can also be applied to the data in order to add categories or create clusters to easily target the data for easy filtering or give them a context (Roldán-Álvarez et al., 2021). By extracting keywords or meaning from the texts, new categories such as: sustainable development goals (SDGs), learning outcomes or research areas can be created. Result of the methods application can be used to support teachers to select the topic and get inspiration to create learning activities.

6. Discussion and conclusions

The variety of websites that share information about CS projects is also a reflection of how variable CS is. Involving citizens also implies adapting to different forms of communication, either because of the language or the region in which it occurs. Websites are great tools for this communication and sharing with others but also for participation. Even so, there is still work to be done in order to increase public access for CS to be well known and to increase citizen's interest in participating in research. CS platforms should consider being aligned with the PPSR_Core and other metadata standards. Normalizing all the data structures and information shared improves the user's experience in the websites along with facilitating them the search. Having the key information about a CS project all together and well documented could also improve citizen's participation and interest and the research analysis of the CS field. In this line, CS Track projects, besides developing knowledge on the CS field, had opened a new perspective on how computational methods can be applied to centralize all the data into a single database for research purposes. There is still work to be done to analyze and apply data mining methods to the data in order to obtain more information for the empty categories. Nevertheless, the text mining methods are useless without a good and detailed CS project description. It is necessary to involve scientists and communication experts (Roche et al., 2020) and follow guidelines already defined by experts (as the one proposed by Veeckman et al., 2019) for good communication action. In order to have educational impact, it is essential to be aligned with the official curriculum of the educational level to which they refer.

6.1. How to identify content with potential educational benefit?

CS platforms and websites can provide content that can be used as a powerful resource for learning and teaching. A first exploratory study developed by Calvera-Isabal et al. (2021) has explored three CS platforms and found that materials and data related to CS projects extracted from websites have the potential to support teachers in their practice (Asensio-Pérez et al., 2014). Previous publications have stated that data exploration has an impact on student's awareness and interest and promotes discussion, opening new perspectives on how to work mathematics in formal education (Saddiqa et al., 2019). From the data classified into categories, teachers will find a powerful source of scientific knowledge for filtering (for instance, based on Research Areas, SDGs or learning skills). Classifying the data by research areas will allow teachers to better understand the field the project is investigating. Information from SDGs, which are addressing world-wide real problems, can be integrated into the learning designs to motivate students to learn more about and also create awareness about, for instance, sustainability or ecology (Massa et al., 2011; Djonko-Moore et al., 2018). Although CS is being integrated in education and has potential to be integrated in many other ways, only 48.61% contain educational material or information related to learning.

For CS platforms there is a more positive result as 55.17% have these resources. The ones that allow citizens participation have specific pages with educational materials and common questions answered (Zooniverse). Specific CS project platforms are used for communication but also as a repository tool for all the information and documentation they develop (Luonto-Liiton Kevätseuranta). The other websites that are not CS platforms, if they are education oriented (i.e., universities or associations about education or learning), then they share specific materials for teachers or educators. If not, it is not common to share these types of resources. The application of advanced computational techniques and having all the information centralized, can be used to support information online: real problems, research areas, scientific disciplines, learning skills, etc.

Finally, it is common for teachers to integrate technology to support learning or enhance it. For this reason, tools and content developed by CS projects might be integrated in the classroom as an instrument to develop an activity, to participate in CS or even to support them during the lesson preparation. Regarding the potential usage of the data in educational contexts covered in this article, it is also necessary to work more on identifying how to communicate (at the level of data/information to be reported on CS websites) to narrow the link that may exist between CS and formal educational contexts. Some opportunities that arise from this analysis are the usage of the CS project information in educational contexts (such as to inspire teachers to create learning design activities) or the participation of schools in the project (such as particular follow-up cases). It is still necessary to analyze the materials teachers need and to what extent all this information and resources supports them for their teaching practice. It is expected that all this data and resources centralized and available to be explored, have an impact on teacher's scientific knowledge and pedagogical skills, which might affect student's attitude toward science (Chan & Yung, 2018). Finally, the application of algorithms and the collection of mass information allows the unification of data in a single source that could potentially be used for educational purposes. For this, as future work, a digital platform could be developed that communicates CS information to support the creation of activities in the classroom.

Authors' Contribution

Idea, M.C, P.S, U.H; Literature review (state of the art), M.C, P.S, U.H, C.S; Methodology, M.C, P.S, U.H; Data analysis, M.C, C.S; Results, M.C, P.S; Discussion and conclusions, M.C, P.S, U.H, C.S; Writing (final draft), M.C, P.S, U.H, C.S; Final revisions, P.S., U.H; Project design and funding agency, M.C, P.S, U.H.

Funding Agency

This work was supported in part by PID2020-112584RB-C33 funded by MCIN/AEI/10.13039/501100011033, by 'CS Track: Expanding our knowledge on Citizen Science through analytics and analysis' H2020-SwafS-2019-1 proj. ref: 872522 and by the Ramón y Cajal programme (P. Santos).

References

Asensio-Pérez, J.I., Dimitriadis, Y., Prieto, L.P., Hernández-Leo, D., & Mor, Y. (2014). From idea to VLE in half a day: METIS approach and tools for learning co-design. In *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 741-745). ACM. https://doi.org/10.1145/2669711.2669983

Bickford, D., Posa, M.R.C., Qie, L., Campos-Arceiz, A., & Kudavidanage, E.P. (2012). Science communication for biodiversity conservation. *Biological Conservation*, *151*(1), 74-76. https://doi.org/10.1016/j.biocon.2011.12.016

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., & Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, *59*(11), 977-984. https://doi.org/10.1525/bio.2009.59.11.9

Bonney, R., Phillips, T.B., Ballard, H.L., & Enck, J.W. (2016). Can citizen science enhance public understanding of science? *Public Understanding of Science*, *25*(1), 2-16. https://doi.org/10.1177/0963662515607406

Bowser, A., Brenton, P., Stevenson, R., Newman, G., Schade, S., Bastin, L., Parker, A., & Oliver, J. (2017). *Citizen Science Association Data & Metadata Working Group: Report from CSA 2017 and Future Outlook*. European Commision. https://bit.ly/3IS85SI

Calvera-Isabal, M., Varas, N., & Santos, P. (2021). Computational techniques for data science applied to broaden the knowledge between citizen science and education. In D. G. Sampson, D. Ifenthaler, & P. Isaías (Eds.), *Proceedings of the 18th International Conference on Cognition and Exploratory Learning in the Digital Age (CELDA 2021)* (pp. 219-226). IADIS Press. https://doi.org/10.1007/978-3-030-65657-7

Chan, K.K.H., & Yung, B.H.W. (2018). Developing pedagogical content knowledge for teaching a new topic: More than teaching experience and subject matter knowledge. *Research in Science Education*, *48*, 233-265.

https://doi.org/10.1007/s11165-016-9567-1

Clarivate analytics (Ed.) (2022). *Web of Science Core Collection Help*. https://bit.ly/3ts2IZl

Derave, T., Sales, T.P., Gailly, F., & Poels, G. (2020). Towards a reference ontology for digital platforms. In G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, & H. C. May (Eds.), *Conceptual modeling. ER 2020. Lecture notes in computer science* (pp. 289-302). Springer. https://doi.org/10.1007/978-3-030-62522-1_21

Diouf, R., Sarr, E.N., Sall, O., Birregah, B., Bousso, M., & Mbaye, S.N. (2019). Web Scraping: State-of-the-Art and Areas of Application. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 6040-6042). https://doi.org/10.1109/BigData47090.2019.9005594

Djonko-Moore, C.M., Leonard, J., Holifield, Q., Bailey, E.B., & Almughyirah, S.M. (2018). Using culturally relevant experiential education to enhance urban children's knowledge and engagement in science. *Journal of Experiential Education*, *41*(2), 137-153. https://doi.org/10.1177/1053825917742164

European Union (Ed.) (2010). *Charter of fundamental rights of the European Union. Official Journal of the European Union C83*, *53*, *380*. https://bit.ly/3PFo605

Fraisl, D., Campbell, J., See, L., Wehn, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J.L., Masó, J., Penker, M., & Fritz, S. (2020). Mapping citizen science contributions to the UN sustainable development goals. *Sustainability Science*, *15*(6), 1735-1751. https://doi.org/10.1007/s11625-020-00833-7

Gruschka, N., Mavroeidis, V., Vishi, K., & Jensen, M. (2018). Privacy issues and data protection in big data: A case study analysis under GDPR. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 5027-5033). IEEE. https://doi.org/10.1109/BigData.2018.8622621

Hiller, S.E., & Kitsantas, A. (2014). The effect of a horseshoe crab citizen science program on middle school student science performance and STEM career motivation. *School Science and Mathematics*, *114*(6), 302-311. https://doi.org/10.1111/ssm.12081

Kasgar, A.K., Agrawal, J., & Sahu, S. (2012). New modified 256-bi MD5 algorithm with SHA. Compression Function. *International Journal of Computer Applications*, *42*(12), 47-51. https://doi.org/10.5120/5748-7956

Kobori, H., Dickinson, J.L., Washitani, I., Sakurai, R., Amano, T., Komatsu, N., Kitamura, W., Takagawa, S., Koyama, K., Ogawara, T., & Miller-Rushing, A.J. (2016). Citizen science: a new approach to advance ecology, education, and conservation. *Ecological Research*, *31*(1), 1-19. https://doi.org/10.1007/s11284-015-1314-y

Kolay, S., D'Alberto, P., Dasdan, A., & Bhattacharjee, A. (2008). A larger scale study of robots. txt. In *Proceedings of the 17th international conference on World Wide Web* (pp. 1171-1172). https://doi.org/10.1145/1367497.1367711

Li, Y., & Manoharan, S. (2013). A performance comparison of SQL and NoSQL databases. In *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)* (pp. 15-19). IEEE. https://doi.org/10.1109/PACRIM.2013.6625441

Lin-Hunter, D.E., Newman, G.J., & Balgopal, M.M. (2020). Citizen scientist or citizen technician: A case study of communication on one citizen science platform. *Citizen Science: Theory and Practice*, *5*(1). https://doi.org/10.5334/cstp.261

Massa, N., Dischino, M., Donnelly, J.F., & Hanes, F.D. (2011). Creating real-world problem-based learning challenges in sustainable technologies to increase the STEM Pipeline. In *2011 ASEE Annual Conference & Exposition*. https://doi.org/10.18260/1-2–17678

Parvez, M.S., Tasneem, K.S.A., Rajendra, S.S., & Bodke, K.R. (2018). Analysis of different web data extraction techniques. In *2018 International Conference on Smart City and Emerging Technology (ICSCET)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICSCET.2018.8537333

Ponti, M., Hillman, T., Kullenberg, C., & Kasperowski, D. (2018). Getting it right or being top rank: Games in citizen science. *Citizen Science: Theory and Practice*, *3*(1). https://doi.org/10.5334/cstp.101

Roche, J., Bell, L., Galvão, C., Golumbic, Y.N., Kloetzer, L., Knoben, N., Laakso, M., Lorke, J., Mannion, G., Massetti, L., Mauchline, A., Pata, K., Ruck, A., Taraba, P., & Winter, S. (2020). Citizen science, education, and learning: challenges and opportunities. *Frontiers in Sociology*, *5*, 613814-613814. https://doi.org/10.3389/fsoc.2020.613814

Roldán-Álvarez, D., Martínez-Martínez, F., & Martín, E. (2021). Citizen science and open learning: A Twitter perspective. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 6-8). IEEE. https://doi.org/10.1109/ICALT52272.2021.00009

Saddiqa, M., Larsen, B., Magnussen, R., Rasmussen, L.L., & Pedersen, J.M. (2019). Open data visualization in Danish schools: A case study. *Complex Systems Informatics and Modeling Quarterly*, *21*, 1-20. https://doi.org/10.24132/CSRN.2019.2902.2.3

Sanz, F., Gold, M., & Mazzonetto, M. (2019). *D2.3: Platform functionality requirements & specification report*. Zenodo. https://doi.org/10.5281/zenodo.3612808

Stocklmayer, S.M., Rennie, L.J., & Gilbert, J.K. (2010). The roles of the formal and informal sectors in the provision of effective science education. *Studies in Science Education*, *46*(1), 1-44. https://doi.org/10.1080/03057260903562284

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*, (2), 1-1. https://doi.org/10.5334/cstp.55

Veeckman, C.M., Talboom, S., Gijsel, L., Devoghel, H., & Duerinckx, A. (2019). *Communicatie bij burgerwetenschap: Een praktische handleiding voor communicatie en betrokkenheid bij citizen science*. SCIVIL. https://bit.ly/3PKQz50

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., & Samson, R. (2021). *The science of citizen science*. Springer Nature. https://doi.org/10.1007/978-3-030-58278-4

Warin, C., & Delaney, N. (2020). *Citizen science and citizen engagement. Achievements in Horizon 2020 and recommendations on the way forward*. European Commision. https://doi.org/10.2777/05286